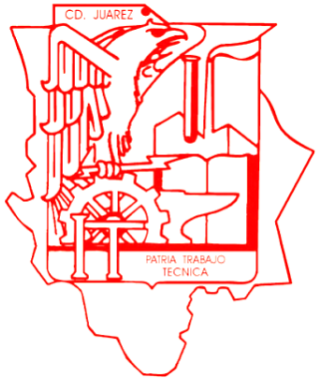
INSTITUTO TECNOLÓGICO DE CD. JUÁREZ
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN



DISEÑO DE UN MODELO DE RIESGO PARA DETERMINAR LA SUPERVIVENCIA DE PACIENTES CON CÁNCER CERVICO UTERINO.

PROYECTO DE INVESTIGACIÓN
QUE PRESENTA:

LUZ ELENA TERRAZAS MATA

DOCTORADO EN CIENCIAS DE LA INGENIERÍA.

CD. JUÁREZ, CHIH. **MAYO 2020**

2. PLANTEAMIENTO DEL PROBLEMA

2.1 Antecedentes

Cáncer Cérvico Uterino (CaCu), séptima neoplasia más frecuente en la población mundial y la cuarta más frecuente entre las mujeres, con un estimado de 528 mil nuevos casos diagnosticados anualmente, 85% de los cuales se registran en países en vías de desarrollo, que varía desde 42.7 en África Oriental, hasta 4.4 por 100, 000 mujeres en Asia Occidental. Con 266,000 defunciones anuales, se convierte en una importante causa de muerte por tumor maligno en países subdesarrollados. En la figura 2.1. (Globocan, 2002) se presenta la información antes mencionada (Globocan, 2002).

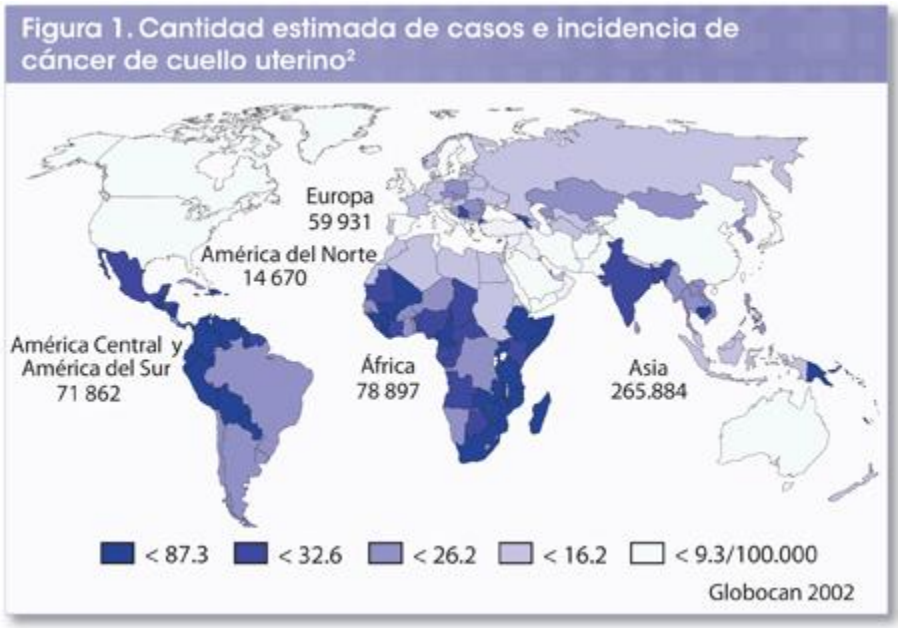


Figura 2.1 Cantidad Estimada de Casos e Incidencia de Cáncer de Cuello Uterino.

En América Latina el CaCu ocupa el segundo lugar como la neoplasia más común con 68,818 casos por año. En esta región la incidencia es de 21.2 casos por 100,000 mujeres alcanzando valores a 30 en países como Perú, Paraguay, Guyana, Bolivia, Honduras, Venezuela, Nicaragua y Surinam. La mortalidad es de 8.7 defunciones por 100,000 mujeres. El 75% de las 28,565 defunciones anuales por esta causa ocurren en los siguientes países: Brasil, México, Colombia, Perú, Venezuela y Argentina. Sin embargo la mortalidad es más alta en Guyana (21.9), Bolivia (21.0) y Nicaragua (18.3).

Para México, el cáncer de cuello uterino es la segunda causa de muerte por cáncer en la mujer. Anualmente se estima una ocurrencia de 13, 960 casos con una incidencia de 23.3 casos por 100,000 mujeres. En el año 2013, en el grupo específico de mujeres de 25 años y más, se registraron 3,771 defunciones en mujeres con una tasa de 11.3 defunciones por 100,000 mujeres. Las entidades con mayor mortalidad por cáncer de cuello uterino son Morelos (18.6), Chiapas (17.2) y Veracruz (16.4).

En ese mismo año ocurrieron 269,332 defunciones en mujeres, de las cuales los tumores malignos representan el 13.8% (37,361) de esas muertes. En la figura 2.2 se presenta la tasa de morbilidad hospitalaria por CaCu según entidad federativa en 2009 (INEGI, 2009).

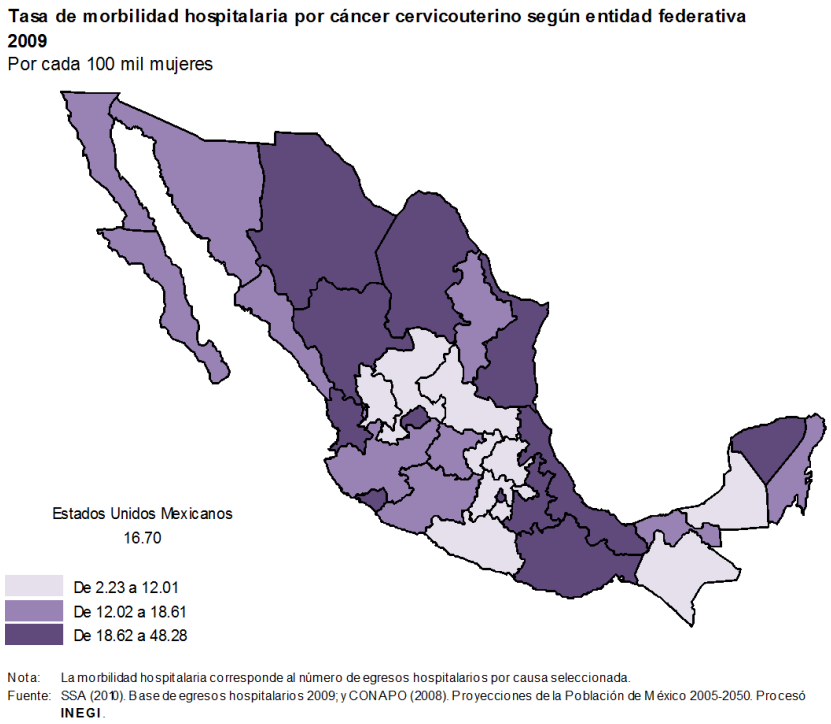


Figura 2.2 Tasa de Morbilidad Hospitalaria por CaCu Según Entidad Federativa 2009.

Dentro de las neoplasias con mayor número de defunciones en mujeres, el cáncer de mama y el cuello uterino ocasionaron en conjunto el 25% de todas las defunciones por cáncer en mujeres. Una de cada 10 muertes por cáncer en mujeres mexicanas es debida a cáncer de cuello uterino.

Dentro del período de 1990 al 2000 fueron reportadas un total de 48,716 defunciones por esta enfermedad lo cual representa un promedio de 12 mujeres fallecidas cada 24 horas, con un crecimiento anual de 0.76% (Tovar-Guzmán, 2008).

Es importante resaltar que en Chihuahua, los tumores malignos presentaron una alta incidencia y es la 4ta. causa de muerte con un total de 2,338 decesos, lo que representa el 10.5% de los fallecimientos del Estado de Chihuahua con una tasa de 68.93 x 100,000 habitantes. En 2009 en Cd. Juárez, de los 725 decesos por tumores malignos entre hombres y mujeres, el 9.25% corresponde al CaCu. Secretaría de Salud, (2009).

2.2 Planteamiento del Problema

La alta incidencia de decesos por CaCu justifica una investigación usando herramientas estadísticas que permitan hacer inferencias con mayor confianza en el análisis de factores significantes; es por esto que en este estudio se propone identificar y evaluar las causas significativas, así como los factores que intervienen en la incidencia del cáncer de cuello uterino en Cd. Juárez Chihuahua y desarrollar un modelo de riesgo para determinar la supervivencia de pacientes con este padecimiento.

2.3 Supuestos

La información proporcionada por médicos y centros de salud, en los expedientes de pacientes de CaCu es confiable.

Las respuestas a las encuestas usadas en este estudio están dadas con toda honestidad y veracidad de los encuestados.

2.4. Preguntas de Investigación

¿Cuáles son las características de mayor significancia que influyen en la supervivencia de las pacientes con CaCu?

¿Cuál es el comportamiento de los tiempos de supervivencia de las pacientes con CaCu, en relación a sus características significantes?

¿Cuáles son las características y la función de riesgo base en el modelo de riesgo proporcional que permita predecir la supervivencia de las pacientes con CaCu?

2.5 Hipótesis

H_1 Existen características generales y específicas propias de las pacientes con CaCu que influyen de manera significante en la supervivencia de las pacientes con CaCu.

H_2 Los tiempos de supervivencia de las pacientes con CaCu tienen una función de densidad de probabilidad relacionada con sus características específicas.

H_3 El modelo de riesgo proporcional será capaz de predecir los tiempos de supervivencia de las pacientes con CaCu.

2.6. Objetivos

2.6.1 Objetivo General.

Identificar los factores significantes que influyen en la supervivencia de las pacientes con CaCu.

2.6.1.1 Objetivos Específicos.

1. Determinar los factores significantes que influyen en la supervivencia de las pacientes con CaCu.
2. Determinar los tiempos de supervivencia de las pacientes con CaCu.

2.7 Justificación

La importancia de este estudio radica en que se identifican aquellas variables que pueden ayudar a prevenir y disminuir la incidencia de muertes por causa del CaCu e identificar la supervivencia en mujeres que padecen esta enfermedad en Cd. Juárez, Chihuahua. Y que además sirva como referencia para la zona norte del país, ya que de acuerdo a la revisión de literatura, no hay un estudio similar.

2.8 Delimitaciones

Este estudio se llevará a cabo en las clínicas del sector público en Cd. Juárez Chih., donde se aplica tratamiento para este tipo de cáncer. Los resultados de esta investigación son aplicables a la situación prevaleciente en esta ciudad durante la duración del estudio.

3. MARCO TEÓRICO

Es este capítulo se presentan una serie de conceptos que funcionan como supuestos teóricos que posibilitan la investigación expuesta.

3.1. Aspecto Legal

La Organización Mundial de la Salud (OMS), establece como objetivos generales para el control del cáncer:(López-Calviño, 2012).

a. Reducir su morbimortalidad
b. Aumentar las tasas de curación
c. Mejorar la calidad de vida, tanto de los enfermos que sobrevive, como de los que fallecerán.
d. Reducir la carga socioeconómica y psicológica que supone esta enfermedad.

En México existe el programa Nacional de Cáncer Cérvicouterino y llama la atención que es ofrecido anualmente a mujeres de 25 a 65 años y forma parte integral de los servicios de salud, sin embargo, actualmente la infraestructura hospitalaria en países en vías de desarrollo solo es suficiente para el análisis de una mujer cada cinco años (Tovar-Guzmán, 2008).

Otro rasgo respecto al marco legal de este padecimiento; en investigación realizada en la ciudad de Chihuahua en 2004 hace mención de los cuatro apartado de la Norma Oficial Mexicana- 014-SSA2-1994.

El primero de ellos corresponde a la identificación de la unidad donde se realiza la prueba; el segundo la identificación del paciente que solicita el examen; en el tercer apartado se encuentran los antecedentes que pueden predisponer a la

mujer a desarrollar CaCu como la situación gineco obstétrica, características del cérvix y factores de riesgo; y en el cuarto apartado se notifica el resultado de la citología cervical dado por el patólogo o citotecnólogo (Salas- Urrutia, 2004).

Dentro de la Ley Estatal de Salud del Estado de Chihuahua, en el historial de reformas en el Decreto No. 964-07 II P.O fecha de aprobación 2007.06.21; entrada en vigor en 2007.08.05 observa que: Se adiciona este capítulo específico debido a la alta incidencia de casos de cáncer mamario y cérvico uterino en la población femenina residente del Estado de Chihuahua. Se adiciona al Título Octavo de la Ley de Salud del Estado, un capítulo III Bis denominado Del cáncer mamario y del cáncer Cérvico-Uterino, con los artículos 123 Bis y 123 Ter.(www.congresochihuahua.gob.mx 2017).

3.2 Cáncer Cévico Uterino

Los cánceres cervicales inician en las células de la superficie del cuello uterino. Existen dos tipos de células en dicha superficie: escamosas y columnares. La mayoría de los cánceres de cuello uterino provienen de células escamosas.

El desarrollo del cáncer cervical generalmente es muy lento y comienza como una afección precancerosa llamada displasia. Esta afección se puede detectar por medio de una citología vaginal y es 100% curable. Pueden pasar años para que los cambios se conviertan en cáncer cervical. La mayoría de las mujeres a quienes se les diagnostica cáncer cervical, en la actualidad no se han sometido a citologías vaginales regulares o no han tenido un seguimiento por resultados anormales a estas.

Casi todos los cánceres cervicales son causados por el virus del papiloma humano (VPH), un virus común que se disemina a través de relaciones sexuales. Existen muchos tipos diferentes (cepas) de VPH, y algunas cepas llevan a cáncer cervical. Otras cepas pueden causar verrugas genitales, mientras que otras no causan ningún problema en absoluto.

3.3 Factores de Riesgo

En lo que respecta a los factores de riesgo para CaCu entre 1980 y 1983, las tasas de incidencia de CaCu son semejantes entre los países de Latinoamérica, esto lo presenta Sierra y Barrantes(1988), hasta 1988 en estudio realizado en Costa Rica en donde demuestran que existen diferencias entre municipios en la atención de salud y el nivel socio económico que puede influir en la frecuencia de cáncer invasor, además de una significativa correlación de este tumor con las tasas de gonorrea por municipio, sugiere que el ambiente, o los factores que favorecen esta enfermedad, posiblemente afecten también la etiología del cáncer de cuello uterino, hecho que ya ha sido señalado por otros autores.

Entre sus conclusiones recomiendan estudios epidemiológicos más amplios para saber hasta qué punto factores tales como la raza, la promiscuidad, el nivel socioeconómico, y la actividad sexual a edad temprana influyen en la incidencia y el comportamiento de éste cáncer en su país.

Rodríguez-Salvá (1999) realizó un estudio en Cuba cuyo propósito es identificar los factores de riesgo del cáncer de cérvix en el municipio Cerro basado en que la distribución geográfica se relaciona con algunos factores del medio

ambiente y estilo de vida como, relaciones sexuales tempranas, multipariedad, higiene inadecuada, niveles socioeconómico y educacional bajos, uso prolongado de anticonceptivos orales, dieta, tabaquismo y comportamiento sexual inadecuado (promiscuidad) en ambos sexos. Además, al parecer existe asociación entre las formas invasivas del CaCu y los del VPH y del herpes simple tipo II.

De donde concluye que la edad no constituye por sí sola como factor de riesgo, pero, que si se estudian los carcinomas invasivos, entonces, a mayor edad, aumenta el riesgo. A menor nivel de escolaridad aumenta el riesgo de padecer CaCu siendo éste significativo en todos los grupos estudiados, sin embargo la relación no es lineal. Respecto a lo factores relacionados con la historia reproductiva, se infiere que cuando aumenta la edad de la menarquía y a menor edad en el primer parto se incrementa significativamente el riesgo de padecer cáncer, así también a mayor número de partos el riesgo aumenta.

De la misma manera, hay asociación significativa entre la edad de inicio de las relaciones sexuales en padecer la enfermedad y la cantidad de compañeros sexuales, el riesgo aumenta cuando han tenido 5 o más.

Usar anticonceptivos orales resultó un factor de riesgo mayor que quienes no lo usan y la realización del coito anal confiere un riesgo mayor que quienes lo hicieron una vez. El coito en el período menstrual eleva el riesgo 3.5 veces.

Respecto a las pacientes con el hábito de fumar se concluye que a mayor tiempo de fumar aumenta el riesgo y es 10.3 veces mayor en las mujeres que fuman o han fumado durante 20 años y más.

Flores-Luna (2000) utilizó las técnicas de análisis de supervivencia para medir el impacto de los factores que determinan una mayor supervivencia en las mujeres con CaCu de un hospital de ginecobstetricia de la ciudad de México. Dentro de los factores estudiados fueron: escolaridad, los antecedentes familiares de cáncer, número de cigarrillos al día, la edad de la menarca, la edad de inicio de la vida sexual, número de partos, el número de parejas sexuales, el grado de diferenciación, patrón clínico, tamaños del tumor, diseminación y tratamiento.

Rosell-Juarte (2004) menciona que la Organización Panamericana de Salud (OPS) y diferentes autores señalan la existencia de diversos factores de riesgo asociados a la aparición del cáncer de cuello uterino, entre las que destacan las enfermedades de transmisión sexual, entre ellas el VPH y el herpes virus tipo II, las relaciones sexuales y el parto precoz; otros autores plantean que el hábito de fumar cigarrillos y las condiciones socioeconómicas adversas, entre, otras predisponen a la neoplasia maligna cervical.

Algunos autores citados por Rosell-Juarte (2004), no encuentran asociación entre el uso continuado de los anticonceptivos orales y el cáncer de cuello uterino, en tanto que otros lo señalan como factor de riesgo desencadenante de esta afección.

El programa Nacional Cubano para el Diagnóstico Precoz del Cáncer cérvico uterino (2004) ha reportado que el inicio precoz de las relaciones sexuales y el parto antes de los veinte años, son causas predisponentes para esta afección. Igual observación se hizo para los sujetos que han tenido múltiples compañeros sexuales,

el tabaquismo y el VPH, los que constituyen en Cuba factores de riesgo para el desarrollo de la neoplasia cervical.

Los resultados de ese estudio dicen que el nivel primario de educación multiplica 5.5 veces la probabilidad de que ocurra el cáncer cervical, por otra parte, coincidiendo con los reportes de literatura donde el hábito de tener múltiples compañeros sexuales, multiplica 32 veces la probabilidad de adquirir una neoplasia maligna cervical, siendo uno de los factores más significativos encontrados en la investigación.

Fumar es un factor de riesgo desencadenante de múltiples dolencias. En el estudio, el cáncer de cuello uterino se halló tres veces más frecuente dentro del grupo de mujeres fumadoras que en el grupo de control.

Se encontraron resultados congruentes con otros autores citados por Rosell-Juarte (2004), referente a que el inicio temprano de la vida sexual incrementa la eventualidad.

Haber tenido el primer parto en la adolescencia, significó una probabilidad 4.5 veces mayor de desarrollar un cáncer cervical, en tanto que para las que tuvieron tres ó más partos, fue 10.4 veces, lo que reafirma también los reportes de otros autores.

En resumen, Rosell-Juarte (2004), concluye que los factores de riesgo asociados al cáncer de cuello uterino que predominaron fueron el nivel escolar bajo, tres o más compañeros sexuales, el tabaquismo, las relaciones sexuales el parto precoz y la multiparidad.

Según el *National Center Institute* (2014), los índices de mortalidad e incidencia del cáncer de cuello uterino varían según la situación socioeconómica y la ubicación geográfica, y los índices de realización de exámenes de detección son diferentes en los distintos grupos raciales, étnicos, socioeconómicos y geográficos.

Ortiz-Serrano (2004) presenta y describe los factores de riesgo para CaCu donde dice que de acuerdo a estudios realizados, el VPH es el principal factor asociado con CaCu y sus precursores. En la actualidad es considerado como virtualmente necesario pero no suficiente como agente causal.

Dentro de éstos también considera las características de la conducta sexual, y dentro de ellas, el número de compañeros sexuales, edad del primer coito, características del compañero sexual respecto su historia sexual y enfermedades de transmisión sexual.

Están también los factores relacionados con la reproducción, como son la paridad, edad del primer parto, partos vaginales y como punto aparte el tabaquismo, donde menciona que existe una relación directa entre el riesgo de lesiones preinvasoras y la duración e intensidad del hábito, pero que aún con todo esto reconocidos estudios no soportan claramente la asociación directa entre el tabaquismo y el CaCu.

Se cree que el uso prolongado de anticonceptivos orales, es otro factor que se asocia con el mayor riesgo de lesión intraepitelial. Sin embargo, es mucho mayor la probabilidad de un embarazo indeseado cuando estos se dejan de tomar, que el riesgo de aparición potencial de neoplasia cervical.

Así mismo están los factores psicosociales, como la condición socio económica donde hay datos controvertidos referente a la asociación que se da entre el cáncer de cuello uterino y condición económica social o educativa. Es aceptable que las mujeres de más bajo nivel social, cultural y educativo, así como las extranjeras tienen mayor incidencia de cáncer del cuello uterino.

Pero existen estudios que muestran a las universitarias como las que más lo sufren. Aparte de esta condición, uno de los aspectos que cada vez se identifican como críticos para el diagnóstico y tratamiento oportuno del cánceres el de la identidad personal que tienen las mujeres así como de su sexualidad, en la medida que esto determina la capacidad para tomar decisiones autónomas.

Hay que mencionar además los factores asociados con la calidad de la atención y el acceso a ella; ya que el cáncer de cuello uterino es una enfermedad previsible cuando su diagnóstico es oportuno y se realiza un tratamiento adecuado.En los países desarrollados donde hay amplia cobertura, el 80% de los casos de lesiones que afectan el cuello uterino son detectados como neoplasia intraepitelial, lo cual sugiere que existen elevados estándares de calidad, en los programas de detección temprana. De ahí lo importante de dichos programas en la disminución del cáncer de cuello uterino en países nórdicos como Canadá y Estados Unidos. Contrario a los países en desarrollo, donde dada la baja cobertura y los bajos estándares de calidad, los índices de mortalidad por cáncer de cuello uterino no han disminuido.

En México, por ejemplo, las deficiencias tienen relación con las bajas coberturas y calidad en el servicio, con los deficientes procedimientos para la

obtención de especímenes adecuados de citología exfoliativa, altos índices de diagnóstico falso negativo en centros de lectura citología ginecológica, agregando a las mujeres que acuden para detección tardíamente. Es importante mencionar que las coberturas más bajas de los programas se dan sobre todo en las áreas rurales.

Así mismo menciona que dentro de los factores la demanda inducida, que se refiere a la acción de organizar, incentivar y orientar a la población hacia la utilización de los servicios de protección específica y detección temprana además de la adhesión a los programas de control. Otro factor es la pertinencia técnica y científica, ésta es el grado en el cual los usuarios obtienen los servicios que requieren, de acuerdo con la evidencia científica, y sus efectos secundarios son menores que los beneficios potenciales. Dentro del conocimiento de los profesionales que intervienen en la norma técnica se destacan la actualización, la capacitación y la educación médica continuada.

También se agrega como factor la oportunidad referida a la posibilidad que tiene el usuario de obtener los servicios que requiere, sin que se presenten retrasos que pongan en riesgo su vida o su salud. Esta característica se relaciona con la organización de la oferta de servicios en relación con la demanda y con el nivel de coordinación institucional para gestionar el acceso a los servicios. Lamadrid (1998) muestra que las mujeres incrementan la utilización de la prueba en presencia de algún síntoma ginecológico como infección. Los tiempos de espera para la toma de la citología así como entrega de resultados impactan en la percepción de las usuarias frente a la calidad de algún servicio de salud. Es tan importante la

oportunidad como atributo cualitativo que da argumentos al cliente para recomendar o no el servicio, pasando por encima de otros, como la amabilidad y la misma pertinencia científica o profesional.

Salas-Urrutia (2004), en estudio realizado en el Hospital Central de Chihuahua, México sobre prevalencia de displasia y CaCu y factores asociados, obtiene que el principal factor de riesgo detectado fue las múltiples parejas sexuales, seguido de la vida sexual activa antes de los 18 años de edad, el tabaquismo y los antecedentes de infecciones de transmisión sexual.

En las investigaciones realizados hace 30 años, Rodriguez-Salvá (1999) recomienda estudiar los factores como raza, promiscuidad, nivel socio económico, actividad sexual temprana, tabaquismo, uso de anticonceptivos como causales de CaCu, sin embargo, en lo expuesto anteriormente por este autor, presume que puede haber asociación entre padecer el VPH y desarrollar CaCu, Rosell-Juarte (2004)menciona que este factor resulta ser el principal causal de esta displasia, dicho lo anterior se puede destacar que los diferentes autores coinciden en los factores de riesgo de CaCu.

3.4 Etapas del Cáncer

El siguiente aspecto trata sobre la clasificación por etapas del CaCu.

La clasificación por etapas (estadios) o estadificación es el proceso para determinar hasta dónde se ha propagado el cáncer. Se utiliza la información de los exámenes y pruebas de diagnóstico para determinar el tamaño del tumor, hasta qué punto éste ha invadido los tejidos en o alrededor del cuello uterino y la propagación

a los ganglios linfáticos u otros órganos distantes (metástasis). Este es un proceso importante porque la etapa del cáncer es el factor más relevante para seleccionar el plan de tratamiento adecuado (www.cancer.org. 2016).

La etapa de un cáncer no cambia con el paso del tiempo, incluso si el cáncer progresa. A un cáncer que regresa o se propaga se le sigue conociendo por la etapa que se le asignó cuando se encontró y diagnosticó inicialmente, sólo se agrega información sobre la extensión actual del cáncer. Una persona mantiene la misma etapa de diagnóstico, pero se agrega más información al diagnóstico para explicar la condición actual de la enfermedad.

Un sistema de estadificación o clasificación por etapas es un método que utilizan los especialistas en cáncer para resumir la extensión de la propagación de un cáncer. Los dos sistemas utilizados para clasificar por etapas (estadios) la mayoría de los tipos de cáncer de cuello uterino, son el sistema FIGO (*International Federation of Gynecology and Obstetrics*) por sus siglas en inglés y el sistema TNM (tumor-nódulo-metástasis) de la AJCC (*American Joint Committee on Cancer*) por sus siglas en inglés que son muy similares. Los ginecólogos y los oncólogos ginecológicos usan el sistema FIGO, pero el sistema TNM se incluye en este documento para proveerle información más completa. El sistema TNM clasifica el cáncer de cuello uterino tomando en cuenta tres factores: El tamaño o el alcance del cáncer (T), si el cáncer se ha propagado a los ganglios (nódulos) linfáticos (N), y si se ha propagado a partes distantes (M). El sistema FIGO usa la misma información. El sistema descrito a continuación es el sistema TNM más reciente, el

cual entró en vigor en enero de 2010. Cualquier diferencia entre el sistema TNM y el sistema FIGO se explica en el texto.

3.4.1 Sistema FIGO

Este sistema clasifica la enfermedad en etapas de 0 a IV. La clasificación por etapas se basa en los hallazgos clínicos en vez de los hallazgos de la cirugía. Esto significa que la extensión de la enfermedad se evalúa por medio del examen físico realizado por el doctor y algunas otras pruebas que se realizan en algunos casos, tales como la cistoscopia y la proctoscopia no se basa en los hallazgos durante la cirugía ni en los estudios por imágenes.

Cuando se hace una cirugía, podría verse si el cáncer se ha propagado más de lo que los médicos pensaban al principio. Esta nueva información podría cambiar el plan de tratamiento, pero no cambia la etapa de la paciente.

Respecto a la extensión del tumor la nomenclatura es la siguiente:

Tis: Las células cancerosas sólo se encuentran en la superficie del cuello uterino (la capa de las células que reviste el cuello uterino) sin crecer hacia (invadir) los tejidos más profundos. (Tis no está incluida en el sistema FIGO).

T1: Las células cancerosas han crecido desde la capa de la superficie del cuello uterino hasta los tejidos más profundos de éste. Además, el cáncer puede estar creciendo hacia el cuerpo del útero, aunque no ha crecido fuera del útero.

T1a:Existe una cantidad muy pequeña de cáncer que sólo se puede observar con un microscopio.

T1a1: El área de cáncer es de menos de 3 milímetros (alrededor de 1/8 de pulgada) de profundidad y de menos de 7 mm (alrededor de 1/4 de pulgada) de ancho.

T1a2: El área de invasión del cáncer es entre 3 mm y 5 mm (alrededor de 1/5 de pulgada) de profundidad y de menos de 7 mm (alrededor de 1/4 de pulgada) de ancho.

T1b: Esta etapa incluye los cánceres de etapa I que se pueden ver sin un microscopio. Esta etapa incluye también los tumores cancerosos que se pueden ver solamente con un microscopio y que se han propagado a más de 5 mm (alrededor de 1/5 de pulgada) de profundidad dentro del tejido conectivo del cuello uterino o que tienen más de 7 mm de ancho.

T1b1: El cáncer se puede ver, pero no tiene más de 4 cm (alrededor de 1 3/5 pulgadas).

T1b2: El cáncer se puede ver y tiene más de 4 centímetros.

T2: En esta etapa, el cáncer ha crecido más allá del cuello uterino y el útero, pero no se ha propagado a las paredes de la pelvis o a la parte inferior de la vagina. Puede que el cáncer haya crecido hacia la parte superior de la vagina.

T2a: El cáncer se ha propagado a los tejidos próximos al cuello uterino.

T2a1: El cáncer se puede ver, pero no tiene más de 4 cm (alrededor de 1 3/5 pulgadas).

T2a2: El cáncer se puede ver y tiene más de 4 centímetros.

T2b: El cáncer se ha propagado a los tejidos adyacentes al cuello uterino (el parametrio).

T3: El cáncer se ha propagado a la parte inferior de la vagina o a las paredes pélvicas. El cáncer puede estar bloqueando los uréteres (conductos que transportan la orina de los riñones a la vejiga).

T3a: El cáncer se ha propagado al tercio inferior de la vagina, pero no a las paredes pélvicas.

T3b: El cáncer ha crecido hacia las paredes de la pelvis y/o está bloqueando uno o ambos uréteres (a esto se le llama hidronefrosis).

T4: El cáncer se propagó a la vejiga o al recto o crece fuera de la pelvis.

Para la propagación a los ganglios linfáticos (N), la clasificación es como sigue:

NX: No se pueden evaluar los ganglios linfáticos cercanos.

N0: No se ha propagado a los ganglios linfáticos adyacentes.

N1: El cáncer se ha propagado a los ganglios linfáticos cercanos.

M0: El cáncer no se ha propagado a otros ganglios linfáticos, órganos o tejidos distantes.

M1: El cáncer se propagó a órganos distantes (como por ejemplo el hígado o los pulmones), a los ganglios linfáticos del pecho o el cuello, y/o al peritoneo (el tejido que cubre el interior del abdomen).

Para asignar una etapa a la enfermedad, se combina la información sobre el tumor, los ganglios linfáticos y cualquier propagación del cáncer. Este proceso se conoce como agrupación por etapas. Las etapas se describen usando el número 0 y con números romanos del I al IV. Algunas etapas se dividen en sub etapas indicadas por letras y números. Las etapas FIGO son las mismas que las etapas TNM, excepto que la clasificación por etapas FIGO no incluye los ganglios linfáticos hasta la etapa III. Además, la etapa 0 no existe en el sistema FIGO.

Etapa 0 (Tis, N0, M0): Las células cancerosas sólo se encuentran en las células de la superficie del cuello uterino (la capa de las células que reviste el cuello uterino) sin crecer hacia (invadir) los tejidos más profundos del cuello uterino. Esta etapa también se llama carcinoma in situ (CIS), y es parte de neoplasia intra epitelial cervical en grado 3 (CIN3). La etapa 0 no está incluida en el sistema de FIGO.

Etapa I (T1, N0, M0): En esta etapa el cáncer creció hacia (invadió) el cuello uterino, pero no fuera del útero. El cáncer no se ha propagado a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

Etapa IA (T1a, N0, M0): Esta es la forma más temprana de la etapa I. Hay una cantidad muy pequeña de cáncer que es visible solamente bajo el microscopio.

El cáncer no se ha propagado a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

Etapa IA1 (T1a1, N0, M0):El cáncer es de menos de 3 milímetros (alrededor de 1/8 de pulgada) de profundidad y de menos de 7 mm (alrededor de 1/4 pulgadas)

de ancho. El cáncer no se ha propagado a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

Etapa IA2 (T1a2, N0, M0): El cáncer es entre 3 mm y 5 mm (alrededor de 1/5 de pulgada) de profundidad y de menos de 7 mm (alrededor de 1/4 pulgadas) de ancho. El cáncer no se ha propagado a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

IB (T1b, N0, M0):Incluye los cánceres en etapa I que se pueden ver sin un microscopio, así como los cánceres que sólo se pueden ver con microscopio, si se han propagado a más de 5 mm (alrededor de 1/5 de pulgada) de profundidad dentro del tejido conectivo del cuello uterino o tienen más de 7 mm de ancho. Estos cánceres no se han propagado a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

IB1 (T1b1, N0, M0): El cáncer se puede ver, pero no tiene más de 4 cm (alrededor de 1 3/5 pulgadas). El cáncer no se propagó a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

IB2 (T1b2, N0, M0):El cáncer se puede ver y mide más de 4cm. No se ha propagado a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

Etapa II (T2, N0, M0): En esta etapa, el cáncer ha crecido más allá del cuello uterino y el útero, pero no se ha propagado a las paredes de la pelvis o a la parte inferior de la vagina.

Etapa IIA (T2a, N0, M0): El cáncer se ha propagado a los tejidos próximos al cuello uterino (parametrio). Puede que el cáncer haya crecido hacia la parte superior

de la vagina. El cáncer no se propagó a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

IIA1 (T2a1, N0, M0): El cáncer se puede ver, pero no tiene más de 4 cm (alrededor de 1 3/5 pulgadas). El cáncer no se propagó a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

IIA2 (T2a2, N0, M0): El cáncer se puede ver y tiene más de 4 centímetros.

IIB (T2b, N0, M0): El cáncer se ha propagado a los tejidos adyacentes al cuello uterino (el parametrio).

Etapa III (T3, N0, M0):El cáncer se ha propagado a la parte inferior de la vagina o a las paredes pélvicas. El cáncer puede estar bloqueando los uréteres (conductos que transportan la orina de los riñones a la vejiga). El cáncer no se propagó a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

IIIA (T3a, N0, M0): El cáncer se ha propagado al tercio inferior de la vagina, pero no a las paredes pélvicas. El cáncer no se propagó a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

IIIB (T3b, N0, M0; O T1-3, N1, M0): El cáncer ha crecido hacia las paredes de la pelvis y/o ha bloqueado uno o ambos uréteres (una afección llamada hidronefrosis). O el cáncer se propagó a los ganglios linfáticos en la pelvis (N1), pero no a sitios distantes (M0). El tumor puede ser de cualquier tamaño y pudo haberse propagado a la parte inferior de la vagina o a las paredes pélvicas (T1a T3).

Etapa IV: Ésta es la etapa más avanzada del cáncer de cuello uterino. El cáncer se ha propagado a órganos adyacentes o a otras partes del cuerpo.

IVA (T4, N0, M0): El cáncer se propagó a la vejiga o al recto, que son órganos cercanos al cuello uterino (T4). El cáncer no se propagó a los ganglios linfáticos cercanos (N0) ni a sitios distantes (M0).

IVB (cualquier T, cualquier N, M1): El cáncer se propagó a órganos distantes más allá del área pélvica, tales como los pulmones o el hígado.

3.5 Grupos de Riesgo

Por lo que se refiere a los grupos de riesgo, en publicación de artículo de la web por *Centers for Disease Control and Prevention*, en abril 2013, en Estados Unidos, en donde muestra resultados de estudio se usó para su investigación dos sistemas federales para la vigilancia del cáncer el Programa Nacional de Registros del Cánce ry el Programa de Vigilancia, Epidemiología y Resultados Finales para describir las tasas del cáncer de cuello uterino en las mujeres menores de 40 años arroja estos resultados, la tabla 3.1 resume los datos de este estudio .

Tabla 3.1 Tasas del Cáncer de Cuello Uterino en las Mujeres Menores de 40 Años.*CDCP*(2013)

Edad	Porcentaje
30-39	78
20-29	21

>a 20	1

Dr. Sanffilipo et al. (2007), concluye que en México, este tipo de cáncer es la primera causa de muerte por neoplasias malignas entre las mujeres de 25 a 64 años de edad.

Aunque el cáncer de cuello uterino es muy poco frecuente en las mujeres jóvenes, las pruebas de Papanicoláu anormales son comunes en estas mujeres debido a una infección frecuente de transmisión sexual, el VPH. Una prueba de Papanicoláu anormal puede llevar a que se tomen procedimientos adicionales que podrían causar daños y llevar también a un tratamiento innecesario. La mayoría de las anormalidades en las mujeres jóvenes mejoran sin tratamiento. Debido a que el cáncer de cuello uterino es poco frecuente en las mujeres jóvenes, los médicos deben seguir las recomendaciones de comenzar a realizar las pruebas de detección de dicho cáncer a los 21 años de edad.

Sierra (1988) muestra un resumen por grupos de edad de pacientes con CaCu en estudio realizado en Costa Rica y en la tabla 3.3 un resumen de la distribución de pacientes que padecen esta neoplasia según grupo etario en investigación realizada en Isla de la Juventud de Cuba.

Tabla 3.2 Incidencia de Cáncer de Cuello Uterino por Grupos de Edad. Costa Rica 1980-1983.

		Cáncer de cuello uterino		
	In situ		**Invasor**	
Grupo de Edad	**Casos (no.)**	**Tasa[a]**	**Casos (no.)**	**Tasa[a]**
15-24	123	11.7	12	1.1

25-34	642	86.3	136	18.3
35-44	464	104.3	203	45.4
45-54	154	48.7	247	77.8
55-64	55	25.7	211	98.7
65-74	42	33.4	155	122.6
>a 75	12	19.9	82	136.0
Desconocida	5	--	6	--
Todas las edades	1497	36.3[b]	1052	33.6[b]

[a] Por 100 000 mujeres

[b] Tasa de incidencia ajustada por edad con la población mundial

Tabla 3.3 Distribución de Pacientes con CaCu Según Grupo Etario. Isla de la Juventud. 2003-2009.

Grupos etarios (años)	No.	%
<= 25	15	9.74
26-33	23	14.94
34-41	34	22.08
42-49	41	26.62
50-57	21	13.64
58-65	12	7.79
66-73	4	2.60
>= 74 años	4	2.60
Total	154	100

En las investigaciones realizadas a partir de 1980 a la fecha respecto al CaCu, los diferentes autores Sierra (1988), Sanffilipo (2007), Dávila (2010) y CDCP (2013), concluyen que dentro de las características de la población que lo padecen el grupo de riesgo está entre mujeres que van de los 20 a 65 años de edad, como lo muestra Dávila (2010) en la tabla 1.3. A continuación se presentan resultados de investigaciones donde verifica esta incidencia, cabe mencionar que en algunos países de latino América este parámetro se abre desde los 15 años de edad. En la

tabla 3.4 se puede observar que incluye en el primer rango de edad el período que va de los 0 a 19 años, en la incidencia de lesiones cancerígenas de cuello uterino distribuido por edad.

Tabla 3.4 Incidencia de Lesiones Malignas y Premalignas del Cuello Uterino en el Área Metropolitana de Bucaramanga Cuba, 2000-2001.

Edad	0-19	20-24	25-29	30-34	35-39	40-44	45-49
Incidencia Por 100,000 hab.	14.81	238.7	394.74	492.64	737.19	770.19	836.98

Edad	50-54	55-59	60-64	65-69	70-74	75-79	80+
Incidencia Por 100,000 hab.	549.62	397.52	336.96	263.42	167.45	115.55	131.47

3.6 Supervivencia

En el portal de Significados (2017) se indica que Supervivencia es la acción y efecto de sobrevivir**.** Mientras que Fernández (2001) dice que la supervivencia es una medida de tiempo a una respuesta, fallo, muerte, recaída o desarrollo de una determinada enfermedad o evento. El término supervivencia se debe a que en las primeras aplicaciones de este método de análisis se utilizaba como evento la muerte de un paciente.

En PLATEA (2017) se indica que la supervivencia es la probabilidad que tienen al nacer los individuos de una población de alcanzar una determinada edad. La probabilidad decrece desde 1 para los individuos nacidos vivos hasta hacerse 0 a la edad máxima de la especie.

Al representar gráficamente el valor de supervivencia frente al tiempo (edad que alcanza) se obtiene la curva de supervivencia para esa población representada en la figura 3.1. En general, las curvas de supervivencia se ajustan, a tres tipos:

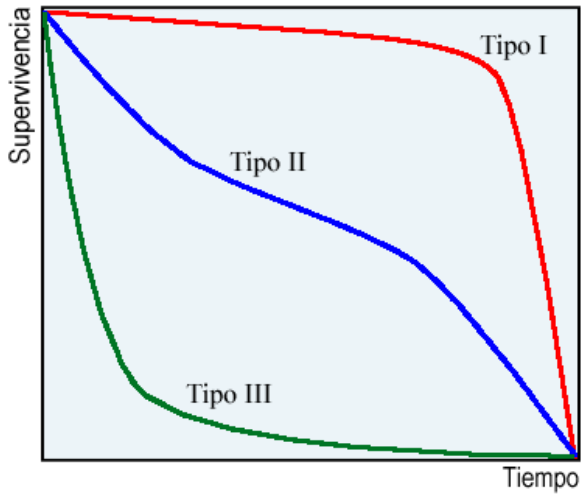


Figura 3.1 Curvas de Supervivencia

Tipo I. Las curvas tipo I o convexas caracterizan a las especies con baja tasa de mortalidad hasta alcanzar una cierta edad en que aumenta rápidamente. Tal es el caso de la mayor parte de los grandes mamíferos, incluido el hombre.

Tipo II. Si la tasa de mortalidad varía poco con la edad, como ocurre en la mayoría de las aves, la curva tiene la forma de una diagonal descendente, normalmente con forma sigmoidea si el número de individuos que muere en cada tramo de edad es más o menos constante.

Tipo III. Las especies r-estrategas sufren una elevada mortalidad en las primeras etapas de la vida, larvaria o juvenil, teniendo luego una mayor probabilidad de supervivencia. La curva muestra un pronunciado descenso inicial seguido de una fase más estable.

3.6.1Análisis de Supervivencia.

El análisis de supervivencia es una herramienta imprescindible en la investigación clínica y epidemiológica. Aunque el origen del nombre se debe a la construcción de tablas de vida en el siglo XVII, sus aplicaciones se extienden al estudio del tiempo transcurrido entre dos eventos de interés, como podría ser la aparición de un acontecimiento adverso tras una intervención terapéutica ó el tiempo transcurrido entre el inicio de una infección y su diagnóstico (Gómez y Cobo, 2004).

Según el concepto de Fernández (2001) el análisis de supervivencia es un área estadística en la que la variable respuesta es el tiempo que transcurre entre un evento inicial (que determina la inclusión del individuo en el estudio) y un evento final (genéricamente llamado falla) que ocurre cuando el individuo presenta la característica para terminar el estudio (muerte, alta de la enfermedad, etc.). En este tipo de estudios de cohorte puede ocurrir que algún individuo lo abandone antes de que le ocurra el evento de interés, registrándose sólo información parcial (censura) sobre la variable de interés (tiempo de falla). El objetivo principal del análisis de supervivencia es incorporar esta información parcial que proporcionan los individuos censurados mediante métodos desarrollados para ese fin. En otras áreas estadísticas esta información parcial es ignorada (datos faltantes); esta manera de proceder es incorrecta, ya que se desconoce el aporte parcial de estos individuos al estudio, lo que es contrario a la filosofía estadística de incorporar toda la información disponible dentro del análisis.

Los métodos estadísticos en el análisis de supervivencia son similares a los que se utilizan en otras áreas donde no se presentan datos censurados, por ejemplo: análisis descriptivo, comparación de poblaciones, modelos tipo regresión, etcétera.

Gómez y Cobo (2004) publican que en el proceso de obtención de pruebas empíricas en las que basar las decisiones clínicas, el análisis de supervivencia aborda las tres situaciones clásicas que resuelve la inferencia estadística.

Estudio univariante: Descripción y resumen de los tiempos de vida, usualmente a partir de la estimación e interpretación de las funciones de supervivencia y de riesgo. Estos análisis permiten a su vez predecir el comportamiento futuro de pacientes de características similares. Por ejemplo, habiendo observado que 9 de 10 pacientes intervenidos de una neoplasia gástrica avanzada la superaban los 6 meses de vida, ¿qué se sabe lo que sucederá en los fututos pacientes?

Estudio bivariante: Comparación del patrón de supervivencia de dos poblaciones:

¿Es cierto que los miembros de cierto grupo viven más tiempo y tienen mayor esperanza de vida o es simplemente un resultado casual observado en esta muestra, pero que no se repetirá en el futuro?

Estudio multivariante: Construcción de un modelo que, teniendo en cuenta las características de los pacientes, ayude a predecir su tiempo de vida y a

seleccionar los factores de riesgo que contribuyen a esta predicción una vez tenida en cuenta la influencia de otros factores más relevantes.

A continuación se presentan, en primer lugar, dos características de las variables del tipo tiempo entre dos eventos: la asimetría y la censura, ya que requieren un análisis específico.

3.6.1.1 Asimetría de la variable tiempo

A diferencia de otras variables, el tiempo solo se desplaza en una dirección: el colesterol puede aumentar o disminuir, pero el tiempo nunca retrocede. Las distribuciones que gobiernan las variables del tipo tiempo presentan unas formas diferentes a las del resto de variables biológicas, que aconsejan planteamientos alternativos de análisis. La media, la desviación típica y la distribución normal no resumen bien el tiempo de supervivencia.

3.6.1.2 Censura

La segunda característica que presenta la variable tiempo, es que requiere un largo período para ser observada en su totalidad. La censura se produce cuando el tiempo de supervivencia supera el período de seguimiento.

3.6.1.3 Normalidad de la variable tiempo

Gran parte de los análisis estadísticos se basan en la hipótesis de que la variable tiempo se distribuye normalmente. En este caso, la inferencia se reduce a

la estimación de los dos parámetros que rigen esta distribución, la media y la varianza.

Como este método no se adapta suficientemente bien a muestras formadas por tiempos de vida, hay que recurrir a otras distribuciones o a otras técnicas. Los procedimientos no paramétricos, es decir aquellos que no imponen a los datos una distribución concreta, son la alternativa más usada en estudios de supervivencia. Además, las técnicas no paramétricas permiten tener en cuenta el carácter secuencial de los datos, de forma que cada individuo sólo contribuye al estudio mientras no aparece la censura.

En resumen, que la variable de interés es el tiempo y que éste se mida secuencialmente tiene como consecuencia una distribución asimétrica y la presencia de censura. Estas circunstancias desaconsejan el uso de la distribución normal, que también caracterizan la media y la desviación típica.

3.6.2 Función de Supervivencia

Para resumir y analizar tiempos de vida se usa la denominada función de supervivencia. Esta función puede calcularse en cada instante y da cuenta de la probabilidad de un paciente sobreviva el mismo.

La función de supervivencia indica la probabilidad de que un paciente supere cierto tiempo de vida.

3.6.3 Función de Riesgo

La función de riesgo, del inglés (*hazard function*), *HR* representa la probabilidad, por unidad de tiempo, de presentar el episodio en el lapso subsiguiente, condicionado a que éste no se había presentado antes.

Cuando la función de riesgo es constante se denomina tasa de riesgo.

La forma de la función de riesgo dependerá del fenómeno estudiado. Un riesgo creciente corresponde a una población que envejece y se encuentra, por ejemplo en el análisis de tiempo de vida de pacientes con leucemia que no responden al tratamiento. Un riesgo decreciente corresponde a poblaciones en las que los individuos se fortalecen con el paso del tiempo, por ejemplo, después de una operación quirúrgica. La función de riesgo puede adoptar otras formas, como la de la bañera o de la joroba. La forma de bañera es apropiada como modelo para poblaciones que se controlan desde el nacimiento: al principio están las enfermedades infantiles, después se estabiliza y luego sigue un proceso creciente.

Posibles procedimientos para estimar las curvas de supervivencia al CaCu.

En el análisis de supervivencia, el manejo de los datos puede ser realizado utilizando técnicas paramétricas y no paramétricas (Fernández, 2014).

- Paramétricas: (las más frecuentes)
    a. Distribución Exponencial.
    b. Distribución de Weibull.
    c. Distribución Lognormal.

d. Distribución Gamma

e. Distribución Beta

- No paramétricas:

a. Kaplan-Meier.

b. Logrank.

c. Regresión de Cox.

3.6.4 Modelos de Estimación Paramétrica de la Supervivencia

Pérez-Hoyos, (2017) *(bvsde.ops-oms.org.)* publica sobre la estimación paramétrica de la supervivencia y en él se asume que la variable aleatoria T , que cuantifica el tiempo desde el origen hasta un evento, sigue una determinada distribución, y que se pueden cuantificar las probabilidades de sobrevivir a un determinado tiempo *t*. Habitualmente los tiempos de supervivencia siguen una distribución asimétrica con lo cual se deberían utilizar distribuciones de esta forma. Una de las principales distribuciones de probabilidad en el análisis de supervivencia, es la distribución Weibull, introducida en 1951 por Wilson Weibull en el contexto de la confiabilidad industrial. Esta distribución juega con el análisis de supervivencia, el mismo papel que la distribución normal en la estadística tradicional. Más sencilla que la distribución Weibull, es la distribución exponencial que tal como se verá es una Weibull especial. Otro tipo de distribuciones que se utilizan son la log-normal, Gamma etc.

3.6.4.1 <u>Distribución Exponencial</u>

El modelo más simple de análisis de supervivencia es el modelo exponencial. Este modelo asume que la tasa de riesgo *h(t)* es constante.

$$h(t) = \lambda = e^{-\beta} \tag{3.1}$$

La reparametrización de λ se justifica en la comparación de curvas con modelos paramétricos. Si la tasa de riesgo es constante, la tasa de riesgo acumulada se puede expresar como:

$$H(t) = \Lambda(t) = te^{-\beta} \tag{3.2}$$

Dado que la supervivencia se puede obtener a partir de la tasa de riesgo acumulada, bajo un modelo exponencial ésta será.

$$S(t) = e^{-H(t)} = e^{-te^{-\beta}} \tag{3.3}$$

3.6.4.2 Distribución Weibull

La tasa de riesgo no tiene por qué ser constante a lo largo del tiempo. El modelo Weibull añade un parámetro de escala σ al modelo exponencial para que la tasa de riesgo no sea constante.

La expresión de la tasa acumulada viene dada por la misma que la tasa exponencial pero corregida por un factor de escala.

$$\Lambda(t) = H(t) = te^{\beta^{1/\alpha}} \tag{3.4}$$

De este modo la tasa de riesgo se puede obtener derivando como

$$\lambda(t) = h(t) = \frac{\Lambda(t)}{\sigma(t)} = \frac{te^{-\beta^{1/\alpha}}}{\sigma(t)} \tag{3.5}$$

A partir de la tasa de riesgo acumulada se obtiene la siguiente expresión para la supervivencia.

$$S(t) = e^{-H(t)} = -t_p e^{-\beta^{-1/\alpha}} \tag{3.6}$$

Tal como en el caso del modelo exponencial, con el modelo Weibull se puede obtener una expresión para obtener los percentiles de la supervivencia, sin más que despejar el tiempo t_p de la ecuación:

$$S(t_p) = 1 - p = e^{(-H(t))} = e^{-\beta^{-1/\alpha}} \tag{3.7}$$

$$(\log(1-p)^{\sigma} = t_p(-\beta) \tag{3.8}$$

$$t_p = e^{(\beta)log\left(\frac{1}{1-p}\right)^{\sigma} \quad log} \tag{3.9}$$

3.6.4.3 Distribución Log-normal

Tal como se ha comentado, el tiempo T desde el origen hasta el evento no tiene una distribución simétrica y su rango de variación es entre 0 y el máximo

tiempo posible. Por ello la distribución normal no es apropiada para describir el fenómeno al poder tomar valores negativos. Puede suceder que aunque los tiempos T no sean normales, el logaritmo de los tiempos sí siga una distribución normal. Así se define como modelo log-normal aquel en el que los tiempos de supervivencia sigan una distribución normal con media β y desviación típica σ.

En este caso los valores de los tiempos de los casos perdidos en el origen no pueden ser asumidos ya que $log\ t$ no está definido cuando $t=0$.

La supervivencia se ha definido como $1 - F(t)$ donde $F(t)$ es la función de la variable aleatoria T , que no es más que la probabilidad acumulada de que el tiempo de supervivencia sea menor o igual que t. Esta probabilidad es lo mismo que obtener la probabilidad de que la distribución normal de $Log(T)$ sea menor que $\log(t)$, es decir

$$F(t) = \Phi\left(\frac{Log\ t - \beta}{\sigma}\right) \tag{3.10}$$

Donde Φ es la distribución normal estándar $N(0,1)$.

Así la supervivencia se obtiene como:

$$S(t) = 1 - F(t) = 1 - \Phi\left(\frac{Log\ t - \beta}{\sigma}\right) \tag{3.11}$$

A partir de la relación que existe entre la supervivencia y la tasa de riesgo instantánea se puede obtener esta como el cociente entre la densidad de la distribución y la supervivencia $h(t) = \frac{f(t)}{S(t)}$. Así la tasa de riesgo es:

$$h(t) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \frac{e^{\frac{-logt^{(-\beta/\sigma)^2}}{2}}}{1 - \Phi(\frac{logt - \beta}{\sigma})} \quad (3.12)$$

Este tipo de distribución es apropiada para períodos de incubación de enfermedades infecciosas y crónicas como por ejemplo el SIDA.

La tasa de riesgo acumulada a partir de la relación con la supervivencia acumulada $H(t) = -\log\big(S(t)\big)$.

$$H(t) = -\log\big(S(t)\big) = -\log(1 - \Phi(\frac{logt - \beta}{\sigma}) \quad (3.13)$$

3.6.4.4 Distribución Beta Generalizada.

Gonzáles (2014), en estudio de sobrevivencia publica que la distribución beta representa una familia de distribuciones de probabilidad continuas con soporte en el intervalo (0,1). La densidad beta es caracterizada por dos parámetros positivos, indicados generalmente por α y β ó u y v, que son parámetros de localización y de escala. La distribución beta ha sido aplicada para modelar el comportamiento de variables aleatorias limitadas a intervalos de amplitud finita, en una gran variedad de áreas.

Su densidad es:

$$f(x) = \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)}(x)^{u-1}(1-x)^{v-1}; x \in (0,1) \quad (3.14)$$

Que puede tener variados comportamientos dependiendo de los valores de los parámetros, desde comportamientos simétricos hasta totalmente asimétricos.

La distribución beta generalizada nació de manera natural para dar mayor flexibilidad al soporte acotado, donde su función de densidad es definida como:

$$f(x) = \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)}(x)^{u-1}(1-x)^{v-1}; x \in (0,1) \quad (3.15)$$

Donde, del mismo modo que en el modelo estándar, los parámetros u y v son positivos. La distribución Beta estándar es ahora una situación particular de la distribución Beta Generalizada, cuando $(a,b) = (0,1)$.

Si X es una variable aleatoria con distribución Beta Generalizada, entonces la notación será $X \sim BG_a(u,v)$ o equivalentemente $X \sim BG_{(u,v,a,b)}$.

3.6.4.5 La Distribución Gamma

Este modelo es una generalización del modelo Exponencial ya que, en ocasiones, se utiliza para modelar variables que describen el tiempo hasta que se produce p veces un determinado suceso, su función de densidad es

$$f(x) = \begin{cases} \frac{1}{\beta^p \Gamma(\alpha)} e^{-\frac{x}{\beta}} x^{\beta-1} & si\ x > 0 \\ 0 & si\ x \leq 0 \end{cases} \quad (3.16)$$

Como vemos, este modelo depende de dos parámetros positivos: α y β. La función $\Gamma(\alpha)$ es la denominada Función Gamma de Euler que representa la siguiente integral:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}\, e^{-x} dx \tag{3.17}$$

Que verifica $\Gamma(p+1) = p\Gamma(p)$, con lo que, si p es un número entero positivo, $\Gamma(\alpha+1) = \alpha$¡

Dentro de las propiedades de la distribución Gamma están:

1 Su esperanza matemática es $\alpha\beta$

2. Su varianza es $\alpha\beta^2$

3. La distribución Gamma,$\beta, \alpha = 1)$ es una distribución Exponencial de parámetro β. Es decir, el modelo Exponencial es un caso particular de la Gamma con $\alpha = 1$.

4. Dadas dos variables aleatorias con distribución Gamma y parámetro α común.

$$X \sim G(\alpha, p_1)\; y\; Y \sim G(\alpha, p_2)$$

Se cumplirá que la suma también sigue una distribución Gamma

$$X + Y \sim G(\alpha, p_1 + p_2).$$

Una consecuencia inmediata de esta propiedad es que, si tenemos k variables aleatorias con distribución exponencial de parámetro α (común) e independientes, la suma de todas ellas seguirá una distribución $G(\alpha, k)$ (www.ub.edu,2017)

Los modelos estadísticos más utilizados para el análisis de la supervivencia son los no paramétricos y dentro de ellos están el Método Kaplan-Meier, LogRank y Regresión de Cox.

3.6.4.6 Método Kaplan-Meier

El método Kaplan-Meier calcula la supervivencia cada vez que un paciente muere. Y da proporciones exactas de supervivencia debido a que utiliza tiempos de supervivencia precisos (www.fisterra.com2017).

Este método se conoce como el límite-producto. Caracterizándose porque la proporción acumulada que sobrevive se calcula para el tiempo de supervivencia individual de cada paciente y no se agrupan los tiempos de supervivencia en intervalos.

La probabilidad de ser censurado debe ser independiente del efecto de interés. Es decir, no puede aplicarse el método de Kaplán-Meier.

3.6.4.7 Logrank*Test*

Es quizá la más popular prueba para demostrar igualdad en las funciones de riesgo, este test usa $W(T)=1$, esto es igual ponderación. Este test tiene una potencia óptima cuando los índices de riesgo son proporcionales uno al otro.

(ncss_wpengine.netdna-ssl.com, 2017)

Es un contraste de hipótesis para comparar curvas de supervivencia.

Se pueden comparar dos o más de dos.

$$H_0 \cong S_1(t) \cong (t) \tag{3.18}$$

$$H_i = S_1(t) \neq S_2(t) \tag{3.19}$$

$$X^2 = \frac{(\varphi_1 - E_1)^2}{E_1} + \frac{(\varphi_1 - E_1)^2}{E_1} \equiv X_1^2 \tag{3.20}$$

3.7 Regresión Lineal Múltiple

Mongomery (2004), Un modelo de regresión donde interviene más de una variable regresora se llama modelo de regresión múltiple.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_K + \varepsilon \tag{3.21}$$

Se llama modelo de regresión lineal múltiple con k regresores. Los parámetros $\beta_j, j = 0,1,2,\ldots,k$ se llaman coeficientes de regresión. Este modelo describe a un hiperplano en el espacio de k dimensiones de las variables regresoras x_j . el parámetro β_j representa el cambio esperado en la respuesta y por cambio unitario en x_j cuando todas la demás variables regresoras $x_i(i \neq j)$ se mantienen constantes. Por esta razón, a los parámetros $\beta_j, j = 0,1,2,\ldots,k$ se les llama coeficientes de regresión parcial. Los modelos de regresión parcial múltiple se usan con frecuencia como modelos empíricos o como funciones de aproximación, ya se desconoce la relación funcional real entre y y $x_{1,}x_2,\ldots,x_k$ pero dentro de ciertos márgenes de las variables regresoras, el modelo de regresión lineal es una

aproximación adecuada a la función verdadera desconocida. En general, todo modelo de regresión que es lineal en los parámetros, es un modelo de regresión lineal, independientemente de la forma de la superficie que genera.

Fuente-Hernández (2017) Los coeficientes β van a indicar el incremento en la variable independiente por el incremento unitario de la correspondiente variable explicativa. Por lo tanto, estos coeficientes van a tener las correspondientes unidades de medida, este es el modelo para las observaciones individuales.

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k X_{1k} + \varepsilon_1 \qquad k = 1 \\
Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k X_{2k} + \varepsilon_2 \qquad k = 2 \\
Y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k X_{nk} + \varepsilon_n \qquad k = n
\end{aligned}
$$

Se utiliza regresión lineal múltiple cuando se estudia la posible relación entre variables independientes (predictoras o explicativas)$X = (x_1, x_2, \ldots, x_n)$

y una variable dependiente (criterio, explicada, respuesta) Y

Se considera que los valores de la variable dependiente Y han sido generados por una combinación lineal de los valores de una o más variables explicativas y un término aleatorio

Los coeficientes son elegidos de forma que la suma de cuadrados entre los valores observados sea mínima, es decir se va a minimizar la varianza residual.

Esta ecuación recibe el nombre de hiperplano, pues cuando tenemos dos variables explicativas, en vez de recta de regresión tenemos un plano.

En la práctica se elige cuidadosamente las variables a considerar como explicativas las cuales deben cumplir con ciertos criterios: tener sentido numérico,

no debe haber variables repetidas o redundantes, las variables introducidas en el modelo deben tener una cierta justificación teórica, la relación entre variables explicativas en el modelo y casos debe de ser como mínimo de 1 a 10, la relación de las variables explicativas con la variable dependiente debe ser lineal, es decir proporcional.

Montgomery (2004) Es más cómodo manejar modelos de regresión múltiple cuando se expresan en notación matricial. Eso permite presentar en forma muy compacta al modelo, los datos y los resultados. En forma matricial el modelo expresado por la siguiente ecuación:

$$y = \boldsymbol{X}\beta + \varepsilon \tag{3.22}$$

en donde:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \; X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots X_{1k} \\ 1 & X_{21} & X_{13} & \cdots X_{2k} \\ \vdots & \vdots & \vdots & \cdots \vdots \\ \vdots & \vdots & \vdots & \cdots \vdots \\ 1 & X_{n1} & X_{n2} & \cdots X_{nk} \end{bmatrix} \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3.24}$$

En general, y es un vector de n x 1 de las observaciones, $\boldsymbol{X}$ es una matriz de n x p de los niveles de las variables regresoras, β es un vector de p x 1 de los coeficientes de regresión y ε.es un vector de p x 1.de errores aleatorios.

El estimador β por mínimos cuadrados es:

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'y \tag{3.25}$$

Siempre y cuando exista la matriz inversa $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. La matriz $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ siempre existe si los regresores son linealmente independientes, esto es, si ninguna columna de la matriz $\boldsymbol{X}$ es una combinación lineal de las demás columnas. Siendo

$\boldsymbol{X}'$ la matriz transpuesta de diseño.

Modelo de regresión ajustado estima a $E(Y\backslash X)$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k \tag{3.26}$$

Valores ajustados para $k = 1, \ldots n$.

$$\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X_{k1} + \cdots + \hat{\beta}_k X_{nk} \tag{3.27}$$

Residuos

$$\hat{\varepsilon}_k = Y_k - \hat{Y}_k \; para \; k = 1, \ldots n \tag{3.28}$$

3.7.1 Hipótesis

3.7.1.1. <u>Linealidad</u>: Los valores de la variable dependiente están generados por el siguiente modelo lineal:

$$E(u_i) = 0 \tag{3.29}$$

3.7.1.2. Normalidad:

$$u_i \varepsilon N(0, \sigma^2) \quad U \approx N(0, \sigma^2) \tag{3.30}$$

La distribución de la perturbación aleatoria tiene distribución normal.

3.7.1.3.Homocedasticidad: Todos los errores aleatorios tienen las misma varianza.

$$V(u_i) = \sigma^2 \tag{3.31}$$

3.7.1.4.Independencia: Los errores aleatorios son independientes entre sí:

$$E(u_i.\ u_j) = 0, \forall i \neq j \tag{3.32}$$

3.7.2 Varianza Residual

Se descompone la variabilidad de la variable dependiente en dos componentes, una explicada por el modelo de regresión y la variabilidad no explicada atribuida a factores aleatorios.

Considerando la variabilidad de la variable dependiente como:

$$n^* \sigma^2 = \sum (y_i - \bar{Y})^2 \tag{3.33}$$

Y sumando y restando el valor pronosticado por el modelo de regresión obtenemos la siguiente expresión.

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2 \quad (3.34)$$

Teniendo en cuenta que el último término representa la varianza no explicada tenemos:

$$VT = VE + VNE \quad (3.35)$$

3.7.3 Contraste de Regresión

Se denomina contraste de regresión al estudio de la posibilidad de que el modelo de regresión sea nulo, es decir los valores de las variables explicativas *X* no van a influir en la variable dependiente *Y*

$$H_0 \equiv \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 \equiv \exists \beta_j \neq 0 \quad (3.36)$$

La hipótesis nula es que todos los coeficientes menos $\beta_0$son nulos y la hipótesis alternativa o complementaria es que al menos uno que es distinto de 0, puede haber varios que sean nulos, pero al menos existe uno distinto de cero.

3.7.4 Construcción del Contraste

El cociente entre la varianza explicada y la varianza no explicada será aproximadamente 1. Además, al seguir una distribución F, podemos asignar una medida de probabilidad (p-value) a la hipótesis de que la varianza explicada es igual a la varianza no explicada.

Si el p-value es menor de 0.05 se acepta que el modelo de regresión es significativo; en caso contrario no podemos hablar de regresión, pues el modelo sería nulo.

3.7.5. Coeficiente de Determinación R^2

Es la proporción de la varianza explicada (VE)

entre la varianza total (VT).

Por ser cociente de sumas de cuadrados, este coeficiente será siempre positivo.

Si todos los puntos están sobre la recta de regresión, la varianza no explicada será 0, por lo tanto.

$$R^2 = \frac{VE}{VT} = 1 - \frac{0}{VT} = 1 \tag{3.37}$$

Este coeficiente es muy importante pues determina qué porcentaje (en tantos por uno) de la varianza dependiente es explicado por el modelo de regresión.

En general, se pueden clasificar los valores de R^2 como lo muestra la tabla 3.5

Tabla 3.5 Clasificación de los valores de R^2

Menor de 0.3	0.3 a 0.4	0.4 a 0.5	0.5 a 0.85	Mayor de 0.85
Muy malo	Malo	Regular	Bueno	Sospechoso

Además, a diferencia de la varianza residual, este coeficiente es adimensional; esto quiere decir que no está afectado por transformaciones lineales de las variables; lo que quiere decir, si cambiamos las unidades de medida, el coeficiente de determinación permanecerá invariante. (www.halweb.uc3m.es2017).

3.8 Modelos Lineales Generalizados

Modelos lineales generalizados (GLM) es una generalización flexible de la regresión ordinaria. Relaciona la distribución aleatoria de la variable dependiente en el experimento (la función de distribución), con la parte sistemática no aleatoria o predictor lineal a través de una función llamada función de enlace.

En un GLM, se asume que la viable dependiente Y está generada por una función de distribución de la familia exponencial. La media μ de la distribución depende las variables independientes X a través de la fórmula.

$$E(Y) = \mu = g^{-1}(X\beta) \tag{3.38}$$

Donde:

$E(Y)$ es el valor esperado de Y

$X\beta$ es el predictor lineal, una combinación lineal de los ´parámetros desconocidos β .

g^{-1} Es la función de enlace.

Con esta notación, la varianza es típicamente una función V de la media.

$$Var(Y) = V(\mu) = (g^{-1}(X\beta)) \quad (3.39)$$

Es conveniente que si V proviene de una distribución en la familia exponencial, pero podría simplemente ser que la varianza es una función de valor ajustado.

3.8.1 Componentes de un GLM.

Los componentes de un GLM que se describen a continuación, son los siguientes: Componente Aleatoria, Componente Sistemática y Función Link.

3.8.1.1 <u>Componente Aleatoria:</u> Identifica la variable respuesta y su distribución de probabilidad y consiste en una variable aleatoria Y con observaciones independientes $(y_1 \dots y_n)$.

En muchas aplicaciones, las observaciones de Y son binarias, y se identifican como éxito y fracaso. Aunque de modo más general, cada Y_i indicaría el número de éxitos de entre un número fijo de ensayos, y se modelizaría como una distribución binomial. En otras ocasiones cada observación es un recuento, con lo que se puede asignar a Y una distribución de Poisson o una distribución binomial negativa. Finalmente, si las observaciones son continuas se puede asumir para Y una distribución normal.

Todos estos modelos se pueden incluir dentro de la llamada familia exponencial de distribuciones.

$$f(y_i \backslash \theta_i) = a(\theta_i).b(y_i).e^{[y_i Q(\theta_i)]} \quad (3.40)$$

De modo que $Q(\theta)$ recibe el nombre de parámetro normal.

3.8.1.2 Componente Sistemática

La componente sistemática de un GLM específica las variables explicativas, que entran en forma de efectos fijos en un modelo lineal, es decir, las variables x_j se relacionan como:

$$\alpha + \beta_1 x_1 + \cdots \beta_k x_k \tag{3.41}$$

Esta combinación lineal de variables explicativas se denomina predictor lineal. Alternativamente, se puede expresar como un vector $(\eta_i, \ldots, \eta_N)$ tal que:

$$\eta_i = \sum_j \beta_j x_{ij} \tag{3.42}$$

Donde x_{ij} es el valor *j-ésimo* predictor en el *i-ésimo* individuo, e $i=1, \ldots N$. el témino independiente α se obtendría con esta notación haciendo que todos los x_{ij} sean igual a 1 para todos los i

En cualquier caso, se pueden considerar variables que están basadas en otras variables como $x_3 = x_1 x_2$ ó $x_3 = x_2^2$, para modelizar interacciones entre variables o efectos curvilíneos de x_2.

3.8.1.3 Función Link

Se denota el valor esperado de Y como $\mu = E(Y)$, entonces, la función link especifica una función $g(.)$ que relaciona μ con el predictor lineal como:

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \tag{3.43}$$

Así, la función link $g(.)$ relaciona las componentes aleatoria y sistemática.

De este modo para $i=1, \ldots N$,

$$\mu_i = E(Y_i)$$

$$\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij} \tag{3.44}$$

La función g más simple es $g(\mu) = \mu$, esto es, la identidad que da lugar al modelo de regresión lineal clásico.

$$\mu = E(Y) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \tag{3.45}$$

Los GLM permiten la unificación de una amplia variedad de métodos estadísticos como la regresión, los modelos ANOVA y los modelos de datos categóricos.

En realidad se usa el mismo algoritmo para obtener los estimadores de máxima verosimilitud en todos los casos. Este algoritmo es la base del procedimiento de GENMOD de SAS y de la función glm de R.

3.8.2 Modelos Lineales Generalizados para datos binarios

En muchos casos las respuestas tiene solo dos categorías del tipo sí/no de modo que se puede definir una variable Y que tome dos posible valores 1(éxito) y 0 (fracaso), es decir $Y \sim Bin(1, \pi)$.

En respuestas binarias, un modelo análogo al de regresión lineal es:

$$\pi(x) = \alpha + \beta x \tag{3.46}$$

Que se denomina modelo de probabilidad lineal, ya que la probabilidad de éxito cambia linealmente con respecto a x.

El parámetro β representa el cambio en probabilidad por unidad de x.. Este modelo es un GLM con un componente aleatorio binomial y con función link igual a la identidad.

Sin embargo, este modelo tiene el problema de que aunque las probabilidades deben estar entre 0 y 1, el modelo puede predecir a veces valores

$\pi(x) > 1 \; y \; \pi(x) < 0.$

3.9 Regresión Logística

Montgomery (2004) Se considerará el caso en el que la variable de respuesta, es un problema de regresión sólo asume dos valores posibles: 0 y 1; esos números podrían ser asignaciones arbitrarias a una respuesta cualitativa. Por ejemplo, la respuesta podría ser el resultado de una prueba de funcionamiento eléctrico para un dispositivo semiconductor, que da como resultado un “éxito”, que

indica que el dispositivo trabaja bien, o un “fracaso” que podría deberse a un corto, un circuito abierto u otro problema de funcionamiento.

Supóngase que el modelo tiene la forma:

$$y_i = X_i'\beta + \varepsilon_i \tag{3.47}$$

En donde $X_i' = [1, x_{i1}, x_{i2}, \dots, x_{ik}], \beta' = [\beta_0, \beta_1, \beta_2, \dots \beta_k]$, y la variable de respuesta y_i toma los valores 0 o 1. Se supondrá que la variable de respuesta y_i es una variable aleatoria de Bernoulli, cuya distribución de probabilidad se muestra en la tabla 3.6.

Tabla 3.6 Distribución de Probabilidad

y_i	Probabilidad
1	$P(y_i = 1) = \pi_i$
2	$P(y_i = 0) = 1 - \pi_i$

Ahora bien, como $E(\varepsilon_i) = 0$, el valor esperado de la variable de respuesta es:

$$E(y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \tag{3.48}$$

Esto implica que:

$$E(y_i) = X_i'\beta = \pi_i \tag{3.49}$$

Que quiere decir que la respuesta esperada, determinada con la función de respuesta $E(y_i) = X_i'\beta$ no es más que la probabilidad de que la variable de respuesta tenga el valor de 1.

Hay algunos problemas sustantivos con el modelo de regresión en la ecuación (3.47).

El primero es que se observa que si la respuesta es binaria, entonces los términos de error ε_i sólo pueden tener dos valores que son:

$$\varepsilon_i = 1 - X_i'\beta \;\; cuando \; y_i = 1$$

$$\varepsilon_i = -X_i'\beta \;\; cuando \; y_i = 0 \tag{3.50}$$

En consecuencia, no es posible que los errores en este modelo sean normales. En segundo lugar, la varianza del error no es constante, ya que

$$\sigma_{yi}^2 = E\{y_i - E(y_i)\}^2$$
$$= (1 - \pi_i)^2\pi_i + (0 - \pi_i)^2(1 - \pi_i) \tag{3.51}$$

$= \pi_i(1 - \pi_i)$

Obsérvese que esta última expresión equivale a

$$\sigma_{yi}^2 = E(y_i)[1 - E(y_i)] \quad (3.52)$$

Porque $E(y_i) = X_i'\beta = \pi_i$, lo que indica que la varianza de las observaciones (que es igual a la varianza de los errores, porque $\varepsilon_i = y_i - \pi_{i,}$ y π_i es constante) es una función de la media. Por último, hay una restricción para la función de respuesta, ya que

$$0 \leq E(y_i) = \pi_i \leq 1 \quad (3.53)$$

Esta restricción puede causar graves problemas en la elección de una función de respuesta lineal, como se ha supuesto al principio en la ecuación (3.47) Sería posible ajustar un modelo con los datos para los cuales los valores predichos de la respuesta salen del intervalo 0,1.

En general cuando la variable de respuesta es binaria, hay bastantes pruebas empíricas que indican que la forma de la función de respuesta debe ser no lineal. Una función monótonamente creciente (o decreciente) en forma de S (o de S invertida), como la de la figura 3.2 es la que se acostumbra emplear; esta función se llama función de respuesta logística y tiene la forma

$$E(y) = \frac{e^{X'\beta}}{1 + e^{X'\beta}}$$

O bien lo que es igual (3.54)

$$E(y) = \frac{1}{1 + e^{-X'\beta}}$$

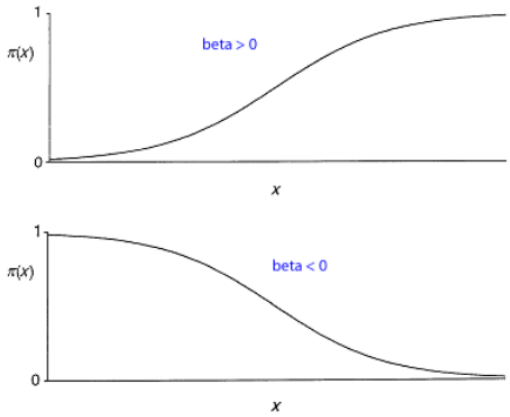


Figura3.2 Función de Respuesta Logística

La función de respuesta logística se puede linealizar con facilidad. Un enfoque consiste en definir la porción estructural del modelo en términos de una función de la media de la función de respuesta.

Sea $\eta = X'\beta$ el predictor lineal estando definida η por la transformación

$$\eta = In\ \frac{\pi}{1 - \pi} \quad (3.55)$$

A esta transformación se le llama con frecuencia transformación logit de la probabilidad π , y la relación $\frac{\pi}{1-\pi}$ en la transformación se llama ventaja; a veces, a la transformación logit se le llama ventaja logarítmica.

Hay otras funciones que tienen la misma forma que la función logística, y también se pueden obtener transformando π , una de ellas es la transformación probit obtenida transformando a π con la distribución normal acumulada. De esta manera se obtiene un modelo de regresión probit, este modelo es menos flexible que el de regresión logística y es probable que no se use tanto, porque no puede incorporar con facilidad más de una variable predictora. Otra tansformación posible es la log-log complementaria de π .

3.9.1 Estimación de parámetros en un modelo de Regresión Logística

El valor estimado del predictor lineal es $\hat{\eta}_i = X_i'\widehat{\beta},$ y el valor esperado del modelo de regresión logística se escribe con frecuencia como sigue:

$$\left.\begin{aligned} \hat{y}_i = \hat{\pi}_i &= \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}} \\ &= \frac{e^{X_i'\widehat{\beta}}}{1+e^{X_\beta'}} \\ &= \frac{1}{1+e^{-X_i'\widehat{\beta}}} \end{aligned}\right\} \quad (3.56)$$

3.9.1.1 <u>Interpretación de los parámetros en un modelo de Regresión Logística</u>

Primero se examinará el caso en el que el predictor lineal solo tiene un regresor, por lo que el valor ajustado del modelo en determinado valor de x por ejemplo

x_i, es $\hat{\eta}(x_i) = \hat{\beta}_0 + \hat{\beta}_i x_i$.

El valor ajustado de $x_i + 1$ es (3.57)

$\hat{\eta}(x_i + 1) = \hat{\beta}_0 + \hat{\beta}_i(x_i + 1)$.

Y la diferencia entre los dos valores predichos es

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = \hat{\beta}_1 \quad (3.58)$$

Ahora, $\hat{\eta}(x_i)$ sólo es el logaritmo de la ventaja cuando la variable regresora es igual a x_i y $\hat{\eta}(x_i + 1)$ es el logaritmo de la ventaja cuando el regresor es igual a $x_i + 1$. Por consiguiente, la diferencia entre los dos valores ajustados es:

$$\hat{\eta}(x_i + 1) - \hat{\eta}(x_i) = In\left(ventaja_{x_{i+1}}\right) - In\left(ventaja_{x_i}\right) \quad (3.59)$$

$$= In\left(\frac{ventaja_{x_{i+1}}}{ventaja_{x_i}}\right) = \hat{\beta}_1$$

Si se sacan antilogaritmos se obtiene el cociente de ventaja

$$\hat{O}_R = \frac{ventaja_{x_{i+1}}}{ventaja_{x_i}} = e^{\hat{\beta}_1} \tag{3.60}$$

Se puede interpretar el coeficiente de ventaja como el aumento estimado en la probabilidad de éxito asociado con un cambio unitario en el valor de la variable predictora, en general el aumento estimado del cociente de ventaja, asociado con un cambio de d unidades en la variable predictora, es $e^{\widehat{d\beta}_1}$

La interpretación de los coeficientes de regresión en el modelo logístico múltiple, se parece al caso en el que el predictor lineal sólo contiene un regresor, que nos indica que la cantidad $e^{\hat{\beta}_j}$ es el cociente de ventaja para el regresor x_j, suponiendo que las demás variables predictoras son constantes.

3.9.2 Pruebas de Hipótesis para los Parámetros del Modelo

La prueba de hipótesis en la regresión logística (y en general, para el modelo lineal general) se basa en pruebas de cociente de máxima verosimilitud, que es un procedimiento para muestras grandes, por lo que los procedimientos de prueba se basan en la teoría asintótica. El método de la razón de verosimilitud conduce a un estadístico llamado desviación.

Si $\lambda(\beta) \leq x^2_{\alpha,n-p}$ se concluye que el modelo ajustado es adecuado

Si $\lambda(\beta) > x^2_{\alpha,n-p}$ se concluye que el modelo ajustado no es adecuado.

La desviación está relacionada con una cantidad muy conocida. Si se considera el error normal estándar del modelo de regresión lineal, sucede que la

desviación es el error de la suma de cuadrados de residuales dividido entre la varianza del error σ^2.

3.10 Regresión Probit

Otro modelo en el que se pueden considerar curvas en forma de S son los modelos *probit* . Una idea natural es que

$\pi(x) = F(x)$, siendo F una función de distribución. Cuando X es una variable aleatoria continua, la función de distribución como función de x , tiene forma de S. Esto sugiere una clase de modelos de dependencia para modelos binarios.

Como caso particular se puede considerar el link *probit* que transforma probabilidades en valores estándar de la función de distribución normal,

$$\begin{aligned} F(x) &= \Phi(x) \\ \pi(x) &= \Phi(\alpha + \beta x) \\ \Phi^{-1}(\pi(x)) &= \alpha + \beta x \end{aligned} \qquad (3.61)$$

Así de define probit $\equiv \Phi^{-1}$.

En la práctica, tanto los modelos probit como logit producen ajustes similares.

3.11 Modelos GLM para recuentos

En muchos casos las variables respuesta son recuentos, y en muchas ocasiones los recuentos aparecen al resumir tablas de contingencia.

En el modelo más simple se asume que el componente aleatorio Y sigue una distribución de Poisson. Esta distribución es unimodal y su propiedad más destacada es que la media y la varianza coinciden.

$$E(Y) = Var(Y) = \mu \tag{3.62}$$

De modo que cuando el número de recuentos es mayor en media, también tienden a tener mayor variabilidad.

En el modelo GLM se usa habitualmente el logaritmo de la media para la función link de modo que el modelo *loglineal* con una variable explicativa X se puede expresar como:

$$\log(\mu) = \mu + \beta x \tag{3.63}$$

De modo que

$$\mu = e^{[\mu + \beta x]} = e^{\alpha}\left(e^{\beta}\right)^{x} \tag{3.64}$$

3.12 GLM binomiales negativos

Si una variable aleatoria Y se distribuye como una binomial negativa, entonces la función de probabilidad es

$$P(y\backslash k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+\ 1)}\left(\frac{k}{\mu+k}\right)^{k}\left(1-\frac{k}{\mu+k}\right)^{y} \tag{3.65}$$

Con $y = 0,1,2,\ldots$ donde k y μ son los parámetros de la distribución.

Se tiene que:

$$E(Y) = \mu \tag{3.66}$$

$$Var(Y) = \mu + \frac{\mu^2}{k} \tag{3.67}$$

El parámetro $\frac{1}{k}$ es un parámetro de dispersión, de modo que si $\frac{1}{k} \to 0$ entonces $Var(Y) \to 0$ y la distribución binomial negativa converge a una distribución Poisson. Por otro lado, para un valor fijo de k esta distribución pertenece a la familia exponencial natural, de modo que se puede definir como un GLM binomial negativo. En general, se usa una función link de tipo logaritmo.

La regresión binomial negativa se puede utilizar para datos sobredispersos de recuentos, es decir cuando la varianza condicional es mayor que la media condicional. Se puede considerar como una generalización de la regresión de Poisson, ya que tiene su misma estructura de medias y además un parámetro adicional para el modelo de sobredispersión.

Si la distribución condicional de la variable observada es más dispersa, los intervalos de confianza para la regresión binomial negativa es probable que sean más estrechos los correspondientes a un modelo de regresión de Poisson.

3.13 Modelo de Riesgos Proporcionales (Regresión de Cox)

El modelo de riesgos proporcionales de Cox (MRPC) constituye un modelo multivariado que puede ponderar el efecto de una serie de variables cualitativas cuantitativas sobre un desenlace dicotómico a través del tiempo.

El MRPC fue creado y publicado en 1972, en la revista *Journal of Royal Statistical Society* por el estadístico inglés Sir David R. Cox, como una alternativa al método Kaplan-Meier en la que se incorporan las características del análisis de regresión en las tablas de supervivencia (Pérez-Rodríguez, 2014).

Al hablar de supervivencia gran parte del ejercicio clínico relacionado con la atención médica, la toma de decisiones implica el conocimiento del curso clínico de la enfermedad, es decir, la estimación del tiempo transcurrido hasta que un evento suceda (Pérez-Rodríguez, 2014).

El análisis de supervivencia es un área estadística en la que la variable de respuesta es el tiempo que transcurre entre un evento inicial (que la inclusión del individuo en el estudio) y un evento final (generalmente llamado falla) que ocurre cuando el individuo presenta la característica para terminar el estudio (muerte, alta de la enfermedad, etc.) (Flores-Luna ,2000).

Este modelo se ha convertido en uno de los métodos más utilizados, siendo los temas relacionados con la salud una de las áreas en las que se ha tenido mayor

aplicación, es particularmente útil para comparar grupos en los que se estudia el tiempo transcurrido hasta la ocurrencia de un evento, pudiendo analizar conjuntamente el efecto de varias variables (Pérez Rodríguez 2014)

El modelo de Cox se adapta fácilmente a situaciones con datos incompletos, algo que ocurre frecuentemente en la investigación (Corpas-Nogales, 2009). Donde hace una aplicación de este modelo a datos de mujeres infectadas con SIDA (*Síndrome de Inmunodeficiencia Adquirida*), donde se estudiaron 1366 mujeres diagnosticadas en Andalucía, España entre 1982 y 2001. Las variables utilizadas fueron edad, fecha de diagnóstico, provincia de residencia y categoría de transmisión.

Flores-Luna (2000) presenta un análisis de supervivencia aplicado a mujeres con CaCu en el Hospital de Ginecobstetricia No. 4 en México entre el año 1984 y 1996, donde utiliza el modelo de regresión proporcional de Cox para determinar el efecto conjunto de los factores pronósticos que resultan significativos de forma individual, este modelo también especifica cómo cambia la función de riesgo básica (individuos con nivel de covariable cero) respecto de aquellos con covariables distintas de cero. Este cambio lo especifica el parámetro asociado a cada factor introducido al modelo y se interpreta como el cambio esperado en el cociente de riesgos entre un individuo en la población básica y uno fuera de ella (Flores-Luna, 2000).

Fueron consideradas, dentro de la población de estudio, las mujeres que fueron canalizadas, diagnosticadas y tratadas en ese hospital con CaCu invasor

confirmadas histopatológicamente y que su entrada al estudio depende del diagnóstico, independientemente de la fecha del mismo.

Pardo (2009) investiga sobre supervivencia de pacientes con cáncer de cuello uterino tratadas en el Instituto Nacional de Cancerología en Bogotá Colombia, con el objetivo de describirla de forma global a cinco años, donde toma como factores de pronóstico el estadío clínico y la presencia de compromiso ganglionar, la invasión del estorma y la presencia de invasión vascular o linfática, tipo histológico y grado de diferenciación y menciona que la edad, el nivel socio económico, la raza y los niveles de hemoglobina, pierden su efecto al incluir a los demás factores en modelos multivariados, por lo que su papel como factores de pronóstico no están plenamente establecidos; para el análisis de supervivencia global se hizo con el método Kaplan-Meier, el análisis multivariado de factores de pronóstico se realizó con un modelo de regresión de Cox, se verificaron los supuestos de riesgo proporcional para cada factor y para el modelo global en el modelo definitivo.

Entre 1998 y 2011 se identificaron a 2205 mujeres de *National Cancer Data Base*, diagnosticadas con edenosarcoma Mulleriana del útero u ovario con el fin de probar asociaciones de potencial y variables explicativas, la supervivencia global y los factores asociados, utilizando el análisis de supervivencia de riesgo proporcional de Cox (Brandon, 2016).

3.13.1 Modelo de Cox

Gómez y Cobo, (2004) publica que el modelo de Cox es el equivalente en análisis de la supervivencia al modelo de regresión lineal. Se trata de un modelo

semiparamétrico, ya que no exige ninguna forma para la función de riesgo (puede ser creciente o decreciente) pero, en cambio, define un parámetro que es la razón entre ambas funciones de riesgo,del inglés, *(hazard ratio)* este planteamiento responde de forma natural a la pregunta del clínico, que no ambiciona conocer cuál es el tiempo exacto de vida de un paciente determinado (parte del modelo no parametrizado, y centra sus esfuerzos en identificar las intervenciones que pueden aumentarlo (parte del modelo parametrizado). El modelo de Cox no impone como premisa una función para la función de riesgo, pero si asume que la razón de riesgo toma el mismo valor durante todo el período de seguimiento.

El modelo de Cox puede considerar simultáneamente la hipotética relación de varias variables con la supervivencia y estudia, de esta manera, si la relación de algunas persiste una vez se ha tenida en cuenta al resto. Para hacerlo, extiende la premisa de riesgos proporcionales a cada variable considerada, tanto variables cuantitativas como categóricas. La construcción del modelo es una tarea muy delicada ya que se debe especificar la relación que cumple con esta premisa de proporcionalidad, al mismo tiempo que se evita introducir variables que tengan entre ellas una elevada colinealidad. Si el modelo está bien elaborado, su interpretación es más simple ya que la razón de riesgo establecida para una variable puede interpretarse independientemente, es decir, de manera fija del resto de las variables. La decisión final sobre si una relación puede interpretarse como causal, debe basarse en la existencia de la correspondiente hipótesis antes de la obtención de los datos, así como de su procedencia; observacional o experimental

Boj del Val (2017) publicó que, el modelo de Cox expresa la función de riesgo

$h(t)$ en función del tiempo t y de un conjunto de covariables /variables explicativas, predictoras, factores de riesgo, variables de confusión $\boldsymbol{X}=(X_{1,\ldots\ldots}X_{p)}$,que definen al sujeto en estudio de acuerdo a la ecuación:

$$h(t,\boldsymbol{X}) = h\big(t, X_1, \ldots X_p\big) = h_{\,0}(t)\, e^{(\sum_{j=1}^{p} \beta_j X_j)} \tag{3.68}$$

Como hipótesis de partida, supondremos que los tiempos de supervivencia tienen distribuciones continuas, que están tomados de forma exacta y que no existe posibilidad de empates. Para cada sujeto $i\ para\ i = 1, \ldots, n$ conoceremos su tiempo de muerte /fallo t_i, su estado de fallo o censura d_1, variable codificada con 1 si el dato no está censurado y con 0 si el dato sí lo está, y las covariables fijas $\boldsymbol{X_i} = (X_{i1\ldots,}X_{ip)}$. Si incluimos el subíndice i para denotar a un sujeto determinado, el modelo (3.64) se podría reescribir como:

$$h\big(t_i,\ \boldsymbol{X_{ip}}\big) = h\big(t_i, X_{i1}, \ldots X_{ip}\big) = h_{\,0} \tag{3.69}$$

A la función $h_0(t)$ se le denomina función de riesgo basal y corresponde al riesgo de un individuo que tiene como valor en todos los predictores 0, el cual sería el individuo de referencia de cara a la interpretación posterior del análisis:

$$h(t, X_1 = 0, \dots, X_p =) = h_0(t)e^{\left(\sum_{j=1}^{p} \beta_j 0\right)} = h_0(t)e^0 = h_0(t)1 = h_0(t) \quad (3.70)$$

También se interpreta que la función de riesgo basal sería aquella función básica del modelo si éste no incorpora predictores.

La función de riesgo basal, $h_0(t)$, es la única parte de la expresión del modelo de Cox que depende del tiempo t . La otra parte $(\sum_{j=1}^{p} \beta_j X_j)$, solo depende del vector de covariables $\boldsymbol{X}=(X_{1,\dots\dots}X_p)$, de los sujetos, que en este apartado supondremos independiente del tiempo.

Una variable independiente del tiempo se define como una variable cuyos valores no varían a lo largo del tiempo. Por ejemplo, el sexo, la raza, o el grupo de tratamiento son variables fijas, sólo toman un valor, el inicial. También podríamos considera variables como el hecho de ser fumador (estado de fumador) como variable independiente del tiempo, ya que, aunque el estado de fumador pude variar en el tiempo, se suele suponer que para el estudio no varía, se parte de un estado inicial y se supone que no cambia hasta el final, y por lo tanto que sólo se toma el valor del individuo. Otro ejemplo de este tipo podría ser la variable estado inicial de la enfermedad. Cabe notar que hay variables cambiantes con seguridad, pero que también se suelen tratar como independientes del tiempo; como lo son la edad y el peso de los sujetos sí varían con el tiempo, pero puede ser inapropiado tratarlas como independientes del tiempo en análisis determinados. Esto es posible siempre que los valores de estos predictores no varíen en exceso a lo largo del tiempo, o bien si el efecto de dichas variables en el riesgo de supervivencia depende esencialmente.

El modelo de Cox se considera un modelo “semiparamétrico debido a que incluye una parte paramétrica y otra parte no paramétrica.

a) La parte paramétrica se corresponde con:$e^{(\sum_{j=1}^{p}\beta_j X_j)}$), es decir, con la exponencial del predictor lineal $\eta = \sum_{j=1}^{p}\beta_1 X_j$. En esta parte del modelo se estiman los parámetros $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1 \ldots., \hat{\beta}_p)$ de la regresión mediante la maximización de la denominada función de verosimilitud parcial que estudiaremos con detalles.

b) La parte no paramétrica es la función de riesgo basal $h_0(t)$, Ésta es una función arbitraria y no especificada y se estima en un segundo estadío, condicionada a la estimación de los parámetros $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1 \ldots., \hat{\beta}_p)$de la regresión.

   Es por ésta componente no paramétrica de la fórmula que el modelo de Cox se considera semiparamétrico.

Una vez estimada la parte paramétrica$e^{(\sum_{j=1}^{p}\widehat{\beta}_j X_j)}$ y posteriormente la no paramétrica, $\hat{h}_0(t)$, tendremos el modelo semiparamétrico completo.

$$\hat{h}(t, \boldsymbol{X}) = \hat{h}_0(t) e^{(\sum_{j=i}^{p}\widehat{\beta}_j X_j)} \qquad (3.71)$$

El hecho de que el modelo de Cox sea un modelo semiparamétrico hace que sea bien recibido en análisis de supervivencia. Al no tener especificada la función de riesgo basal es posible estimar los coeficientes de regresión, calcular las razones de riesgo y ajustar las curvas de supervivencia a una gran variedad de situaciones.

Podemos decir que el modelo de Cox es robusto en el sentido de que los

resultados obtenidos en los ajustes tendrán a aproximarse a los del modelo paramétrico correcto.

Dicho de otro modo, con este modelo de Cox evitamos utilizar un modelo paramétrico incorrecto para el estudio.

Podríamos definir el modelo de regresión de Cox (3.68) de un modo más general en lo referente a la parte paramétrica de acuerdo a la ecuación:

$$h(t, \boldsymbol{X}) = h_0\,(t)\psi(\eta) \tag{3.72}$$

La parte $\psi(\eta)$ se interpreta como el riesgo relativo en el momento t de un individuo con perfil $\boldsymbol{X} = (X_1\,, \ldots, X_p)$respecto de un individuo con $\boldsymbol{X} = (0, \ldots, 0)$. El modelo (3.68) se correspondería con $\psi(\eta) = e^{(\sum_{j=1}^{p} \beta_j\, X_j)}$. Esta parametrización del modelo de Cox se denomina log lineal y es la más popular. También se considera la forma lineal con $\psi(\eta) = 1 + \sum_{j=1}^{p} \beta_j\ X_j$ y la logarítmica con

$\psi(\eta) = \log\left(1 + e^{\left(\sum_{j=1}^{p} \beta_j\, X_j\right)}\right)$, aparte de alguna familia dependiente de otros parámetros adicionales que obtiene a las lineales (log lineal y lineal) como casos particulares.

Si analizamos la expresión paramétrica del modelo de Cox definida en (3.68), observaremos que al suponer que el predictor lineal está relacionado con las curvas de riesgo (o funciones de supervivencia) a través de la exponencial, esto nos asegura que nunca obtendremos valores negativos en la estimación del modelo. Esta es una propiedad deseable que la forma lineal no siempre cumple.

Otra propiedad atractiva del modelo de Cox es, aunque la función de riesgo basal no esté especificada, es posible determinar los parámetros del predictor lineal. Una vez estimados los coeficientes, podremos también estudiar el efecto de las variables explicativas de interés y calcular los denominados *ratios* o razones de riesgo que posteriormente se define, todo ello de forma independiente a la estimación de las funciones de riesgo. En el modelo de Cox se estima en un segundo estado la función de riesgo $h(t, \boldsymbol{X})$(y las correspondientes curvas de supervivencia $S(t, \boldsymbol{X})$) con muy pocos supuestos ya que la función de riesgo basal $h_0(t)$o de supervivencia basal, $S_0(t)$ son funciones no especificadas.

Un último punto sobre la popularidad del modelo de Cox es que es preferido al modelo de regresión logarítmico cuando la variable objeto de estudio son tiempos de supervivencia que pueden estar o no censurados, pues el modelo de Cox utiliza más información sobre estos datos que el logarítmico. El logarítmico únicamente trata con respuestas del tipo 0 y 1 y no tiene en cuenta las censuras.

Comentar que para el modelo de Cox de riesgos proporcionales se procede de forma análoga que, con los modelos lineales o lineales generalizados en lo referente al tratamiento y codificación de predictores, excepto que el término *intercept* o medio queda absorbido por la función de riesgo basal. Se pueden tener en cuenta los efectos principales, aunque estos se vuelvan menos significativos con la introducción de la interacción.

Si la naturaleza del predictor lo permite, se pueden hacer transformaciones de las variables para que entren de un modo más adecuado en el modelo, por ejemplo, la forma cuadrática o cúbica o bien estandarizada por algún método. En el modelo de Cox, y debido a la interpretación que se realiza, interesará tener a

predictores continuos discretizados en clases que sean de interés y susceptibles de interpretación en las denominadas razones de riesgo luego se definirán.

3.13.1.1 Hipótesis de Riesgos Proporcionales

En el modelo de Cox se busca, como primer paso, la relación entre los riesgos de muerte de dos individuos expuestos a factores de riesgo diferentes. Para ello, el modelo parte de una hipótesis fundamental, la de que los riesgos son proporcionales.

HR entre dos sujetos con diferente vector de covariables

$\boldsymbol{X} = (X_1, \ldots . X_p)y(X_1^*, \ldots, X_p^*)$ como:

$$HR = \frac{h(t, \boldsymbol{X}^*)}{h(t, \boldsymbol{X})} \tag{3.73}$$

Al igual que se realiza en los *odds ratios, (probabilidades de riesgo),* típicamente se evalúa en el numerador el grupo de mayor riesgo definido por $\boldsymbol{X}^*$y en el denominador el grupo de menor riesgo definido por $\boldsymbol{X}$ .en tal caso se espera que el *HR* sea mayor que 1, ya que

$h(t, \boldsymbol{X}^*) > h(t, \boldsymbol{X})$y cuantifica cuántas veces es mayor el riesgo de morir con perfil $\boldsymbol{X}^*$ que con $\boldsymbol{X}$ **.**Es más fácil la interpretación si excede el valor base unidad que indica que tienen el mismo riesgo, que si disminuye la unidad. Si sustituimos la expresión del modelo (3.68) en (3.73) obtenemos la ecuación:

$$HR = \frac{h_0(t)\, e^{\left(\sum_{j=1}^{p} \beta_j X_j^*\right)}}{h_0(t)\, e^{\left(\sum_{j=1}^{p} \beta_j X_j\right)}} = \frac{e^{\left(\sum_{j=1}^{p} \beta_j X_j^*\right)}}{e^{\left(\sum_{j=1}^{p} \beta_j X_j\right)}} = e^{\sum_{j=1}^{p} \beta_j \left(X_j^* - X_i\right)} \tag{3.74}$$

Con lo que

$$HR = \mathrm{e}^{\sum_{j=1}^{p} \beta_j \left(x_j^* - x_j\right)} \tag{3.75}$$

Se observa que el resultado de la *HR* (3.75) no depende de la función de riesgo basal, tan sólo del valor de los predictores y de las betas estimadas, i.e., no depende del tiempo. Por lo tanto, en el modelo de Cox se supone la hipótesis de que los riesgos son proporcionales, ya que se suponen covariables no dependientes del tiempo.

La hipótesis de riesgos proporcionales significa explícitamente que la razón de riesgo (3.69) es constante con el tiempo: $h(t, \boldsymbol{X}^*)$= constante x $h(t, \boldsymbol{X})$ Si lo aplicamos a la expresión resultante en el modelo de Cox (3.64) denominado θ a la constante y una vez estimados los coeficientes de la regresión por máxima verosimilitud parcial, tenemos que la razón de proporcionalidad es constante en el tiempo e igual a la ecuación:

$$\hat{\theta} = e^{\left(\sum_{j=1}^{p} \hat{\beta}_j \left(x_j^* - x_j\right)\right)} \tag{3.76}$$

En el caso de dos individuos, i y j que solo se diferencian en la *k-ésima* variable, supongamos que X_k vale cero para, i y 1 para j, entonces, tenemos que para cualquier tiempo t la denominada *HR* es:

$$HR = \frac{h(t, X_{1,} \dots X_{k-1,} 1, X_{k+1,} \dots, X_p)}{h(t, X_{1,} \dots X_{k-1,} 0, X_{k+1,} \dots, X_p)} \tag{3.77}$$

Y vale exactamente $HR = e^{\beta_k}$, ya que:

$$\frac{h_0(t) e^{(\beta_1 X_1 + \cdots \beta_{k-1} X_{k-1} + \beta_k 1 + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p)}}{h_0(t) e^{(\beta_1 X_1 + \cdots \beta_{k-1} X_{k-1} + \beta_k 0 + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p)}} \tag{3.78}$$

$$= \frac{e^{(\beta_1 X_1 + \cdots \beta_{k-1} X_{k-1} + \beta_k 1 + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p)}}{e^{(\beta_1 X_1 + \cdots \beta_{k-1} X_{k-1} + \beta_k 0 + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p)}} = e^{\beta_k} \tag{3.79}$$

Comentar que en caso de covariables continuas el $HR = e^{\beta_j}$representa la *HR* de incrementar en una unidad la covariable continua X_j. Si nos resulta interesante estimar la *HR* al incrementar la covariable $X_j \; en \; c$ unidades, lo haremos mediante $e^{c\beta_j}$.

Al utilizar la regresión de Cox para unos datos determinados será necesario verificar que se cumple esta hipótesis de proporcionalidad de riesgos. Para ello se suele comprobar que el efecto de cada variable es constante en el tiempo.

Existen varios modelos para estudiar el cumplimiento de la hipótesis de proporcionalidad. Por un lado, puede utilizarse un método gráfico: si una variable, por ejemplo toma únicamente los valores 0 y 1, pueden representarse las curvas de supervivencia para los dos grupos de sujetos definidos por dicha variable y estudiar si son paralelas. No obstante, existen métodos estadísticos algo más rigurosos como el estudio de los denominados residuos de Schoenfeld.

3.13.1.2 Función de Verosimilitud Parcial. Estimación de los Coeficientes

En el método de regresión de Cox, los parámetros $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$se estiman maximizando el logaritmo de la denominada función de verosimilitud parcial. La maximización de dicha función se realiza mediante métodos, obteniendo de esta forma la estimación $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$. Con la estimación de estos parámetros ya tendremos la componente paramétrica totalmente especificada según la ecuación

$$h(t, \boldsymbol{X}) = h_{\boldsymbol{0}}(t) e^{(\sum_{j=1}^{p} \widehat{\beta}_j X_j)} \tag{3.80}$$

y consecuentemente podemos hacer inferencia sobre dicho vector de parámetros y calcular los HR de interés para el estudio.

La función de verosimilitud parcial que a continuación se define, se denomina parcial debido a que tiene en cuenta únicamente en la función de verosimilitud las posibilidades de los tiempos de muerte/fallo y no incluye las probabilidades de los tiempos de datos censurados. Sin embargo, en el cálculo de las probabilidades de los tiempos de muertes sí tiene en cuenta a todos los sujetos (censurados o no a *posteriori*) objeto de riesgo al inicio de los diferentes tiempos de muerte.

Denominados $L \equiv L(\beta_1, \dots \beta_p)$ a la función de verosimilitud parcial. Supongamos que tenemos k tiempos de muerte y que no hay empates. Así, tendremos $n - k$ tiempos censurados. Los tiempos de muerte ordenados los denotamos por $(t_1, \dots, t_k)$ y denotamtos por $R(t_i)$ para $i = 1, \dots k$ al conjunto de los sujetos de riesgo en el tiempo t_i. Denominamos por $L \equiv L_{t_i}(\beta_1, \dots, \beta_p)$ para $i = 1, \dots k$

a las porciones de la verosimilitud total anterior debidas a las aportaciones de los diferentes tiempos de muerte $(t_1 , \dots t_k)$.

Se construye la función de verosimilitud total como el producto de cada una de las aportaciones de los k tiempos de muerte:

$$L = \prod_{i=1}^{k} L_i \tag{3.81}$$

Una vez con la verosimilitud ya construida, se hace el logaritmo y se deriva respecto a los parámetros.

$$\frac{\partial \log L}{\partial \beta_j} \tag{3.82}$$

$$\frac{\partial^2 \log L}{\partial \beta_1 \partial \beta_j} \tag{3.83}$$

De (3.35) igualando a 0, $\frac{\partial \log L}{\partial \beta_j} = 0 \; para \; j = 1, \dots p$ obtendremos las ecuaciones linealmente independientes que nos permitirán obtener estimaciones de $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1 \dots \hat{\beta}_p\right)$ mediante la utilización de algún método numérico.

De (3.79) comprobamos que realmente es un máximo y podemos obtener, como ocurre cuando se trabaja en general con una función de verosimilitud, la "matriz de información (observada),$\boldsymbol{I}(\boldsymbol{\beta})$ donde cada elemento se iguala a:

$$I_{ij}\ (\boldsymbol{\beta}) = -\frac{\partial^2 \log L}{\partial \beta_1 \partial \beta_j} \tag{3.84}$$

Así, la matriz de varianzas y covarianzas $p \times p$ estimada es $\widehat{\sum} I^{-1}(\hat{\beta})$ Cabe notar que este estimador, obtenido a partir de la maximización de la verosimilitud parcial es asintóticamente no sesgado, eficiente y normal.

Por un lado $\widehat{\boldsymbol{\beta}}$ estima consistentemente el vector de parámetros β, pero no es completamente eficiente, es decir no alcanza la cota de Cramer –Rao.

Finalmente la distribución de $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1 \dots \hat{\beta}_p\right)$ es aproximadamente normal de media $\left(\hat{\beta}_1 \dots \hat{\beta}_p\right)$y matriz de varianzas y covarianzas$\widehat{\sum}$.

La ecuación $L \equiv L_{t_{(i)}}(\beta_1, \dots, \beta_p)$ muestra el valor exacto para cada una de las L

$i = 1, \dots k$:

$$L_{t_i}(\beta_1, \dots, \beta_p) = \frac{h(t_i, X_i)}{\sum_{l \epsilon R(t_i} h(t_i, X_i)} = \frac{h_{\mathbf{0}}(t_i) e^{(\sum_{j=1}^{p} \beta_j X_{ij})}}{\sum_{l \epsilon R(t_i} h_0(t_i) e^{(\sum_{j=1}^{p} \beta_j X_{ij})}} \tag{3.85}$$

Siendo $\boldsymbol{X_i}$el vector de covariables para el sujeto con tiempo de muerte t_i y $\boldsymbol{X}_l$

Para $l \epsilon R(t_i)$el vector de covariables de cada uno de los sujetos de $R(t_i)$.

Como podemos ver la función de riesgo basal se anula en el numerador y en el denominador, con lo que nos queda la expresión:

$$L_{t_i}(\beta_1, \dots, \beta_p) = \frac{e^{(\sum_{j=1}^{p} \beta_j X_{ij})}}{\sum_{l \epsilon R(t_i} e^{(\sum_{j=1}^{p} \beta_j X_{ij})}} \quad (3.86)$$

Observamos que la función de verosimilitud parcial total así calculada no depende de las cuantías de los tiempos, tan sólo de su ordenación y de si el dato estaba o no censurado. Como consecuencia podríamos obtener las mismas estimaciones de $\widehat{\boldsymbol{\beta}}$ para distintos datos, siempre que éstos tengan el mismo patrón de orden y censura en los tiempos de supervivencia.

3.13.1.3 Contrastes de Hipótesis

Como ya hemos comentado, a partir de la función de verosimilitud parcial obtenemos una estimación de los coeficientes $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1 \dots \hat{\beta}_p)$ cuya distribución es aproximadamente normal de media $(\hat{\beta}_1 \dots \hat{\beta}_p)$y matriz de varianzas y covarianzas $\widehat{\sum}$. Lo que permite utilizar test análogos a los utilizados en un modelo lineal o lineal generalizado.

Para resolver la hipótesis $H_0: \beta_j = 0$ versus $H_1: \beta_j \neq 0$ podemos utilizar el estadístico de Wald

$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{Var}\hat{\beta}_j}} \tag{3.87}$$

3.13.1.3 Ajuste de Curvas de Riesgo/Supervivencia en el Modelo de Cox

El modelo de Cox (3.68) puede expresarse en términos de funciones de supervivencia. El modelo en función de las funciones de riesgo es:

$$h(t, \boldsymbol{X}) = h_{\mathbf{0}}(t) e^{(\Sigma_{j=1}^{p} \beta_j X_j)}$$

Puesto que se cumple la relación $S(t) = e^{(-\int_0^t h(s)ds)}$, se puede comprobar que el modelo de Cox en términos de función de supervivencia es:

$$S(t, \boldsymbol{X}) = S_{\mathbf{0}}(t)^{e^{(\Sigma_{j=1}^{p} \beta_j X_j)}} \tag{3.88}$$

Donde $S_0(t)$ se denomina función de supervivencia basal siguiendo la idea de la función de riesgo basal, $h_0(t)$, que se corresponde con la supervivencia de un individuo base al que le corresponderían covariables todas iguales a 0.

En el modelo de Cox ajustamos las curvas de supervivencia teniendo en cuenta la parte paramétrica ya estimada del modelo$e^{(\Sigma_{j=1}^{p} \widehat{\beta}_j X_j)}$, con lo que el ajuste de las curvas tiene en cuenta los valores de las variables explicativas utilizadas como predictores en el modelo. A estas curvas las denominamos curvas de

supervivencia ajustadas y al igual que las curvas KM se trazan como funciones escalonadas para cada uno de los tiempos de supervivencia.

Las curvas estimadas serán:

$$\hat{S}(t, \boldsymbol{X}) = \hat{S}_{\mathbf{0}}(t)^{e^{(\Sigma_{j=1}^{p} \beta_j X_j)}} \tag{3.89}$$

En este paso del proceso de la estimación, se ajusta la parte no paramétrica relativa a la función de supervivencia basal. Sin embargo, observamos el *input* a introducir para la estimación de la función de supervivencia, depende de los tiempos de supervivencia y también de los valores de las variables explicativas del modelo.

Así, para la estimación de $\hat{S}(t, \boldsymbol{X})$debemos indicar los valores de $\boldsymbol{X}$para los cuáles queremos construir la curva de supervivencia. A no ser que estemos interesados en construir la de un perfil en concreto, una práctica habitual es elegir el valor medio (o mediana) de los predictores:

$$\hat{S}(t, \overline{\boldsymbol{X}}) = \hat{S}_{\mathbf{0}}(t)^{e^{(\Sigma_{j=1}^{p} \hat{\beta}_j \bar{X}_j)}} \tag{3.90}$$

La curva resultante sería una curva resumen de todas las posibles curvas para los diferentes valores de covariables.

Si por ejemplo, la variable $\boldsymbol{X_1}$ es una variable de exposición que puede tomar los valores 0 y 1 para los dos grupos en estudio, y queremos dibujar en el mismo

gráfico las curvas de supervivencia de los dos grupos teniendo en cuenta el resto de las variables de confusión, podemos dibujar:

$$\hat{S}\left(t, 1, \overline{X_2}, \overline{X_{3,}}, \ldots, \bar{X}_p\right) = \hat{S}_{\mathbf{0}}(t)^{e^{(\Sigma_{j=1}^{p} \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3 + \cdots \hat{\beta}_p \bar{X}_p)}} \quad (3.91)$$

y

$$\hat{S}\left(t, 0, \overline{X_2}, \overline{X_{3,}}, \ldots, \bar{X}_p\right) = \hat{S}_{\mathbf{0}}(t)^{e^{(\Sigma_{j=1}^{p} \hat{\beta}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3 + \cdots \hat{\beta}_p \bar{X}_p)}} \quad (3.92)$$

Los métodos de estimación de esta función no paramétrica que a continuación se contemplan, se basan en metodologías que maximizan una función de verosimilitud definida con las aportaciones de las curvas de supervivencia, que con lo que bajo condiciones de regularidad, las curvas siguen una distribución normal de las cual podemos calcular la esperanza y la varianza. Es por ello que podremos calcular intervalos de confianza.

Método Nelson- Aalen-Breslow

$$\widehat{H}_0(t) = \sum_{t_i \leq t} \left[\frac{m_i}{\sum_{l \epsilon R(t_i} e^{(\hat{\boldsymbol{\beta}} \boldsymbol{X}_l')}} \right] \quad (3.93)$$

Donde $\hat{S}_{\mathbf{0}}(t) = e^{[-\widehat{H}_0(t)]}$ y las curvas estimadas serán (3.89)

Método Efron

$$\widehat{H}_0(t) = \sum_{t_i \le t} \left[\sum_{j=1}^{m_i} \frac{1}{\sum_{l \epsilon R t_i} e^{(\widehat{\boldsymbol{\beta}} \boldsymbol{X}_l')} - [(j-i)/m_i] \sum_{l \epsilon U_i^*} e^{(\widehat{\beta} X_l')}} \right] \tag{3.94}$$

Donde $\hat{S}_{\mathbf{0}}(t) = e^{[-\hat{H}_0(t)]}$ y las curvas estimadas serán (3.89).

Método Kalbfleish-Prentice

$$\widehat{S_0}(t) = \prod_{t_i \le t} \alpha_i \tag{3.95}$$

Dónde α_i para $i = 1, \dots, k$ se obtienen de solucionar el sistema de k ecuaciones:

$$\sum_{j \epsilon U_i^*} \frac{e^{(\widehat{\boldsymbol{\beta}} \boldsymbol{x}_j')}}{1 - \widehat{\alpha}_i^{\, e^{(\widehat{\boldsymbol{\beta}} \boldsymbol{x}_j')}}} = \sum_{l \epsilon R(t_i)} e^{(\boldsymbol{\beta}\, \boldsymbol{X}_l')} \text{ para} i = 1, \dots, k \tag{3.96}$$

$$\sum_{j \epsilon U_i^*} \frac{e^{(\widehat{\boldsymbol{\beta}} \boldsymbol{x}_j')}}{1 - \widehat{\alpha}_i^{\, e^{(\widehat{\boldsymbol{\beta}} \boldsymbol{x}_j')}}} = \sum_{l \epsilon R(t_i)} e^{(\boldsymbol{\beta}\, \boldsymbol{X}_l')} \text{ para } i = 1, \dots, k$$

Y las curvas estimadas serán (3.89).

En particular, la solución cuando no hay empates viene dada por:

$$\widehat{S_0}(t) = \prod_{t_i \leq t} \left[\frac{e^{(\boldsymbol{\beta}\, \boldsymbol{X}_i')}}{\sum_{l \epsilon R(t_i)} e^{(\widehat{\boldsymbol{\beta}} \boldsymbol{X}_l')}} \right]^{e^{(\boldsymbol{\beta} \boldsymbol{X}_i')}} \tag{3.97}$$

3.14 Análisis por Componentes Principales

Se menciona en (halwb.uc3m, 2018), que estas técnicas fueron inicialmente desarrolladas por Pearson a finales del siglo XIX y posteriormente fueron estudiadas por Hotelling en los años 30 del siglo XX. Y se hicieron populares hasta la aparición de los ordenadores.

Para estudiar las relaciones que se presentan entre *p* variables correlacionadas (que miden información común) se puede transformar el conjunto original de variables en otro conjunto de nuevas variables incorreladas entre sí (que no tengan repetición o redundancia en la información) llamado conjunto de componentes principales. Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.

De modo ideal, se buscan $m < p$ variables que sean combinaciones lineales de las *p* originales y que estén incorreladas, recogiendo la mayor parte de la información o variabilidad de los datos. Si las variables originales están incorreladas de partida, entonces no tiene sentido realizar un análisis de componentes principales.

En resumen los componentes principales son las combinaciones lineales de las variables originales que explican la varianza de los datos.

El análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de los datos, aunque si esto último se cumple se puede dar una interpretación más profunda de dichos componentes. Ha sido una herramienta estadística ampliamente utilizada en diversas áreas del conocimiento, sobre todo en aquellas donde se tienen un volumen considerable de datos y por tanto aumenta la necesidad de conocer la estructura de los mismos y sus interrelaciones. En muchos casos los supuestos del método no se satisfacen, especialmente los relacionados con el nivel de medición de las variables y la relación lineal entre ellas (Navarro-Céspedes et al. 2010).

3.14.1 Enfoque Descriptivo

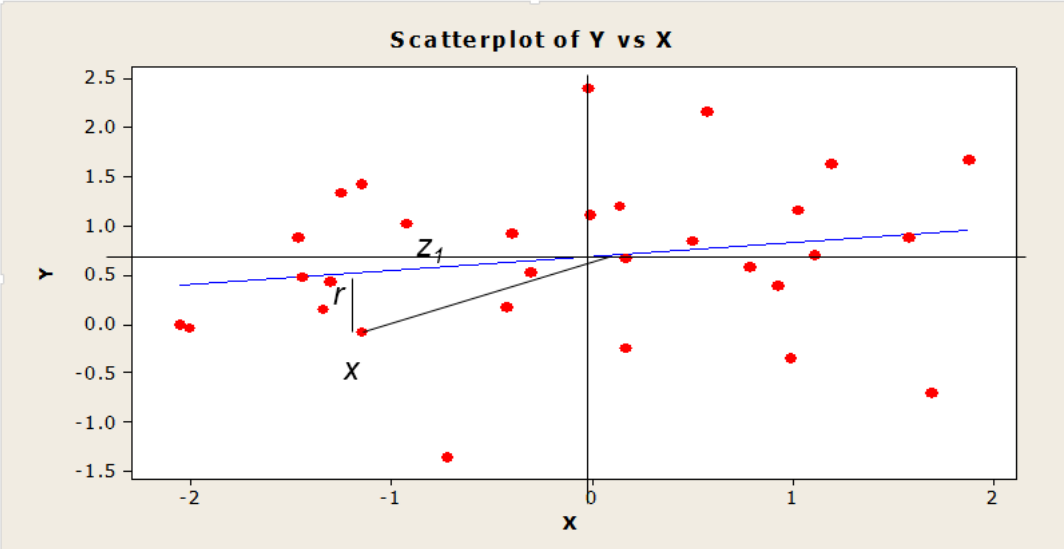


Figura 3.2 Resumen de Datos en Diagrama de Dispersión

Se menciona en (mhe.es,2018) que considerando el caso de dos dimensiones (*p = 2).* la figura 3.2 indica el diagrama de dispersión y una recta que proporciona un resumen de los datos, ya que pasa cerca de todos los puntos y las

distancias entre ellos se mantienen aproximadamente en su proyección sobre la recta.

Si consideramos un punto x_i y una dirección $a_1 = (a_{11}, \dots, a_{1p})$ definida por un vector a_1 de norma unidad, la proyección del punto sobre esta dirección es el escalar:

$z_i = a_{11}x_{i1} + , \dots + a_{1p}x_p = a_1^{'}x_i$ y el vector que representa esta proyección sera $z_i a_1$ Sea r_i distancia entre x_i

y su proyección sobre la dirección a_1 entonces:

$$minimizar \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} |x_i - z_i a_1|^2 \, x_i^{'} x_i = z_i^2 + r_i^2 \tag{3.98}$$

Por el Teorema de Pitágoras podemos escribir:

$$x_i^{'} x_i = z_i^2 + r_i^2 \tag{3.99}$$

y sumando esta expresión para todos los puntos, se obtiene:

$$\sum_{i=1}^{n} x_i' x_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2 \tag{3.100}$$

Como el primer miembro es constante, minimizar , $\sum_{i=1}^{n} r_i$, la suma de las distancias a la recta de todos los puntos, es equivalente a maximizar $\sum_{i=1}^{n} z_i$, la suma al cuadrado de los valores de las proyecciones, las cuales son variables de media cero, y maximizar la suma de sus cuadrados equivale a maximizar su varianza, y

obtenemos el criterio de encontrar la dirección de proyección que maximice la varianza de los datos proyectados.

Si en lugar de buscar la dirección que pasa cerca de los puntos buscamos la dirección tal que los puntos proyectados sobre ella conserven lo más posible sus distancias relativas llegamos al mismo criterio. Si llamamos $d_{ij}^2 = x_i' x_i$ cuadrados de las distancias originales entre puntos, y $\widehat{d_{ij}^2} = (z_i - z_j)^2$ a las distancias entre los puntos proyectados sobre una recta, deseamos que $D = \sum_i \sum_j (d_{ij}^2 - \widehat{d_{ij}^2})$ sea mínima.

3.14.2 Cálculo del Primer Componente

El primer componente principal se define como la combinación lineal de las variables originales que tiene varianza máxima. Se definen los valores en el primer componente: $z_1 = X a_1$, *Varianza* **=** $\frac{1}{n} z_1' z_1 = \frac{1}{n} a_1' X' X_{a1} = a_1' S a_1$

Para maximizar la varianza debemos imponer una restricción al módulo del vector

a_1

Restricción: $a_1' a_1 = 1$ que introduciremos mediante:

Multiplicadores de Lagrange:

$$M = a_1' S a_1 - \lambda (a_1' a_1 - 1) \qquad (3.101)$$

3.15 MTS Mahalanobis-Taguchi System

El sistema Mahalanobis Taguchi (MTS) es usado para encontrar variables significativas usando el reconocimiento de patrones en las variables que forman parte de un sistema multidimensional. Este sistema es una combinación de la distancia Mahalanobis (MD *Mahalanobis Distance*, por sus siglas en inglés) con los conceptos de la metodología Taguchi [20].

2.1 Análisis de Datos de Sistemas Multidimensionales

Ésta investigación es un estudio basado en el reconocimiento de patrones de varias variables con diferentes escalas de medición, útil para toma de decisiones.

Este reconocimiento de patrones se hace en los sistemas multidimensionales, principalmente en aquellos compuestos por factores que se pueden controlar (x1, x2, x3, …, xn) y por situaciones ajenas al experimento no controlables, llamadas condiciones de ruido.

Un sistema multidimensional puede contener un número muy alto de variables; sin embargo, no todas serán útiles para el diagnóstico de la situación, entonces, es importante identificar y elegir aquellas que tienen mayor impacto en el fenómeno que se observa, para así reducir el conjunto de factores a analizar [20].

Encontrar las variables importantes es un proceso extenso y repetitivo debido a que no se puede hacer inferencias acerca de cada variable en forma independiente, ya que éstas pudieran guardar una correlación entre sí.

A lo largo de este proceso iterativo de aprendizaje, las variables a menudo se agregan o suprimen del estudio. Por lo tanto, la complejidad de la mayoría de los

fenómenos requiere que el investigador recoja las observaciones de muchas variables diferentes [21].

2.1.1 Arreglos Ortogonales

La optimización del diseño, como parte del diseño de experimentos, implica el uso de relaciones señal a ruido, gráficas lineales y arreglos ortogonales.

Los arreglos ortogonales (OA) se aplican en la metodología Taguchi para valorar el diseño de un producto con respecto a la robustez en relación al ruido. Esto señala qué tan resistente es un producto a los efectos de los factores no controlables.

La forma convencional de llamar a los arreglos es:

$$\boldsymbol{L_a(b^c)} \quad \mathbf{(1)}$$

Donde:

a es el número de corridas.

b es el número de niveles de cada factor.

c número de columnas en el arreglo.

Un arreglo puede tener factores con muchos niveles, aunque es común que se usen factores de dos o tres niveles. Un arreglo L_8, por ejemplo, puede tener hasta siete factores en dos niveles cada uno, bajo ocho condiciones experimentales [22].

Los arreglos ortogonales se utilizan en MTS para seleccionar las variables o factores importantes que se estudia ya que reducen las diferentes combinaciones del conjunto original de variables.

2.1.2 Distancia Mahalanobis (MD)

La MD (*Mahalanobis Distance*, por sus siglas en inglés) es una medida de distancia entre variables que se dio a conocer en 1936. Es diferente a la distancia euclidiana debido a que toma en cuenta la correlación entre variables. Esta distancia es usada para determinar similitud o patrones entre dos variables aleatorias multidimensionales. Si se tienen dos variables aleatorias con la misma distribución de probabilidad $\vec{x}$ y $\vec{y}$ con matriz de covarianza Σ se define como:

$$\boldsymbol{d_m} = \sqrt{(\vec{x} - \vec{y})^T \boldsymbol{\Sigma}^{-1} (\vec{x} - \vec{y})} \quad (2)$$

Esta ecuación es sumamente sensible a la estructura de correlación del grupo de referencia. En los métodos clásicos, MD es usada para encontrar la cercanía de un punto desconocido a la media del grupo. La observación es clasificada en un punto cuyo centro tiene la menor distancia [23].

En el sistema Mahalanobis Taguchi, la MD se modifica por una escala adecuada. Esto es para definir una base o punto de referencia de la escala y medir las distancias de observaciones desconocidas desde el punto de referencia [24].

2.2 Sistema Mahalanobis Taguchi (MTS)

El MTS es usado para encontrar las variables significativas de un sistema multidimensional fusionando la distancia Mahalanobis, que ayuda a encontrar las

condiciones anormales y la correlación de las variables, con herramientas de la metodología Taguchi que ayudan a reducir el número de variables que se van a analizar, como los arreglos ortogonales y la relación señal a ruido obtenidas de las condiciones anormales.

Una de las ventajas del MTS es introducir una escala basada en todas las características de las variables para medir el grado de anormalidad. A efecto de obtener dicha escala, la distancia Mahalanobis se adecúa dividiendo la distancia original por el número de variables *n*; también minimiza el número de variables requeridas para un diagnóstico efectivo; (el cual es el objetivo principal de este análisis), predice el desarrollo de un sistema multidimensional bajo varias condiciones; establece zonas de tratamiento de un producto o paciente basado en la severidad y costo, así, el que toma la decisión pueda tomar acciones apropiadas [24].

El MTS se muestra superior sobre otros métodos como se puede ver en [25] donde se compara con el modelo Logit para datos binarios y esta muestra varias limitaciones en el análisis de datos.

Entre otros métodos se encuentra el análisis de componentes principales, que se enfoca en reducir los datos que se analizan y en su interpretación, pero cuenta con la desventaja que para el cálculo de los componentes principales se necesitan todas las variables originales, así que este método no ayuda a reducir la dimensión en términos de variables originales [20].

2.3 Cuatro Pasos en MTS

2.3.1 Creación del Espacio Mahalanobis (MS)

Se crea un conjunto de datos estándar con los sujetos saludables, definiendo las variables consideradas como condición saludable.

Calculando las distancias Mahalanobis de todas las observaciones usando la matriz de correlación definiendo así el punto cero. Este punto servirá como punto de referencia o la base para la escala de medición [20,24,26].

Las MDs del conjunto de datos saludables se calculan mediante:

$$\boldsymbol{MD_i = \frac{1}{k} z_i C^{-1} z_i^T z_i} \quad (\mathbf{3})$$

Donde:

$$\boldsymbol{Z_i = \frac{x_i - m_i}{s_i}} \quad (\mathbf{4})$$

$\boldsymbol{x_i}$ = valor de la *i*-ésima característica

$\boldsymbol{m_i}$ **=** media de la *i*-ésima característica

$\boldsymbol{s_i}$ = desviación estándar de la *i*-ésima característica

$\boldsymbol{k}$ = Número de características/variables

T = Transpuesta del vector

$\boldsymbol{C^{-1}}$ = Matriz inversa de la matriz de correlación

2.3.2 Validación de la Escala de Medición

Aquí es necesario identificar y recolectar los datos de las condiciones anormales, calcular las distancias Mahalanobis de cada observación, después se normalizan usando la media y desviación estándar de las variables usadas en el grupo de datos saludable. La matriz de correlación del grupo normal se usa para calcular las MD de las condiciones anormales. Si la escala es buena, las MD del grupo anormal tendrán valores más altos que los saludables y de esta forma la escala está validada [20,24].

2.3.3 Identificación de variables significantes

Para encontrar el conjunto de variables significantes se usan arreglos ortogonales y las relaciones señal a ruido. La relación señal a ruido se obtiene de las MD de los datos anormales, se usa como la respuesta para cada combinación de arreglos ortogonales. El conjunto de variables significantes se obtiene evaluando la ganancia de la relación señal a ruido. [20,24].

Al utilizar las relaciones en los arreglos ortogonales y la relación señal a ruido, se seleccionan características útiles. En MTS, los arreglos ortogonales se utilizan para reconocer características significativas al disminuir el número de combinaciones de características establecidas inicialmente [24].

Con esto se optimiza el sistema, en el experimento cada factor es asignado a una columna en el arreglo ortogonal (OA) y cada renglón representa la combinación del experimento a correr. Un OA de dos niveles es utilizado para representar la inclusión o exclusión. En un OA de dos niveles, el nivel 1 corresponde a la presencia de una variable y el nivel 2 corresponde a la ausencia de la variable.

Cada variable será incluida con respecto a la combinación en el OA, se obtendrán las MD y se calculará la relación *S/N* [14].

2.3.4 Relación Señal a Ruido *(S/N)*

En MTS se trabaja con una sola población separada en datos normales o saludables y datos anormales. El grupo principal de datos que se analiza se llama normal y lo que cae fuera del grupo saludable se le nombra anormal. El grado de anormalidad se mide en referencia al grupo normal.

La S/N es una medida de la funcionalidad del sistema, la cual aprovecha la interacción entre los factores de control y de ruido. En los sistemas multidimensionales esta medida se aplica para minimizar el número de variables que se analizan y juega un rol importante para predecir las condiciones anormales.

Usando la *S/N* se puede obtener un conjunto de variables útiles para toma de decisiones en varios contextos. En el área de salud, por ejemplo, ayuda a encontrar factores importantes para diferentes enfermedades debido a que puede evaluar el desarrollo de una condición con diferentes anormalidades y encontrar un conjunto ventajoso de variables [20].

Hay muchos tipos diferentes de relación *S/N*; sin embargo, MTS usa el *S/N* más grande, mejor o dinámico proporción. En el contexto de MTS, la relación *S/N* se define como la medida de precisión de predicción de escala. Refleja la gravedad de las anormalidades y la diferencia de los valores promedio de *S/N* de cada atributo cuando está incluido y excluido. La Ecuación (5) muestra la relación dinámica *S/N.*

$$S/N = \eta = 10log\left(\frac{\frac{1}{r}(S_\beta - V_e)}{V_e}\right) \quad (5)$$

Donde:

$r =$ Suma de los cuadrados debido a la señal de entrada

$$r = \sum_{i=1}^{t} M_i^2 \quad (6)$$

$S_\beta =$ Suma de cuadrados debido a la pendiente

$$S_\beta = \frac{1}{r}\sum_{i=1}^{t}(M_i y_i)^2 \quad (7)$$

$V_e =$Varianza del error

$$V_e = \frac{S_e}{t-1} \quad (8)$$

$S_e =$Suma de cuadrados del error

$$S_e = S_T - S_\beta \quad (9)$$

$S_T =$Suma total de cuadrados

$$S_T = \sum_{i=1}^{t} y_i^2 \quad (10)$$

para un atributo dado X_i, SN^+representa la relación *S/N* promedio de incluir el atributo X_i , SN^-representa cuando se excluye X_i

Evaluando la ganancia en las relaciones *S/N*, se identifican las características útiles. Con la Ecuación 11 se calcula la ganancia de cada característica. Las características con ganancia positiva se consideran útiles.

$$\boldsymbol{Gain = SN^{+} - SN^{-}} \qquad (\mathbf{11})$$

Se utilizan las variables con la ganancia más grande. Se realiza una ejecución de confirmación construyendo un espacio de Mahalanobis con las variables útiles. Las MD de las observaciones anormales también son calculadas en base al conjunto de variables útiles. El MD promedio del grupo normal se compara con el MD promedio del grupo anormal [20,24,26].

2.3.4.1 Diagnóstico futuro con variables significantes

Al encontrar las variables significantes se pueden usar para diagnóstico de situaciones futuras para la toma de decisiones, y llevar a cabo acciones correctivas o de mejora.

Una gran distancia indica una gran desviación entre las pacientes saludables y las pacientes con CaCu , el MS se reconstruye utilizando las variables registradas en el paso 3 y se calculan los MDs de las variables monitoreadas. Los sujetos están sanos si las MDs están dentro de la MS. Si las MDs están fuera del MS entonces los sujetos revelan comportamientos anormales[20].

3.15 Aplicaciones del Modelo de Riesgo proporcional de Cox

Yamasaki (2017) publicó estudio retrospectivo que investigó las características clinicopatológicas de SCLC (por sus siglas en inglés *Small Cell Lung Cáncer Células pequeñas de cáncer de pulmón*) en personas que nunca fuman. Se condujeron pruebas exactas para evaluar la correlación entre los parámetros clinicopatológicos y el historial de tabaquismo en SCLC. La supervivencia global (OS) se definió como el tiempo desde el inicio del tratamiento antitumoral hasta la fecha del último seguimiento o muerte. Las curvas de supervivencia se estimaron usando el método de Kaplan-Meier y la prueba de rango largo. Las razones de riesgo (HR) están calculadas por los métodos de riesgo proporcional de Cox.

Ocón-Hernández (2010) Describe la supervivencia global y libre de enfermedad a los cinco y diez años del diagnóstico de cáncer de mama en mujeres participantes, las cuales llevaron su caso en control previo y establece las variables pronóstico relacionadas. La estimación de la función de supervivencia se realizó mediante el método de Kaplan –Meier, calculando la supervivencia global y la supervivencia libre de enfermedad para cada una de las características de la enfermedad. Para el análisis de aquellos factores que potencialmente podrían modificar la supervivencia se empleó la regresión de Cox.

Hughes (2015) publica investigación realizada para determinar los predictores de recurrencia y patrones de fracasos entre pacientes tratados con nefroureterectomía por carcinoma urotelial de vías superiores.

Las tasas de recurrencia después de la nefroureterectomía (NU) para el carcinoma urotelial del tracto superior permanecen altas

Como tal, la terapia medica dirigida a sitios de alto riesgo puede mejorar los resultados a largo plazo. Se describen patrones y predictores de recurrencia UTUC según el paciente, factores relacionados con la enfermedad y el tratamiento.

Los datos de todos los momentos del evento se describieron con gráficos de Kaplan Meier y las diferencias entre los estratos se probaron mediante la prueba de rango largo. Se realizó un análisis de Riesgos proporcionales Cox multivariado y univariado para identificar los factores potenciales asociados con la recurrencia Las interacciones y el cumplimiento con los supuestos de riesgos proporcionales se probaron para todas las combinaciones potenciales de covariables.

Flores-Luna (2007) en investigación realizada para determinar los factores pronósticos relacionados con la supervivencia del cáncer de mama, utiliza un análisis mediante la técnica de Kaplan Meier en la cual toma en cuenta la información que proporcionan las mujeres que murieron como las de aquellas que se censuraron. Para factores pronósticos se efectuó un análisis multivariado de las variables significativas en el análisis estratificado y otras biológicamente plausibles mediante el modelo de riesgos proporcionales de Cox.

Segura-Noguera (2003) determina las características y hace análisis de supervivencia de las personas atendidas en el programa de atención domiciliaria del área básica de salud en Raval Nord, Barcelona, haciendo un análisis descriptivo de las características principales de esta población que son enfermos cónicos

incluidos en este programa. Con l modelo de riesgo s proporcionales de cox se valoraron los factores asociados con la supervivencia, y así mismo se elaboraron las correspondientes curvas de probabilidad de supervivencia acumulada utilizando el método Kaplan Meier.

Herrera-Villalobo (2007) mediante un análisis de regresión de Cox evalúa el comportamiento de los factores analizados para determinar la influencia de la demora en el diagnóstico y el tratamiento en la supervivencia de pacientes con cáncer pulmonar mientras que las curvas de supervivencia se analizaron utilizando el método Kaplan Meier.

La influencia de la temporada y la distancia a un centro de cáncer en las tasas de tratamientos de cáncer de pulmón, fueron analizadas mediante el modelo de riesgo proporcional de Cox para identificar los factores asociados con la recepción de la terapia sistémica en 2012 por Gotfrit. Publicado en 2017.

Estimación de confiabilidad para la prueba de vida acelerada basada en un modelo de riesgo proporcional de Cox con efecto de error publicado por Rodríguez-Borbón (2107).

En este estudio, se propone el modelo de riesgo proporcional de Cox con la adición de un parámetro de efecto de error bajo la intención de modelar el estrés no constante mientras se realiza la ALT. (Accelerated Life Testing) prueba de vida acelerada. Este modelo de Cox se utiliza en un ALT aplicado a un sensor de detonación (la vibración de amortiguación del bloque de cilindros se detecta como presión de vibración por el sensor de detonación). en el que se censuran algunas

observaciones en los datos Se supone que los tiempos de falla en un nivel de estrés constante pertenecen a una distribución de Weibull.

Báez (2014) Modelo de Confiabilidad Humana para trabajadores manuales en líneas de montaje, el objetivo de este estudio fué desarrollar un modelo de confiabilidad humana para descifrar el comportamiento defectuoso del operador, teniendo en cuenta el efecto del entorno operativo y el tiempo transcurrido antes de informar los errores utilizando el modelo de riesgo proporcional de Cox.

4. MATERIALES Y MÉTODOS.

En este capítulo se describen los materiales métodos adoptados para llevar a cabo el proyecto de investigación.

4.1 Materiales y Métodos

Como parte de la metodología se analizan los formatos de análisis de la última citología aplicada a pacientes de un organismo del Sector Salud, con el fin de identificar y analizar la significancia los factores que estos consideran dentro de sus programas de detección y prevención al CaCu aquí en México.

Las herramientas estadísticas utilizadas para este efecto son Regresión Lineal Múltiple para datos categóricos y el análisis de Componentes Principales.

Para el análisis estadístico se utiliza el *software* Minitab 18®.

A continuación se describe el manejo de los datos para su estudio.

a. Separación de los registros de las pacientes Saludables y No Saludables.

b. Seleccionar, como referencia, dentro de los resultados de los exámenes y utilizando las herramientas estadísticas mencionadas, los factores que de acuerdo a la literatura revisada tiene injerencia en padecer o no la enfermedad.

Tabla 4.1 Factores en análisis citológico

TABLA DE FACTORES EN ANALISIS CITOLOGICO	
FACTOR	**PARÁMETROS**
Estatus	(0)No Saludables (1) Saludables.,
Edad	Edad de la paciente
Postmenopausia	(1) Si, (2) No.
Citologia	(1) Primera vez en la vida, (2) Un año o menos (3)2 años (4)3 o más años.
Otro	(1) Si, (2) No.
Anticonceptivo Hormonal	(1) Si lo utiliza, (2) No lo utiliza
Flujo	(1)Presenta, (2) No presenta:
Histerectomía	(1) Si, (2) No.
Sangrado	(1)Presenta, (2) No presenta:
Metaplasia	(1) Si, (2) No.
Cuadros Marca	(4) Displasia moderada(NicII) (5)Displasia grave (NicIII) (6) In situ (NicIII), (7) Microinvasor (8) Invasor, (9)Adenocarcinoma, (10)Maligno no Especificado.
Virus Papiloma	(1) Si, (2) No.

La tabla 4.1 muestra los factores y parámetros con los que registran y miden en el formato de resultados de los análisis practicados a las mujeres.

4.2 Resultados Preliminares

En esta sección se presenta una parte los resultados obtenidos con las diferentes pruebas aplicadas a la información.

4.2.1 Análisis por Componentes Principales

Análisis de los valores y vectores propios de la matriz de correlación

En la tabla 4.2 se muestran los resultados de los componentes principales, las primeras cinco columnas muestran los valores propios mayores que 1 y explican el 72.4% de la variación de los datos.

Tabla 4.2 Valores Propios y Proporción de los Componentes Principales

Valor propio	3.043	1.852	1.345	1.273	1.172	.992	.758	.65	.403	.316	.187	.006
Proporción	.254	.154	.112	.106	.098	.083	.063	.054	.034	.026	.016	.001
Acumulada	.254	.408	.520	.626	.724	.807	.870	.924	.957	.984	.999	1.00

La figura 4.1 muestra la gráfica de sedimentación, compara visualmente el tamaño de los valores propios y ayuda a determinar el número de componentes con base en el tamaño de los valores propios, y aquí podemos notar el punto de inflexión que marca del primero al segundo componente y de ahí al resto. Debemos resaltar los siguientes tres componentes los cuales forman la mayor parte de la variabilidad, por lo que la línea comienza a enderezarse. El resto de estos representan una porción muy pequeña de ésta y probablemente son poco importantes.

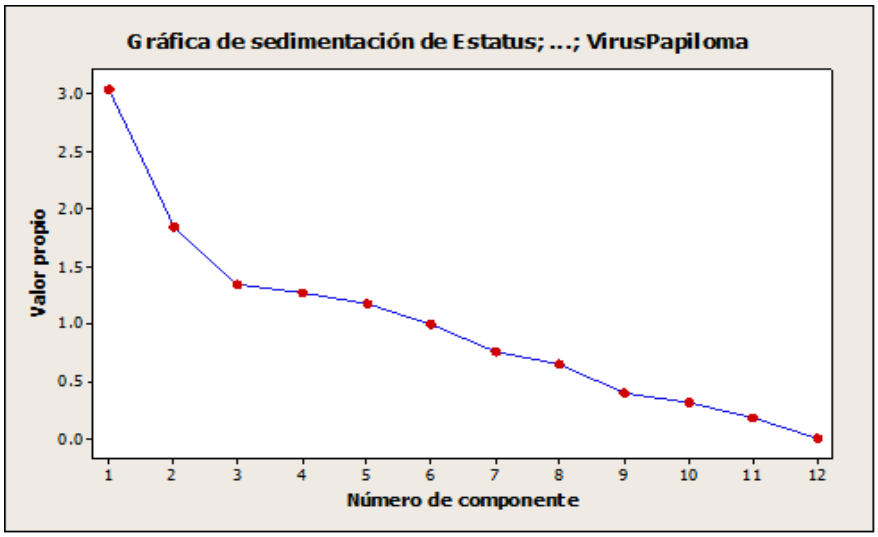


Figura 4.1 Gráfica de Sedimentación

La tabla 4.3 corresponde al cálculo de los componentes principales, donde se examina la magnitud y dirección de los coeficientes de las variables originales.

Cuanto mayor es el valor absoluto del coeficiente más importante es la variable correspondiente en el cálculo del componente.

Tabla 4.3 Cálculo de Componentes Principales

CÁLCULO DE COMPONENTES PRINCIPALES					
Variable	PC1	PC2	PC3	PC4	PC5
Estatus	-0.552	-0.092	0.025	-0.036	0.040
Edad	0.056	-0.601	-0.333	-0.014	-0.091
Ultima Citología	0.170	0.169	0.071	-0.610	0.138
Postmenopausia	-0.164	0.527	0.174	-0.247	0.207
Hormonal	0.074	-0.006	0.119	-0.278	-0.601
Histerectomía	0.082	0.184	0.331	0.575	-0.122
Otro1	0.100	-0.281	0.382	-0.017	0.530
Flujo	0.046	-0.402	0.464	-0.346	-0.085
Sangrado	0.164	0.168	-0.502	-0.162	-0.046
Metaplasia	0.188	-0.077	-0.279	0.034	0.508
CuadrosMarca	-0.550	-0.119	0.017	-0.009	0.031
VirusPapiloma	0.498	-0.011	0.196	0.101	-0.057

En el Componente Principal 1 (PC1) la relación de factores se concentra en el padecimiento de la enfermedad que corresponde al Estatus dado los Cuadros Marca, así como el Virus del Papiloma.

Para el PC2, resalta la relación entre la edad de la paciente, la postmenopausia y el flujo blanco, el cual es una excreción anormal procedente de las vías genitales de la mujer. De la misma manera, la relación en el PC3 está concentrado en el flujo y el sangrado anormal.

Para el PC4, la última citología y de manera significante contrapuesta, la histerectomía, que es la extirpación total o parcial del útero.

En el PC5 la relación de factores se concentra en los anticonceptivos Hormonales, otro tipo de anticonceptivos y Metaplasia epidermoide del cérvix o cervical, que es la transformación de un tejido adulto por otro en la capa externa de la piel del cuello uterino.

En la figura 4.2 podemos observar la estructura de los datos y la influencia del PC1 y el PC2 correspondientes al padecimiento de la enfermedad, cuadros marca y el virus de papiloma, y por otro lado los factores de influencia en el segundo componente edad, postmenopausia y flujo.

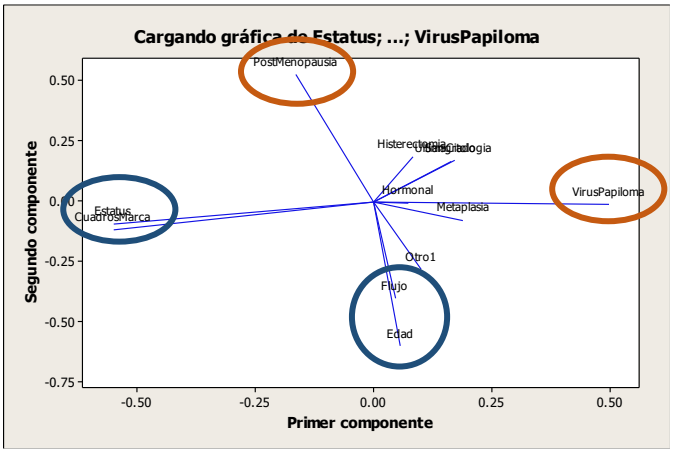


Figura 4.2 Gráfica de Estatus

La figura 4.3 muestra la gráfica de puntuación, en donde podemos destacar que las influencias que se aproximan a -1 ó 1 indican que la variable afecta considerablemente al componente. Las influencias cercanas a 0 indican que la variable tiene poca influencia. De tal manera que entonces esta gráfica divide en este caso al Componente 1 en cuanto al Estatus de las pacientes, es decir si pertenecen al grupo Saludable ó No Saludable.

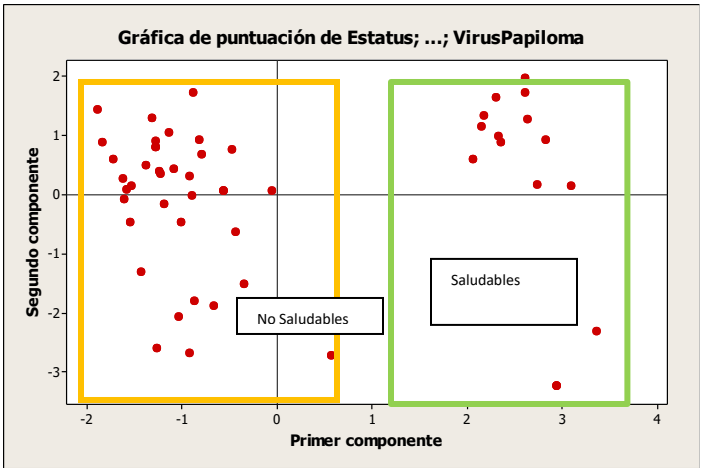


Figura 4.3 Gráfica de Puntuación

Analizando la figura 4.4; los puntos indicados pertenecen al Primer Componente, y son las pacientes No Saludables que no padecen VPH es por eso que la ausencia de esa variable los acerca a cero.

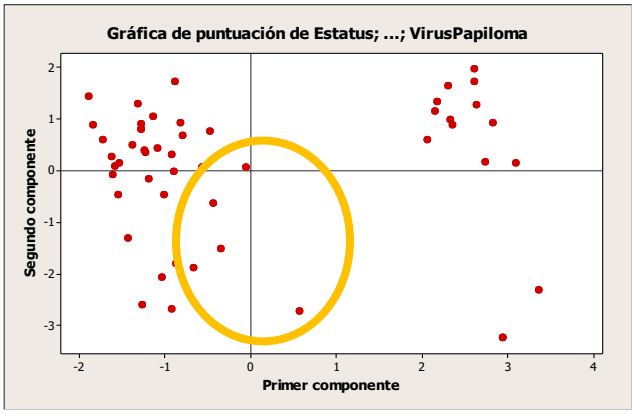


Figura 4.4 Gráfica de Puntuación Primer Componente

La figura 4.5 indica que dentro del análisis del eje del Segundo Componente donde una variable considerada es la presencia de flujo blanco, ubica a las pacientes Saludables y No Saludables cercanas a 1.

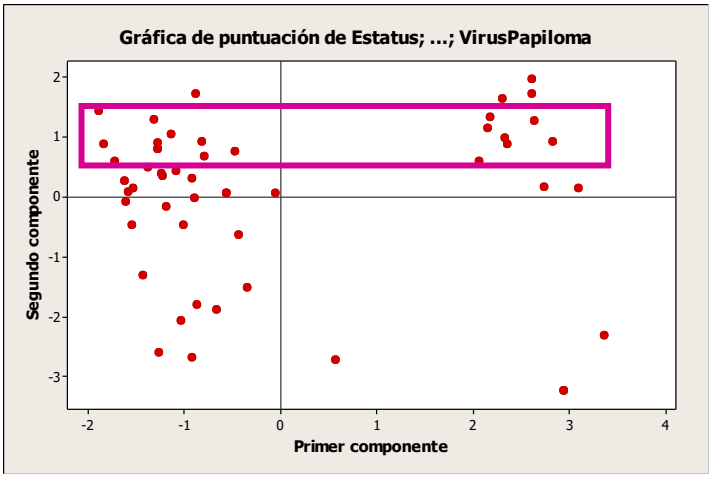


Figura 4.5 Gráfica de Puntuación Segundo Componente

La gráfica de valores atípicos representada en la figura 4.6, muestra el comportamiento lineal de los componentes y está ponderado con respecto a la variación de estos. Para este análisis los valores fuera del promedio corresponden a pacientes No saludables, las cuales una utilizaba anticonceptivo hormonal y las otras dos tenían histerectomía, esto significa que se habían retirado la matriz o parte de ella.

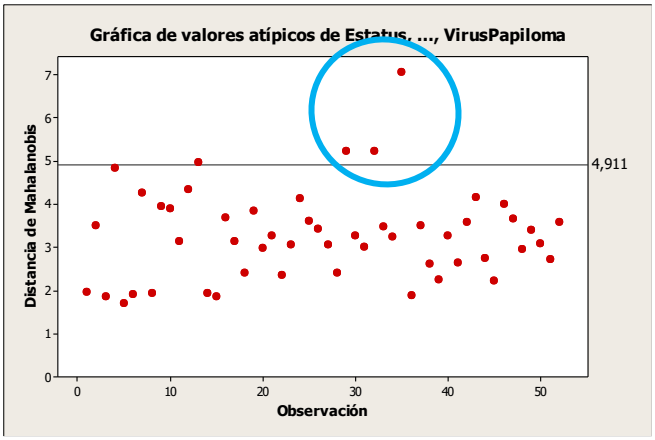


Figura 4.6 Gráfica de valores atípicos

El Análisis de Componentes Principales puede emplearse como método de selección de variables. En los datos analizados se obtiene que los principales factores están concentrados en los cuadros marca y el virus del papiloma humano y son significantes dentro del grupo de registros analizados, por otro lado nos permite identificar que los factores analizados están correlacionados, así como la variabilidad que influye para la composición de los componentes.

4.2.2. Regresión Lineal Múltiple para Datos Categóricos

A continuación, se exponen los resultados de estudiar mediante un análisis de regresión múltiple para datos categóricos la información obtenida antes mencionada.

Respecto a la construcción del contraste, el cociente entre la varianza explicada y la varianza no explicada será aproximadamente 1. Además, al utilizar una distribución F, podemos asignar una medida de probabilidad (p-value) a la hipótesis de que la varianza explicada es igual a la varianza no explicada.

Si el p-value es menor de 0.05 se acepta que el modelo de regresión es significativo; en caso contrario no podemos hablar de regresión, pues el modelo seria nulo.

Tabla 4.4 Análisis de Varianza

Codificación de predictores categóricos (1, 0)

Análisis de Varianza

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	25	10.3957	0.415826	38.97	0.000
Edad	1	0.0011	0.001066	0.10	0.754
UltimaCitologia	3	0.0304	0.010128	0.95	0.431
PostMenopausia	1	0.0000	0.000002	0.00	0.989
Hormonal	1	0.0011	0.001058	0.10	0.755
Histerectomia	1	0.0020	0.001987	0.19	0.670
Otro1	1	0.0189	0.018853	1.77	0.195
Flujo	1	0.0145	0.014476	1.36	0.255
Sangrado	1	0.1413	0.141284	13.24	0.001
Ninguno1	1	0.0037	0.003693	0.35	0.561
SeObserva	3	0.1064	0.035467	3.32	0.035
Derivada	2	0.0107	0.005370	0.50	0.610
Presencia	1	0.0307	0.030704	2.88	0.102
Metaplasia	1	0.0581	0.058084	5.44	0.028
DiaCitologico	3	0.2931	0.097689	9.16	0.000
VirusPapiloma	1	0.5061	0.506127	47.43	0.000
Bacterias	1	0.0158	0.015847	1.49	0.234
Otros	1	0.0004	0.000365	0.03	0.855
Ninguno2	1	0.1875	0.187457	17.57	0.000
Error	26	0.2774	0.010670		
Total	51	10.6731			

El análisis de varianza de la tabla 4.4 muestra que sí hay asociación entre la variable dependiente y las independientes dado el valor de p=.000 < a 0.05.

Así mismo los factores sangrado, Se observa, Diagnóstico citológico, virus papiloma y ninguno2 indican con un valor p<.05 que son de gran influencia para la variable dependiente que en este caso es categórica, Saludable, No saludable.

Tabla 4.5 Resumen del Modelo

Resumen del modelo

S	R-cuad.	R-cuad. (ajustado)	R-cuad. (pred)
0.103297	97.40%	94.90%	*

El coeficiente $R^2 = 97.40\%$ como se muestra en la tabla 4.5, determina que porcentaje dependiente de la varianza es explicado por el modelo de regresión. Esto demuestra que los predictores son muy significativos

Tabla 4.6 Ecuación de Regresión

Ecuación de regresión

```
salud = 1.569 - 0.00070 Edad + 0.0 UltimaCitologia_1 - 0.0162 UltimaCitologia_2
        - 0.0265 UltimaCitologia_3 + 0.0741 UltimaCitologia_4
+ 0.0 PostMenopausia_1
        + 0.0012 PostMenopausia_2 + 0.0 Hormonal_1 - 0.037 Hormonal_2
+ 0.0 Histerectomia_1
        - 0.064 Histerectomia_2 + 0.0 Otro1_1 - 0.0490 Otro1_2 + 0.0 Flujo_1
+ 0.0810 Flujo_2
        + 0.0 Sangrado_1 - 0.2620 Sangrado_2 + 0.0 Ninguno1_1 + 0.0486 Ninguno1_2
        + 0.0 SeObserva_1 - 0.0343 SeObserva_2 - 0.108 SeObserva_3 -
 0.2310 SeObserva_8
        + 0.0 Derivada_1 + 0.0416 Derivada_2 + 0.0217 Derivada_8
+ 0.0 Presencia_1
        + 0.195 Presencia_2 + 0.0 Metaplasia_1 - 0.289 Metaplasia_2
+ 0.0 DiaCitologico_2
        + 0.3229 DiaCitologico_3 + 0.573 DiaCitologico_4 + 0.384 DiaCitologico_9
        + 0.0 VirusPapiloma_1 - 0.5726 VirusPapiloma_2 + 0.0 Bacterias_1
+ 0.0633 Bacterias_2
        + 0.0 Otros_1 + 0.0154 Otros_2 + 0.0 Ninguno2_1 - 0.531 Ninguno2_2
```

Analizando la ecuación de regresión de la tabla 4.6 Se confirma la hipótesis alternativa del contraste de regresión en donde las variables explicativas si influyen en la variable dependiente que en este caso es la condición de estatus denominado en este análisis como No Saludable.

Tabla 4.7 Coeficientes de Regresión

Coeficientes

Término	Coef	EE del coef.	Valor T	Valor p	VIF
Constante	1.569	0.410	3.83	0.001	
Edad	-0.00070	0.00221	-0.32	0.754	3.41
UltimaCitologia					
2	-0.0162	0.0538	-0.30	0.766	1.84
3	-0.0265	0.0606	-0.44	0.665	2.08
4	0.0741	0.0520	1.42	0.166	1.89
PostMenopausia					
2	0.0012	0.0821	0.01	0.989	3.36
Hormonal					
2	-0.037	0.119	-0.31	0.755	1.29
Histerectomia					
2	-0.064	0.149	-0.43	0.670	3.98
Otro1					
2	-0.0490	0.0369	-1.33	0.195	1.64
Flujo					
2	0.0810	0.0696	1.16	0.255	4.64
Sangrado					
2	-0.2620	0.0720	-3.64	0.001	3.62
Ninguno1					
2	0.0486	0.0826	0.59	0.561	3.39
SeObserva					
2	-0.0343	0.0434	-0.79	0.436	1.81
3	-0.108	0.192	-0.56	0.579	3.39
8	-0.2310	0.0735	-3.14	0.004	2.69
Derivada					
2	0.0416	0.0416	1.00	0.327	2.00
8	0.0217	0.0519	0.42	0.679	2.58
Presencia					
2	0.195	0.115	1.70	0.102	8.36
Metaplasia					
2	-0.289	0.124	-2.33	0.028	7.63
DiaCitologico					
3	0.3229	0.0730	4.42	0.000	6.15
4	0.573	0.134	4.28	0.000	3.23
9	0.384	0.158	2.43	0.022	2.30
VirusPapiloma					
2	-0.5726	0.0831	-6.89	0.000	7.97
Bacterias					
2	0.0633	0.0519	1.22	0.234	3.11
Otros					
2	0.0154	0.0830	0.19	0.855	2.39
Ninguno2					
2	-0.531	0.127	-4.19	0.000	4.25

En la tabla 4.7 se indican los coeficientes de regresión, las variables explicativas con los más altos coeficientes son aquellos que tienen mayor influencia en la variable dependiente que para nuestro caso es No Saludable, y estos factores son: sangrado, se observa, diagnóstico citológico, virus del papiloma y ninguno.

La descripción de los factores que a continuación se exponen, están presentadas de acuerdo al detalle que expone el formato de registro del análisis. Respecto al sangrado 2, este refiere a que es un sangrado anormal; otro factor destacado es, se observa 8; indica que a la exploración realizada a la paciente se observa la apariencia del cuello uterino y la descripción 8 corresponde a que no se observa cuello, hablando del diagnóstico citológico 4, este corresponde a que el que tiene más influencia sobre la variable independiente es displasia moderada NicII; el virus del papiloma figura entre los coeficientes con más influjo, y por último el factor destacado como ninguno, este significa que el resultado del análisis citológico practicado a la paciente si presenta algún flujo, prurito y/o sangrado, de lo contrario se presentaría como ninguno 1.

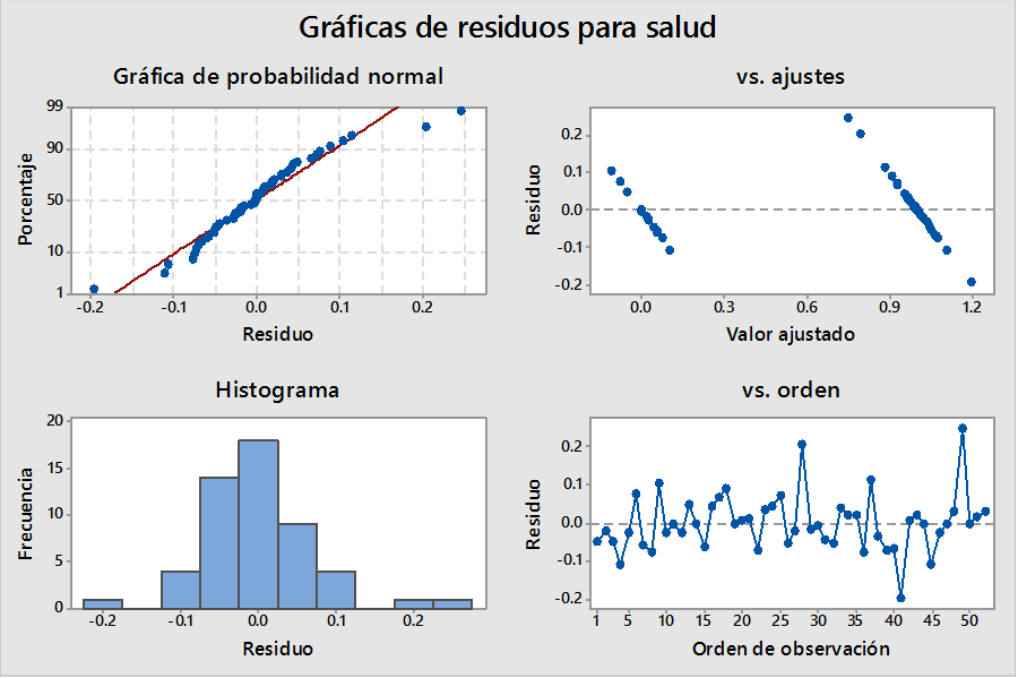


Figura 4.7 Gráficas de Residuos

A continuación se describen las gráficas contenidas en la figura 4.7.

La gráfica de probabilidad normal muestra el comportamiento de la mayoría de los datos aproximadamente normal. Y resalta los puntos no apegados a ella, que son los valores atípicos, estos valores corresponden a las evaluaciones de las mujeres que utilizan anticonceptivo hormonal y se habían practicado la histerectomía. La dispersión es normal a lo largo del recorrido.

El histograma muestra las características generales de los residuos, incluyendo los valores atípicos. Las barras alejadas de las demás así lo demuestran.

La gráfica de Residuos vs. Orden, muestra los datos en el orden que se colectaron, ayuda a revisar el supuesto que establece que los residuos no se correlacionan unos con otros.

En la gráfica Residuos vs. Ajuste se muestra un patrón aleatorio a ambos lados de cero. Los valores alejados del resto de los puntos son valores atípicos.

4.3 Análisis Preliminar de Datos Jurisdicción Sanitaria II

Se presenta un análisis preliminar efectuado a los datos obtenidos de los expedientes de las pacientes registradas y detectadas con CaCu en la Clínica de Colposcopía de la Jurisdicción Sanitaria II en Cd. Juárez Chihuahua en el período de 2013 a 2017.

La tabla 4.8 presenta los factores registrados en los diferentes formatos que contienen los expedientes, y dentro de estos, como edad, edad de la menarca, inicio de la vida sexual, cantidad de parejas y número de gestas, se hizo un análisis de correlación por cada año revisado; considerando como significante un coeficiente de correlación mayor o igual a .±0.3

Se utilizó el *software* libre R studio® para el análisis y gráficas, solo las ecuaciones de regresión se calcularon con *software* Minitab 18®.

Tabla 4.8 Factores Registrados en Período 2013-2017

FACTORES REGISTRADOS		
Edad	Ritmo menstrual	Fecha de detección
Estado civil	Inicio vida sexual activa	Fecha de atención
Religión	Cantidad de parejas sexuales	Diagnóstico inicial
Antecedentes Familiares	Número de gestas	VPH
Ocupación Esposo	Cantidad de partos	Prevención citológica
Hipertensa	Cantidad de cesáreas	Exploración
Diabetes	Cantidad de abortos	Antecedentes ETS
Tabaquismo	Síntomas	Resultado Biopsia
Tiempo tabaquismo	Metrorragia	Fecha Liberación
Cantidad cigarrillos/día	Dolor	Fecha muerte
Edad de la menarca	Tipo de anticonceptivo	

4.3.1 Grupos de Riesgo

Como resultado del análisis preliminar, la tabla 4.9 muestra los grupos de riesgo o grupo etario por período de las pacientes detectadas con CaCu en la Clínica de Colposcopía.

Tabla 4.9 Distribución de Pacientes con CaCu según grupo Etario. Jurisdicción Sanitaria II. Cd. Juárez Chih.

Grupos etarios (años)	2013	2014	2015	2016	2017
<= 25	2	0	2	0	0
26-33	9	6	8	1	2
34-41	9	7	5	3	4
42-49	8	6	5	6	7
50-57	3	2	4	5	1
58-65	3	4	3	1	0
66-73	2	0	1	0	4
>= 74 años	0	1	0	3	0
Total	36	26	28	19	18

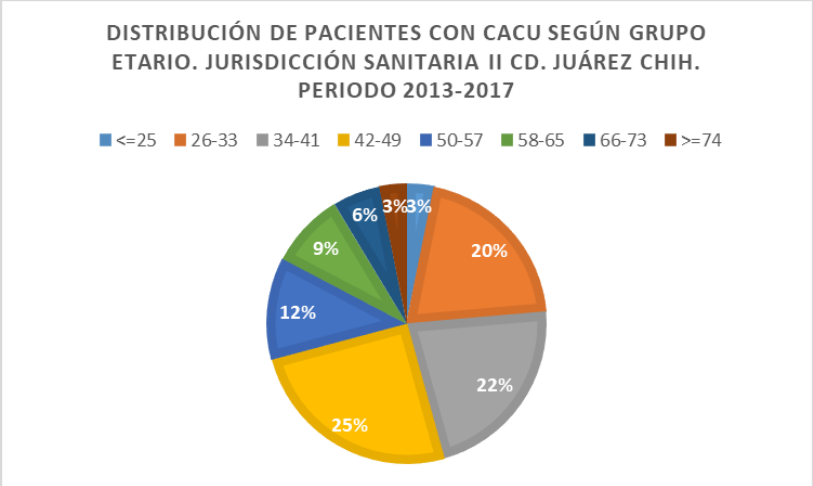


Figura 4.8 Distribución de Pacientes con CaCu según Grupo Etario. Jurisdicción Sanitario II. Cd. Juárez Chih.

En la figura 4.8 el 67% de las pacientes están en edades que van de los 26 a los 49 años, aunque debe considerarse también el 12% que representa el grupo de los 50 a 57 años.

4.3.2 Análisis Preliminar para Factores Relacionados con la Edad

El hecho de descubrir que dos variables están correlacionadas, a menudo revela un análisis de regresión que intenta describir detalladamente este tipo de relación. Un valor de correlación bajo no significa que no exista relación, simplemente no existe una relación lineal.

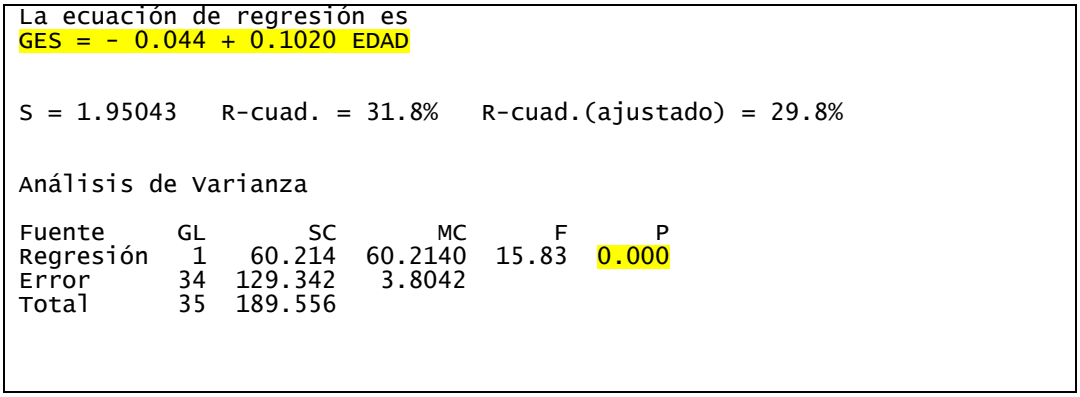
A continuación, se presentan los resultados de los análisis de correlación de los períodos del 2013 al 2017 y sus graficas de dispersión resultantes, así también las ecuaciones de regresión y los valores p.

4.3.2.1 Factores Relacionados con la Edad; Período 2013

```
datos2013<-read.csv("prueba2013.csv",head=T)
> datos22013<-datos2013[,c(2,7,8,9,10)]
> cor(datos22013)
           EDAD       MENA       IVSA        PAR        GES
EDAD  1.0000000  0.2958054  0.4325431 -0.3738984  0.5636125
MENA  0.2958054  1.0000000  0.2665249 -0.3337725  0.1870714
IVSA  0.4325431  0.2665249  1.0000000 -0.3234470 -0.1537154
PAR  -0.3738984 -0.3337725 -0.3234470  1.0000000 -0.2941698
GES   0.5636125  0.1870714 -0.1537154 -0.2941698  1.0000000

La ecuación de regresión es
IVSA = 12.34 + 0.1117 EDAD


S = 3.03616   R-cuad. = 18.7%   R-cuad.(ajustado) = 16.3%


Análisis de Varianza

Fuente     GL       SC       MC     F      P
Regresión   1   72.135  72.1350  7.83  0.008
Error      34  313.421   9.2183
Total      35  385.556
```

Figura 4.9 Correlación Período 2013

En la figura 4.9 se muestra el período 2013, los coeficientes significantes son edad vs. inicio de la vida sexual y edad vs cantidad de gestas. Respecto a la edad como predictor del inicio de la vida sexual, esta correlación tiene un coeficiente de 0.4325 y un valor p=.008, resultado del análisis de regresión, lo que nos indica que para este período las mujeres iniciaron su vida sexual a más temprana edad.



La ecuación de regresión es
GES = - 0.044 + 0.1020 EDAD

S = 1.95043 R-cuad. = 31.8% R-cuad.(ajustado) = 29.8%

Análisis de Varianza

Fuente	GL	SC	MC	F	P
Regresión	1	60.214	60.2140	15.83	0.000
Error	34	129.342	3.8042		
Total	35	189.556			

Figura 4.10 Ecuación de Regresión para Cantidad de Gestas

Para la correlación entre la edad y la cantidad de gestas, representada en la figura 4.10, con un coeficiente de 0.5636 y un valor p=.000 resultado del análisis de regresión, indican que en este año la edad es predictor para la cantidad de embarazos. Las mujeres más jóvenes tuvieron menos gestas que las mujeres de mayor edad.

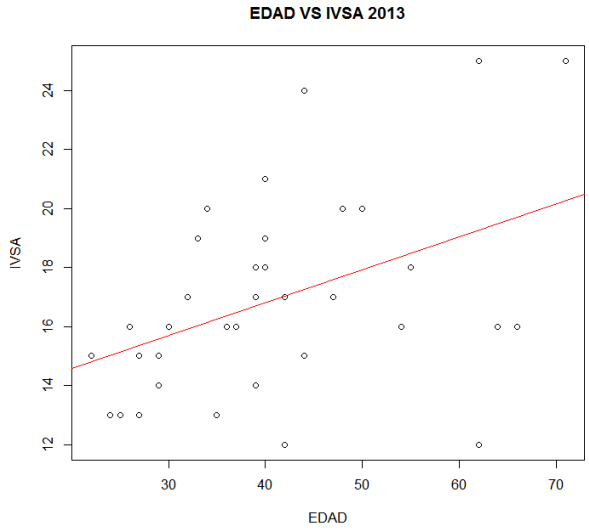


Figura 4.11 Gráfica de Dispersión Edad vs IVSA 2013

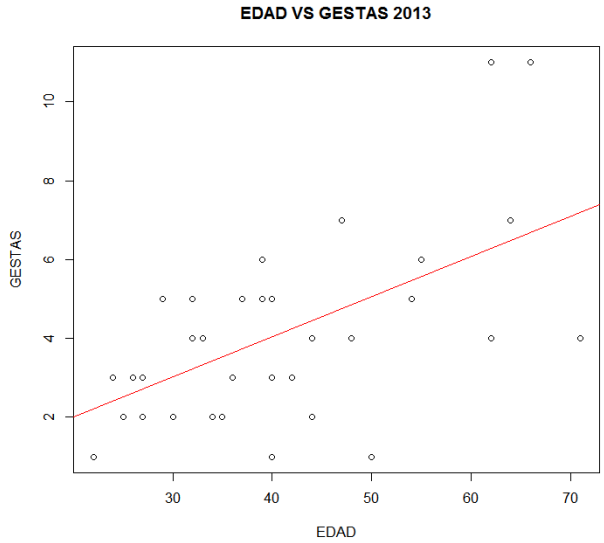


Figura 4.12 Gráfica de Dispersión Edad vs Gestas 2013

4.3.2.2 Factores Relacionados con la Edad; Período 2014

```
datos2014<-read.csv("prueba2014.csv",head=T)
> datos22014<-datos2014[,c(2,7,8,9,10)]
> cor(datos22014)
             EDAD         MENA        IVSA          PAR          GES
EDAD  1.00000000 -0.02872311 -0.1964810  0.05701899 -0.23646690
MENA -0.02872311  1.00000000  0.3348239  0.08875654 -0.07779862
IVSA -0.19648098  0.33482386  1.0000000 -0.36731624 -0.23400120
PAR   0.05701899  0.08875654 -0.3673162  1.00000000  0.07742324
GES  -0.23646690 -0.07779862 -0.2340012  0.07742324  1.00000000

La ecuación de regresión es
PAR = 6.091 - 0.1905 IVSA

S = 1.65554   R-cuad. = 13.5%   R-cuad.(ajustado) = 9.9%


Análisis de Varianza

Fuente     GL       SC        MC     F      P
Regresión   1  10.2592  10.2592  3.74  0.065
Error      24  65.7793   2.7408
Total      25  76.0385
```

Figura 4.13 Correlación Período 2014

Para este período el análisis de correlación indica con un coeficiente de -0.3673 la relación del inicio de la vida sexual con la cantidad de parejas, aunque el valor p=.065 indicador en el análisis de regresión, señala que la edad de inicio de la vida sexual no es predictor para la cantidad de parejas, sin embargo, indica que entre menos edad tengan cuando inicien su vida sexual más parejas tienen las pacientes.

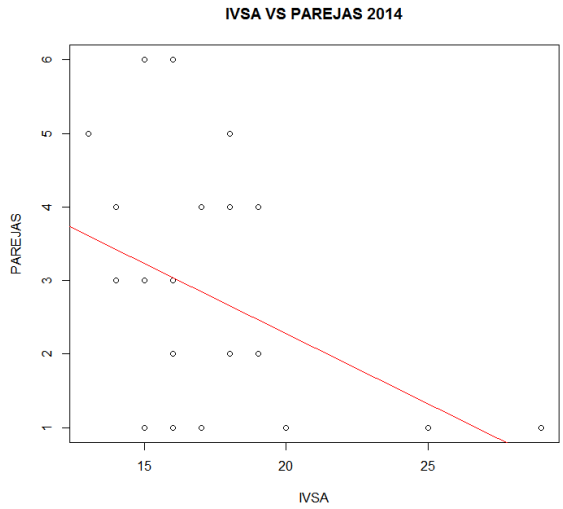


Figura 4.14 Gráfica de Dispersión IVSA vs Cantidad de Parejas 2014

4.3.2.3 Factores Relacionados con la Edad; Período 2015

```
datos2015<-read.csv("prueba2015.csv",head=T)
> datos22015<-datos2015[,c(2,7,8,9,10)]
> cor(datos22015)
             EDAD        MENA         IVSA         PAR         GES
EDAD   1.00000000  0.02810536  0.37514865 -0.3810334  0.2692284
MENA   0.02810536  1.00000000 -0.01516196  0.1059873  0.3184337
IVSA   0.37514865 -0.01516196  1.00000000 -0.2687527 -0.2290243
PAR   -0.38103340  0.10598728 -0.26875272  1.0000000 -0.1432357
GES    0.26922838  0.31843367 -0.22902425 -0.1432357  1.0000000

La ecuación de regresión es
PAR = 4.872 - 0.05232 EDAD

S = 1.68651   R-cuad. = 14.5%   R-cuad.(ajustado) = 11.1%


Análisis de Varianza

Fuente     GL       SC       MC     F      P
Regresión   1  12.0774  12.0774  4.25  0.050
Error      25  71.1078   2.8443
Total      26  83.1852
```

Figura 4.15 Correlación Período 2015

En este período solo destaca la correlación entre edad vs cantidad de parejas, con un coeficiente de -0.3810 y un valor p=.05, lo que indica que la edad de las pacientes es un predictor para la cantidad de parejas, que como indica la gráfica las mujeres mayores tienen menos parejas.

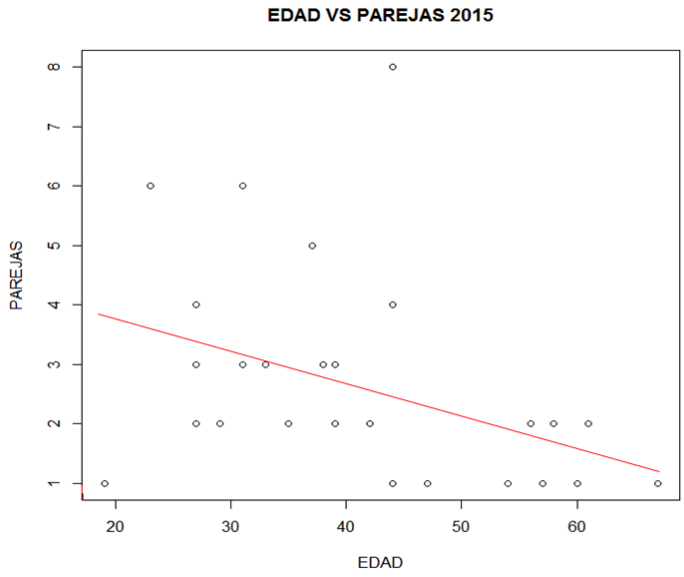


Figura 4.16 Gráfica de Dispersión Edad vs Cantidad de Parejas 2015

4.3.2.4 Factores Relacionados con la Edad; Período 2016

Para el análisis de la información correspondiente al período 2016, los factores con más alto coeficiente de correlación son edad vs cantidad de parejas con -0.6622. La gráfica de dispersión muestra una pendiente negativa, y el valor p=.002 nos indica para este período que la edad si es predictor para la cantidad de parejas. Es decir, las mujeres más jóvenes han tenido más parejas sexuales que las de mayor edad.

```
datos2016<-read.csv("Prueba2016.csv",head=T)
> datos22016<-datos2016[,c(2,7,8,9,10)]
> cor(datos22016)
            EDAD       MENA        IVSA          PAR          GES
EDAD  1.00000000  0.1936685 -0.06144871 -0.66224821  0.43936039
MENA  0.19366849  1.0000000 -0.38312391 -0.10276889  0.25250917
IVSA -0.06144871 -0.3831239  1.00000000 -0.12151457 -0.50740314
PAR  -0.66224821 -0.1027689 -0.12151457  1.00000000 -0.02818576
GES   0.43936039  0.2525092 -0.50740314 -0.02818576  1.00000000

Ecuación de regresión

 PAR  =  6.55 - 0.0775 EDAD
Análisis de Varianza

 Fuente             GL  SC Ajust.  MC Ajust.  Valor F  Valor p
 Regresión           1     18.605     18.605    13.28    0.002
   EDAD              1     18.605     18.605    13.28    0.002
 Error              17     23.816      1.401
   Falta de ajuste  14     17.316      1.237     0.57    0.798
   Error puro        3      6.500      2.167
 Total              18     42.421
```

Figura 4.17 Correlación Período 2016

Otra correlación destacada en este período es la edad vs cantidad de gestas con un coeficiente de 0.4393, con una pendiente positiva, pero con un valor p=0.60, lo que nos indica que la edad no es predictor significante para la cantidad de gestas, pero su comportamiento indica que a mayor edad de las enfermas mayor es la cantidad de hijos.

Análisis de regresión: GES vs. EDAD

Análisis de Varianza

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	1	20.56	20.564	4.07	0.060
EDAD	1	20.56	20.564	4.07	0.060
Error	17	85.96	5.057		
Falta de ajuste	14	52.46	3.747	0.34	0.933
Error puro	3	33.50	11.167		
Total	18	106.53			

Resumen del modelo

S	R-cuad.	R-cuad. (ajustado)	R-cuad. (pred)
2.24870	19.30%	14.56%	4.25%

Coeficientes

Término	Coef	EE del coef.	Valor T	Valor p	FIV
Constante	0.03	2.11	0.02	0.987	
EDAD	0.0815	0.0404	2.02	0.060	1.00

Ecuación de regresión

GES = 0.03 + 0.0815 EDAD

Figura 4.18 Análisis de Regresión para Cantidad de Gestas en el período 2016

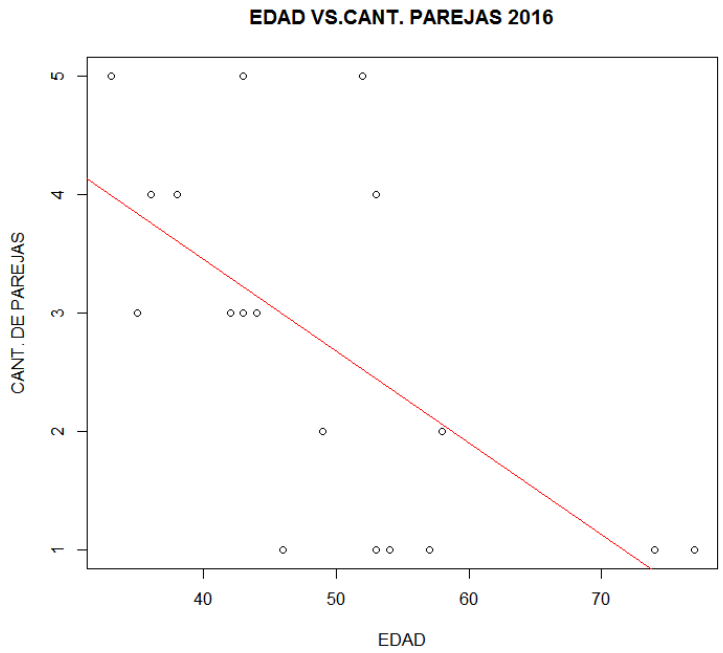


Figura 4.19 Gráfica de Dispersión Edad vs Cantidad de Parejas en Período 2016

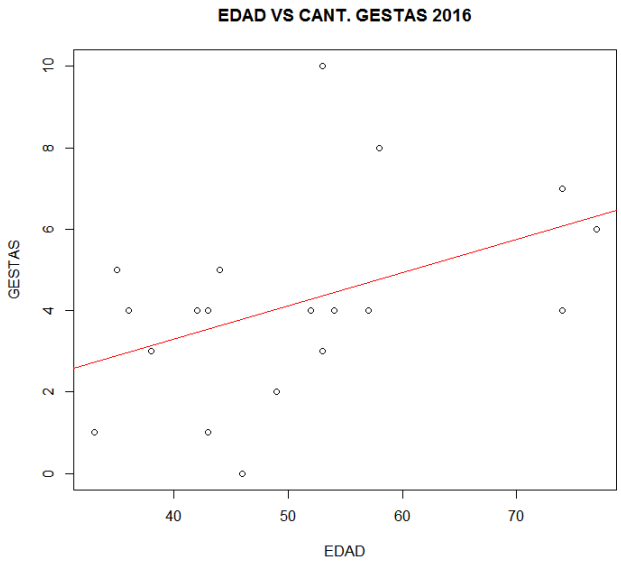


Figura 4.20 Gráfica de Dispersión Edad vs Cantidad de Gestas en Período 2016

4.3.2.5 Factores Relacionados con la Edad; Período 2017

```
datos2017<-read.csv("prueba2017.csv",head=T)
> datos22017<-datos2017[,c(2,7,8,9,10)]
> cor(datos22017)
           EDAD        MENA        IVSA         PAR         GES
EDAD  1.0000000  0.24599815  0.22987530 -0.45298883  0.34058690
MENA  0.2459981  1.00000000 -0.05080661  0.04708710  0.44634389
IVSA  0.2298753 -0.05080661  1.00000000 -0.31100362 -0.30445197
PAR  -0.4529888  0.04708710 -0.31100362  1.00000000 -0.00567733
GES   0.3405869  0.44634389 -0.30445197 -0.00567733  1.00000000
Análisis de regresión: PAR vs. EDAD
Análisis de Varianza

                                                        Valor
 Fuente              GL  SC Ajust.  MC Ajust.  Valor F      p
 Regresión            1       8.721     8.7210     4.13  0.059
   EDAD               1       8.721     8.7210     4.13  0.059
 Error               16      33.779     2.1112
   Falta de ajuste   13      32.279     2.4830     4.97  0.106
   Error puro         3       1.500     0.5000
 Total               17      42.500
Resumen del modelo

                      R-cuad.  R-cuad.
       S  R-cuad.  (ajustado)   (pred)
 1.45299   20.52%      15.55%    0.00%
Coeficientes

                    EE del
 Término      Coef   coef.  Valor T  Valor p   FIV
 Constante    5.90    1.39     4.25    0.001
 EDAD      -0.0584  0.0287    -2.03    0.059  1.00
Ecuación de regresión

 PAR  =  5.90 - 0.0584 EDAD
```

Figura 4.21 Correlación Período 2017

Para este período el coeficiente de correlación entre la edad y la cantidad de parejas es de -0.4552 con un valor *p*=.059 resultado del análisis de regresión, lo que indica que la edad pudiera ser predictor para la cantidad de parejas sexuales de las mujeres registradas, esto quiere decir, y también lo podemos observar en la figura 4.21 que las pacientes más jóvenes han tenido más parejas sexuales que las mujeres de mayor edad.

Análisis de regresión: GES vs. MENA

Análisis de Varianza

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Regresión	1	16.16	16.159	3.98	0.063
MENA	1	16.16	16.159	3.98	0.063
Error	16	64.95	4.059		
Falta de ajuste	5	22.75	4.550	1.19	0.376
Error puro	11	42.20	3.836		
Total	17	81.11			

Resumen del modelo

S	R-cuad.	R-cuad. (ajustado)	R-cuad. (pred)
2.01482	19.92%	14.92%	0.00%

Coeficientes

Término	Coef	EE del coef.	Valor T	Valor p	FIV
Constante	-2.04	3.45	-0.59	0.563	
MENA	0.529	0.265	2.00	0.063	1.00

Ecuación de regresión

GES = -2.04 + 0.529 MENA

Figura 4.22 Análisis de Regresión para Cantidad de Gestas período 2017

El coeficiente de correlación entre la edad de la menarca vs la cantidad de gestas es de 0.4463, el valor *p*=0.063 dentro del análisis de regresión nos indica que la edad de la menarca no es predictor significante para la cantidad de gestas.

La gráfica de dispersión muestra una tendencia positiva, que en este caso representa que conforme la edad de la menarca es mayor, el número de gestas se incrementa.

En los períodos de 2016 y 2017 los factores correlacionados en cuanto a edad vs cantidad de parejas sexuales, que, aunque la significancia no es la misma, esta correlación se mantiene.

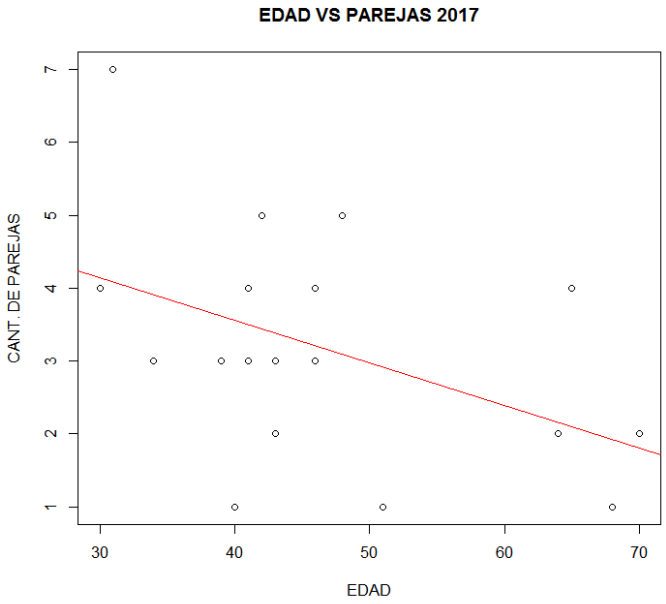


Figura 4.23 Gráfica de Dispersión Edad vs Cantidad de Parejas 2017

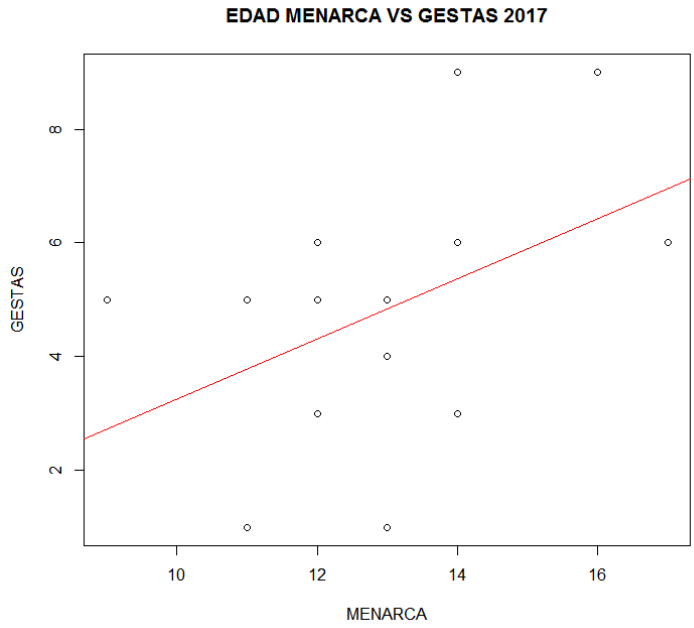


Figura 4.24 Gráfica de Dispersión Menarca vs Cantidad de Gestas 2017

4.3.3 Distribución de Estado Civil vs Edad, identificado por diagnóstico

En el siguiente análisis por período anual mediante gráficos de caja se separa a las pacientes que están diagnosticadas con cáncer en etapa In situ y las pacientes con cáncer en etapa diferente a ésta, lo que en este caso tenemos como datos binarios, relacionado con el estado civil y edad de las pacientes, el estado civil está catalogado de la siguiente manera: soltera, casada, divorciada, unión libre y viuda.

Adicional a las gráficas de caja se realizó un análisis de regresión logística, también por período anual, considerando como factores predictores de padecer la enfermedad en etapa in situ o diferente al que se denomina en este caso, como invasor, edad de la paciente, edad de la menarca y edad de inicio de vida sexual

Los gráficos como los análisis de regresión logística se programaron en el *software* libre R studio.

4.3.3.1 Distribución de Estado Civil vs Edad, Identificado por Diagnóstico; Período 2013

```
rlog<-glm(datos2013$BIO~datos2013$EDAD+datos2013$MENA+datos2013$IVSA,family = binomial)
> summary(rlog)

Call:
glm(formula = datos2013$BIO ~ datos2013$EDAD + datos2013$MENA +
    datos2013$IVSA, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9934  -0.8720  -0.8147   1.4748   1.6812

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.465430   2.666411  -0.175    0.861
datos2013$EDAD  0.009734   0.031876   0.305    0.760
datos2013$MENA  0.031101   0.190070   0.164    0.870
datos2013$IVSA -0.068316   0.125513  -0.544    0.586

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.316  on 35  degrees of freedom
Residual deviance: 43.987  on 32  degrees of freedom
AIC: 51.987

Number of Fisher Scoring iterations: 4
```

Figura 4.25 Análisis Regresión Logística Período 2013

El análisis de regresión logística, indica de acuerdo a los valores p>0.05 que estos factores no son predictores de padecer la enfermedad en etapa in situ o invasor.

```
datos2013$BIO<-as.factor(datos2013$BIO)
p2013<-ggplot(datos2013,aes(x=E.CIVIL,y=EDAD,fill=BIO))+geom_boxplot()
p2013
p2013+theme(legend.position="bottom")
```

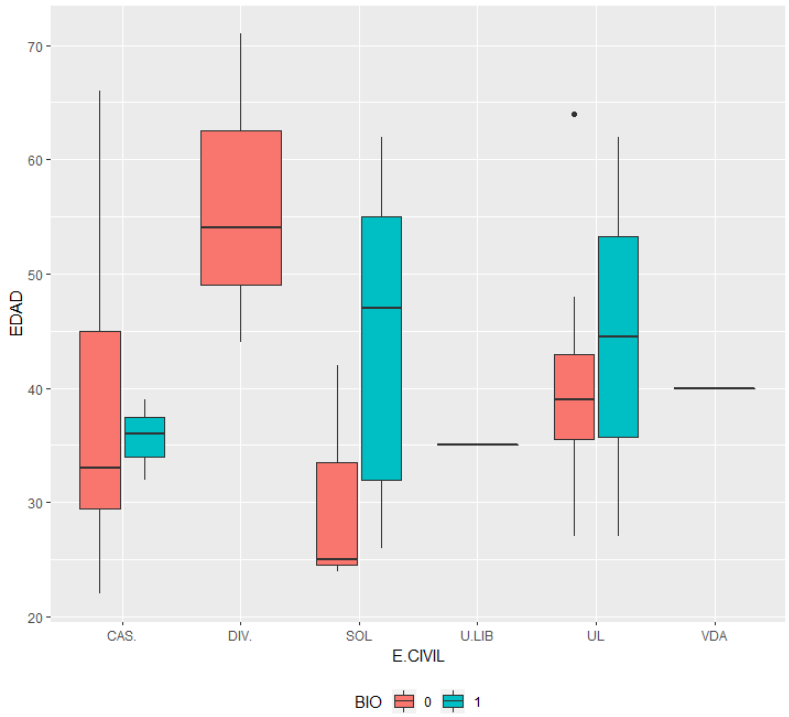


Figura 4.26 Gráfico de Caja Estado Civil vs Edad Período 2013

En la figura 4.26 hay que destacar que dentro del grupo de las pacientes divorciadas solo hay pacientes en etapa in situ, distribuida en edades de entre 44 a 71 años, así mismo se identifica solo una paciente viuda en esta misma etapa.

En el grupo de las casadas se presentan los dos estatus de la enfermedad, para el estatus in situ la edad mínima es 22 años y la máxima 66 años, con mayor dispersión que las pacientes que padecen cáncer invasor, en el cual la edad mínima es de 32 años y la máxima 39 años. Para el grupo de las solteras y similar al grupo

de las de unión libre, la mayor dispersión se presenta en las pacientes con cáncer invasor.

4.3.3.2 Distribución de Estado Civil vs Edad, Identificado por Diagnóstico; Período 2014

```
rlog2014<-
glm(datos2014$BIO~datos2014$EDAD+datos2014$MENA+datos2014$IVSA,family =
binomial)
summary(rlog2014)

Call:
glm(formula = datos2014$BIO ~ datos2014$EDAD + datos2014$MENA +
    datos2014$IVSA, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0969   0.5036   0.5478   0.6714   0.9371

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.68361    4.90684   0.139    0.889
datos2014$EDAD -0.02200    0.03816  -0.577    0.564
datos2014$MENA  0.24924    0.35694   0.698    0.485
datos2014$IVSA -0.09758    0.16213  -0.602    0.547

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.457  on 25  degrees of freedom
Residual deviance: 24.637  on 22  degrees of freedom
AIC: 32.637

Number of Fisher Scoring iterations: 4
```

Figura 4.27 Análisis de Regresión Logística Período 2014

El análisis de regresión logística representado en la figura 4.27, indica de acuerdo a los valores p>0.05 que estos factores no son predictores de padecer la enfermedad en etapa in situ o invasor.

```
datos2014<-read.csv("prueba2014.csv",head=T)
datos2014$BIO<-as.factor(datos2014$BIO)
p22014<-ggplot(datos2014,aes(x=E.CIVIL,y=EDAD,fill=BIO))+geom_boxplot()
p22014
p22014+theme(legend.position = "bottom")
```

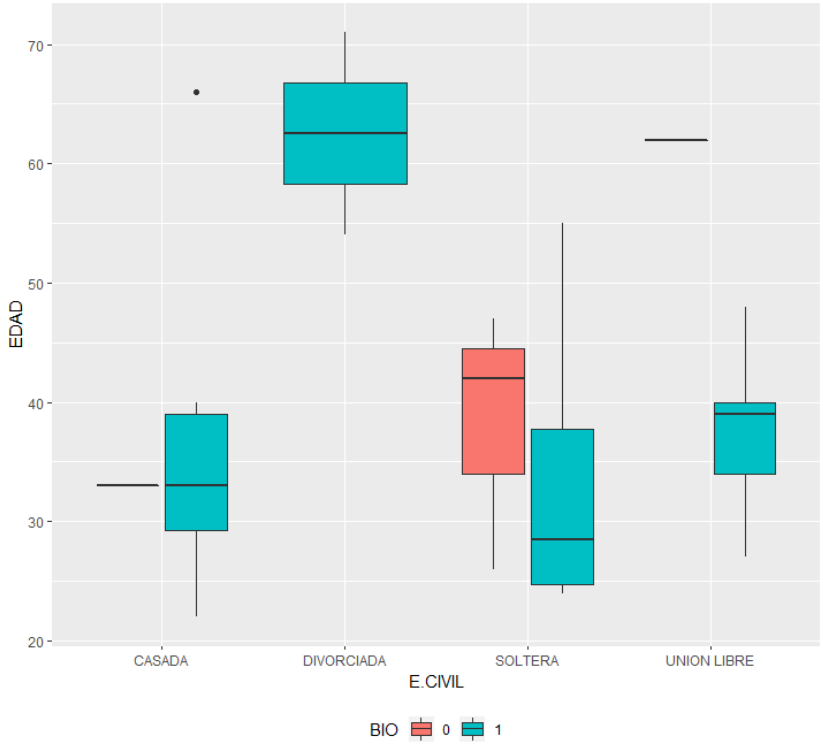


Figura 4.28 Gráfico de Caja Estado Civil vs Edad, Período 2014

En este período como se ilustra en la figura 4.28, dentro del grupo de las pacientes casadas, solo hay una diagnosticada con cáncer In situ de 34 años de edad, el resto de las pacientes de este grupo están diagnosticadas con cáncer invasor.

Destaca en el grupo de mujeres divorciadas que solo se presentan diagnosticadas con cáncer invasor, y las edades van entre los 54 años y 72 años.

El grupo de solteras presenta pacientes con los dos diagnósticos, las edades de las pacientes diagnosticadas con cáncer in situ son de 26 a 43 años; mientras que, para las diagnosticadas con cáncer invasor, las edades de 24 a 36 años.

Entre las mujeres con estado civil unión libre solo hay una mujer diagnosticada con cáncer in situ de 64 años, y el resto del grupo diagnosticado con cáncer invasor de las cuales las edades fluctúan entre los 24 y 48 años de edad.

4.3.3.3 Distribución de Estado Civil vs Edad, Identificado por Diagnóstico; Período 2015

```
rlog2015<-
glm(datos2015$BIO~datos2015$EDAD+datos2015$MENA+datos2015$IVSA,family =
binomial)
summary(rlog2015)

 Call:
 glm(formula = datos2015$BIO ~ datos2015$EDAD + datos2015$MENA +
     datos2015$IVSA, family = binomial)
 Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -2.0472  -0.8977   0.3128   0.8949   1.8547
 Coefficients:
                Estimate Std. Error z value Pr(>|z|)
 (Intercept)    -8.69552    5.76906  -1.507    0.132
 datos2015$EDAD  0.02813    0.04182   0.673    0.501
 datos2015$MENA -0.13447    0.25319  -0.531    0.595
 datos2015$IVSA  0.58296    0.36734   1.587    0.113

 (Dispersion parameter for binomial family taken to be 1)

     Null deviance: 37.096  on 26  degrees of freedom
 Residual deviance: 29.365  on 23  degrees of freedom
 AIC: 37.365
 Number of Fisher Scoring iterations: 5
```

Figura 4.29 Análisis de Regresión Logística Período 2015

El análisis de regresión logística representado en la figura 4.29, indica de acuerdo a los valores p>0.05 que estos factores no son predictores de padecer la enfermedad en etapa in situ o invasor.

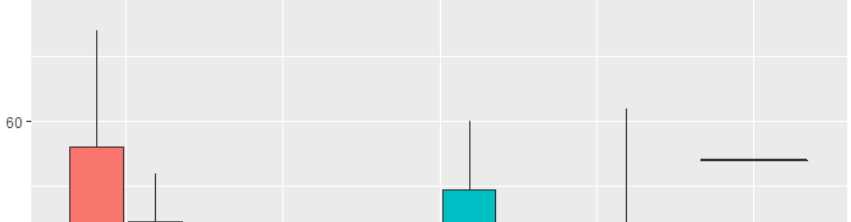


Figura 4.30 Gráfica de Caja Estado Civil vs Edad, Período 2015

En la figura 4.30 se muestra el período 2015 donde el grupo de las mujeres casadas diagnosticadas con cáncer in situ tienen edades entre 29 y 68 años y las diagnosticadas con cáncer invasor están entre 27 y 56 años de edad.

En el grupo de las divorciadas solo una mujer esta diagnosticada con cáncer in situ y tiene 31 años, mientras que el resto del grupo padecen cáncer invasor y están distribuidas entre los 26 y 44 años.

Dentro del grupo de las solteras la dispersión es menor entre las mujeres diagnosticadas con cáncer in situ, las edades dentro del grupo diagnosticado con cáncer invasor son de 38 a 60 años.

El grupo de pacientes con estado civil unión libre, las pacientes diagnosticadas en etapa in situ van de los 18 a los 44 años, y para cáncer invasor

entre los 38 y 62 años, finalmente una paciente viuda diagnosticada en etapa invasor con 57 años de edad.

4.3.3.4 Distribución de Estado Civil vs Edad, Identificado por Diagnóstico; Período 2016

```
Call:
glm(formula = datos2016$BIO ~ datos2016$EDAD + datos2016$MENA +
    datos2016$IVSA, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7290  -1.3082   0.8407   0.9812   1.2352
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.022370   5.439340   0.739    0.460
datos2016$EDAD -0.006225   0.038234  -0.163    0.871
datos2016$MENA -0.018336   0.315207  -0.058    0.954
datos2016$IVSA -0.178846   0.146195  -1.223    0.221

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 25.864  on 18  degrees of freedom
Residual deviance: 23.732  on 15  degrees of freedom
AIC: 31.732
Number of Fisher Scoring iterations: 4
```

Figura 4.31 Análisis de Regresión Logística Período 2016

El análisis de regresión logística de la figura 4.31, indica de acuerdo a los valores p>0.05 que estos factores no son predictores de padecer la enfermedad en etapa in situ o invasor.

```
library(ggplot2)
datos2016<-read.csv("prueba2016.csv",head=T)
p2016<-
ggplot(datos2016,aes(x=E.CIVIL,y=EDAD,fill=E.CIVIL))+geom_boxplot()
```

```
p2016
p2016+theme(legend.position="bottom")
```

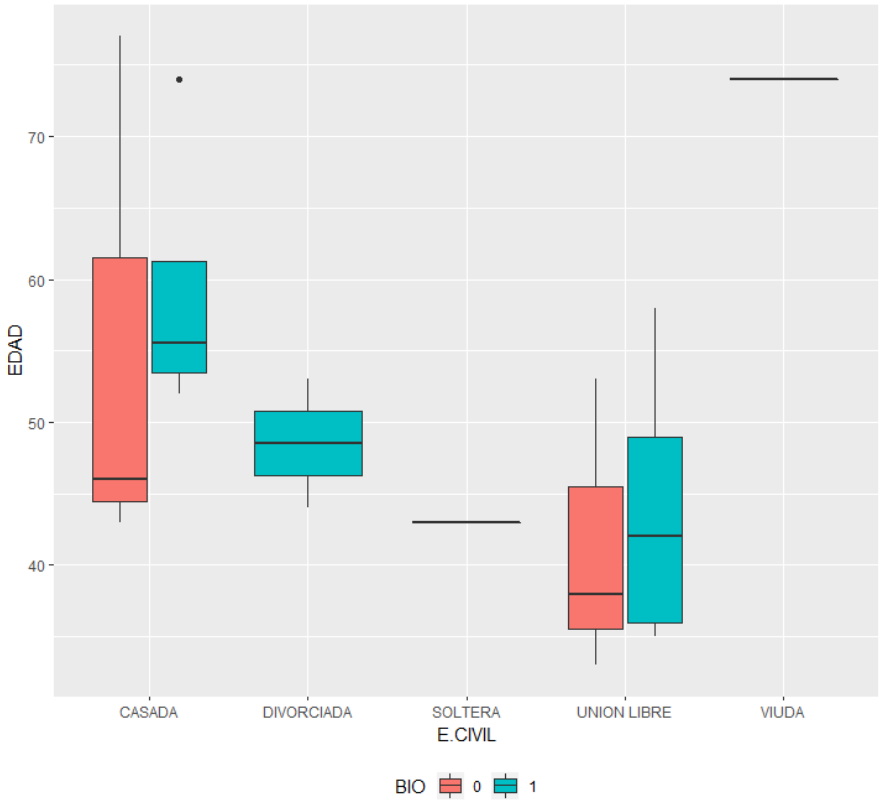


Figura 4.32 Gráfica de Caja Estado Civil vs Edad, Período 2016

La figura 4.32 muestra que, en este período, solo se presenta una paciente diagnosticada en etapa in situ para el estado civil soltera y viuda respectivamente.

Se destaca el estado civil divorciada que solo presenta pacientes diagnosticadas con cáncer invasor con edades entre 44 y 53 años.

Dentro del grupo de mujeres casadas mujeres de 43 a 78 años presentan el diagnóstico de cáncer in situ y menor dispersión con las que padecen cáncer invasor de entre 52 a 62 años de edad.

El grupo estado civil unión libre se presenta más uniforme en cuanto las edades, mujeres de entre 25 a 53 años con cáncer in situ, y de 30 a 57 años con cáncer invasor.

4.3.3.5 Distribución de Estado Civil vs Edad, Identificado por Diagnóstico; Período 2017

```
rlog2017<-glm(datos2017$BIO~datos2017$EDAD+datos2017$MENA+datos2017$IVS
A,family = binomial)
> summary(rlog2017)
Call:
glm(formula = datos2017$BIO ~ datos2017$EDAD + datos2017$MENA +
    datos2017$IVSA, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4145   0.2175   0.3531   0.4596   0.7708
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.26224   15.38837  -0.212    0.832
datos2017$EDAD -0.09986    0.08177  -1.221    0.222
datos2017$MENA  0.32105    0.75560   0.425    0.671
datos2017$IVSA  0.40366    0.65902   0.613    0.540

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 12.558  on 17  degrees of freedom
Residual deviance: 10.520  on 14  degrees of freedom
AIC: 18.52
```

Figura 4.33 Análisis de Regresión Logística Período 2017

El análisis de regresión logística representado en la figura 4.33, indica de acuerdo a los valores p>0.05 que estos factores no son predictores de padecer la enfermedad en etapa in situ o invasor.

```
datos2017<-read.csv("prueba2017.csv",head=T)
datos2017$BIO<-as.factor(datos2017$BIO)
p22017<-ggplot(datos2017,aes(x=E.CIVIL,y=EDAD,fill=BIO))+geom_boxplot()
p22017
p22017+theme(legend.position = "bottom")
```

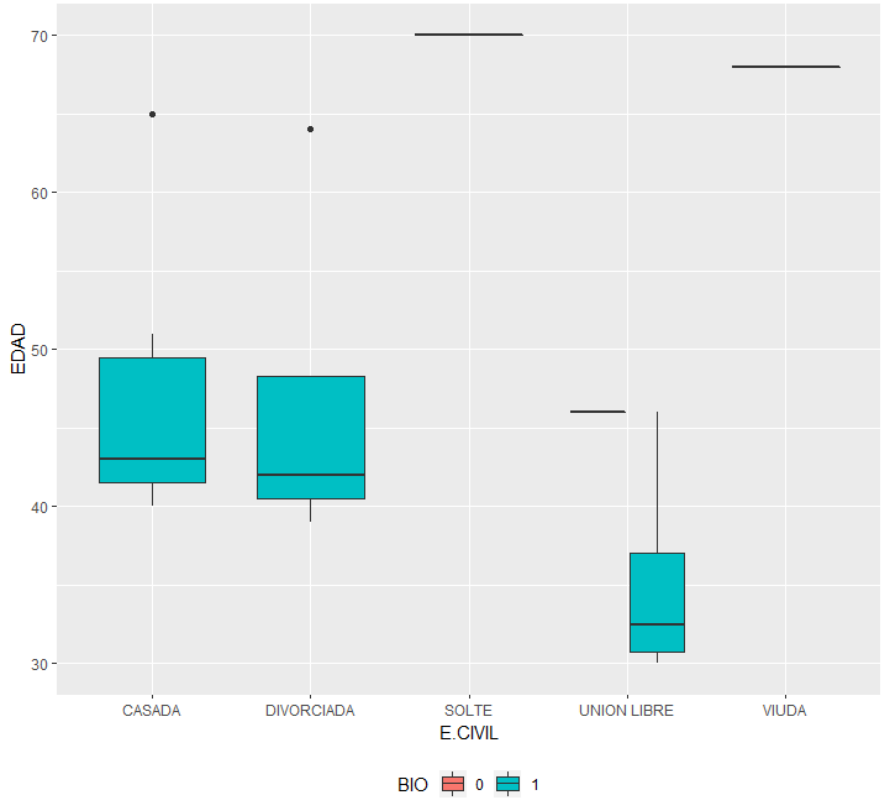


Figura 4.34 Gráfico de Caja Estado Civil vs Edad, Período 2017

La figura 4.34 muestra que para el período 2017, cabe destacar los grupos de estado civil casada, divorciada y soltera ya que únicamente presentan el estatus cáncer invasor en las pacientes diagnosticadas, las del grupo casadas van de los 40 a los 52 años, el grupo divorciadas las edades son de 39 a 49 años, muy similar a las casadas, y la soltera con 70 años; la mujer viuda esta diagnosticada con cáncer in situ y tiene 68 años.

El estatus unión libre presenta pacientes con los dos diagnósticos, pero de manera particular solo una mujer de 46 años con cáncer in situ y el resto del grupo con cáncer invasor en edades de 30 a 46 años.

4.3.4 Distribución Tipo de Anticonceptivo vs Edad Identificado por Diagnóstico

Se presenta el análisis de los gráficos de cajas en donde se muestran el tipo de anticonceptivo utilizado por las pacientes, también separadas por diagnóstico de cáncer en etapa in situ o cáncer en etapa diferente a esta, así como el grupo de edades de las mismas.

4.3.4.1 Distribución Tipo de Anticonceptivo vs Edad Identificado Diagnóstico, Período 2013

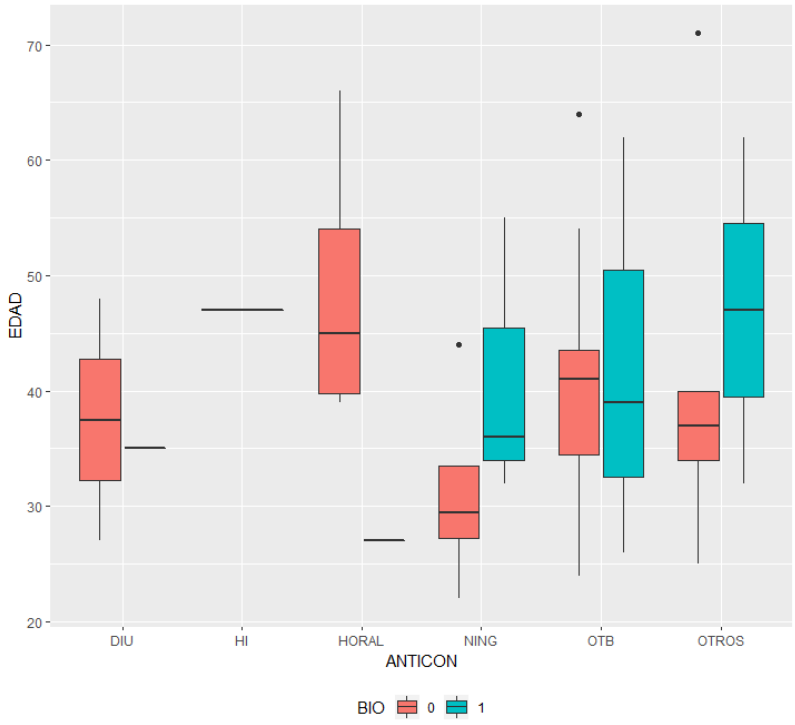


Figura 4.35 Gráfico de Caja Tipo de Anticonceptivo vs Edad, Período 2013

En la figura 4.35 podemos apreciar de manera general que en el uso de los diferentes anticonceptivos registrados en este estudio se tienen pacientes diagnosticadas con cáncer in situ e invasor, sin embargo, el hormonal inyectable solo lo utiliza una mujer de 47 años diagnosticada con cáncer invasor, el dispositivo intra uterino (DIU), presenta que es utilizado por mujeres diagnosticadas con cáncer in situ en edades de 27 a 48 años y una sola paciente de 35 años de edad. De manera semejante el anticonceptivo hormonal oral es utilizado por solo una paciente de 27 años diagnosticada con cáncer invasor, y con cáncer in situ hay mujeres de edades entre 39 a 66 años. En la categoría de las mujeres que no utilizaron ningún anticonceptivo, las diagnosticadas con cáncer in situ van de los 23 a los 33 años y las diagnosticadas con cáncer invasor de los 37 a los 55 años.

Pacientes que se practicaron la obstrucción tubaria bilateral (OTB) como anticonceptivo presenta pacientes de 24 a 54 años diagnosticadas con cáncer in situ y de 26 a 63 años con cáncer invasor.

Y en el grupo de otro tipo de anticonceptivos utilizados por las pacientes, las edades de las mismas dentro del grupo de diagnóstico de cáncer in situ van de los 25 a 40 años y para el cáncer invasor de los 33 a los 63 años.

4.3.4.2 Distribución Tipo de Anticonceptivo vs Edad Identificado Diagnóstico Período 2014

```
datos2014<-read.csv("prueba2014.csv",head=T)
datos2014$BIO<-as.factor(datos2014$BIO)
p22014<-ggplot(datos2014,aes(x=ANTICON,y=EDAD,fill=BIO))+geom_boxplot()
p22014
p22014+theme(legend.position = "bottom")
```

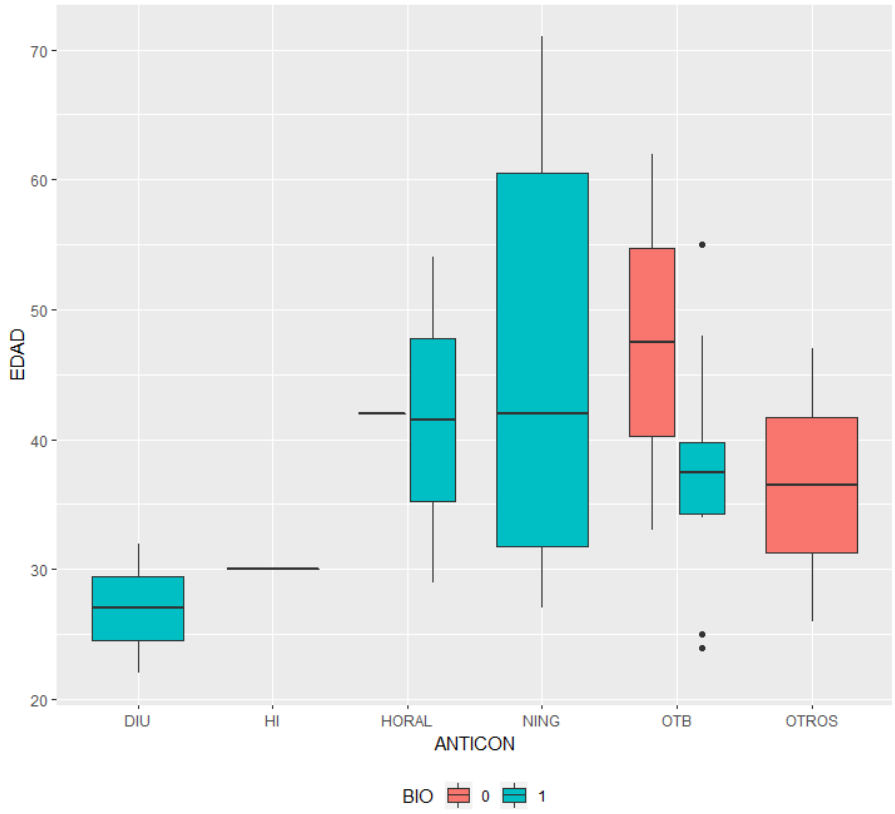


Figura 4.36 Gráfico de Caja Tipo de Anticonceptivo vs Edad, Período 2014

Se muestra en la figura 4.36, donde cabe destacar que los anticonceptivos DIU, hormonal inyectable y quienes no utilizaron anticonceptivo solo hay pacientes con diagnóstico de cáncer invasor con las siguientes edades; el DIU fue utilizado por mujeres de 23 a 33 años, el hormonal inyectable por una sola mujer de 30 años y ningún anticonceptivo por mujeres de 27 a 72 años.

Las mujeres que se practicaron OTB diagnosticadas con cáncer in situ están en edades de 33 a 63 años y las afectadas por cáncer invasor fluctúan entre los 44 y 47 años de edad.

Para los anticonceptivos hormonales orales su uso se presentó en una mujer diagnosticada con cáncer in situ de 42 años y para las diagnosticadas con cáncer invasor que utilizaron este tipo de anticonceptivo las edades van de los 28 a 54 años de edad.

Mujeres con edades entre 26 y 47 años de edad diagnosticadas con cáncer in situ, utilizaron el anticonceptivo catalogado como otro.

4.3.4.3 Distribución Tipo de Anticonceptivo vs Edad Identificado Diagnóstico, Período 2015

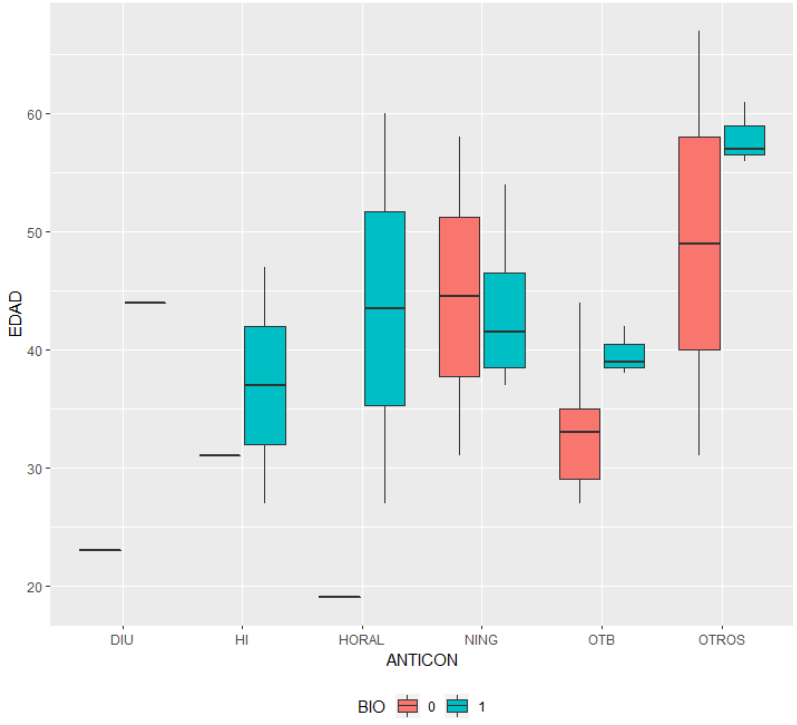


Figura 4.37 Gráfico de Caja Tipo Anticonceptivo vs Edad, Período 2015

Muestra la figura 4.37 que el uso del DIU solo se presenta en un caso de cáncer in situ para una mujer de 23 años y uno en cáncer invasor, para este caso la edad de la mujer es de 44 años.

Del anticonceptivo hormonal inyectable tenemos una mujer con cáncer in situ de 31 años de edad, y para el cáncer invasor, el período de las edades va de los 37 a 47 años de edad.

El uso de anticonceptivo hormonal oral se presentó en una mujer diagnosticada con cáncer in situ con edad de 18 años contra un grupo de edad de los 27 a 60 años con cáncer invasor.

Ningún tipo de anticonceptivo fue utilizado en mujeres con diagnóstico de cáncer in situ con edades entre los 32 y 58 años de edad, muy similar al grupo de edades de las pacientes con cáncer invasor en el mismo grupo de ningún tipo de anticonceptivo, cuyas edades son de 38 a 53 años de edad.

La OTB fue el anticonceptivo de un grupo de mujeres diagnosticadas con cáncer in situ cuyas edades van de los 27 a 44 años de edad contra las diagnosticadas con cáncer invasor que van de los 37 a 42 años.

El señalado como otro tipo de anticonceptivo concentra en el diagnóstico in situ a mujeres entre los 32 y 67 años de edad, así como a otro grupo de entre 56 y 62 años diagnosticadas con cáncer invasor.

4.3.4.4 Distribución Tipo de Anticonceptivo vs Edad Identificado Diagnóstico, Período 2016

```
library(ggplot2)
datos2016<-read.csv("prueba2016.csv",head=T)
datos2016$BIO<-as.factor(datos2016$BIO)
p22016<-ggplot(datos2016,aes(x=ANTICON,y=EDAD,fill=BIO))+geom_boxplot()
p22016
p22016+theme(legend.position = "bottom")
```

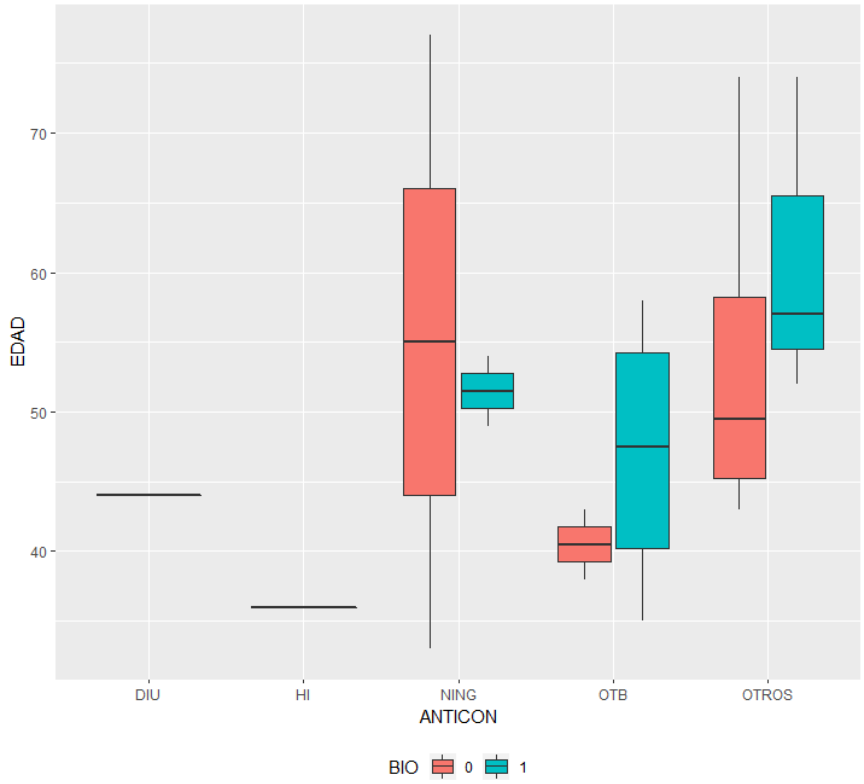


Figura 4.38 Gráfico de Caja Tipo Anticonceptivo vs Edad, Período 2016

La figura 4.38 presenta al período 2016 con diagnóstico de cáncer invasor para paciente de 48 años de edad que utilizó DIU y otra única de 32 años en el grupo de anticonceptivo hormonal inyectable.

La dispersión de edades en las pacientes que se diagnosticaron con cáncer in situ que no utilizaron ningún tipo de anticonceptivo va de 25 a 78 años, en contraste con las diagnosticadas con cáncer invasor con edades entre 50 y 53 años.

Las edades de las mujeres que se practicaron la OTB como método anticonceptivo y que fueron diagnosticadas con cáncer in situ son de 38 a 42 años mientras que las diagnosticadas con cáncer invasor con este mismo método anticonceptivo van de 30 a 58 años.

La utilización de otro tipo de anticonceptivos concentra un grupo de pacientes diagnosticadas con cáncer in situ en edades de 43 a 74 años y otro grupo de las diagnosticadas con cáncer invasor de 52 a 74 años de edad.

4.3.4.5 Distribución Tipo de Anticonceptivo vs Edad Identificado Diagnóstico, Período 2015

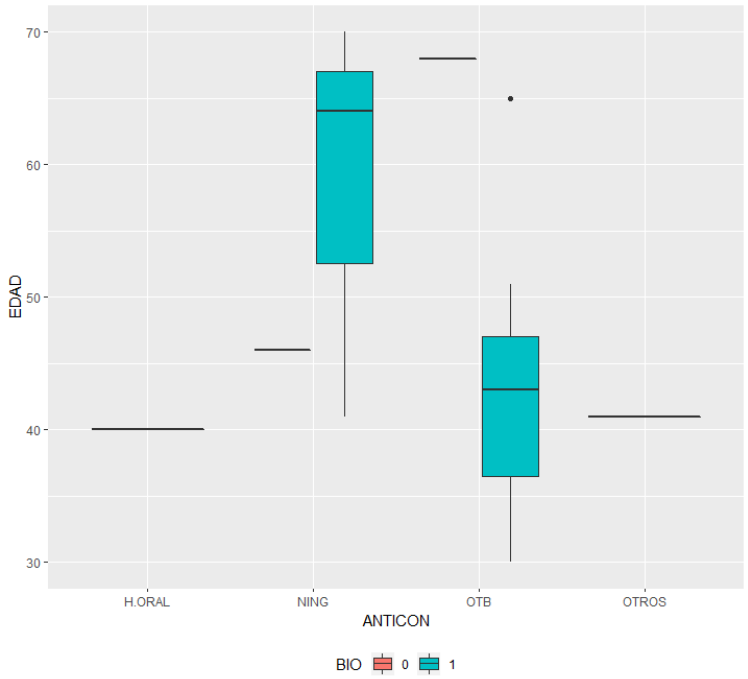


Figura 4.39 Gráfico de Caja Tipo de Anticonceptivo vs Edad, Período 2017

Presentado en la figura 4.39 se explica que los anticonceptivos hormonales orales y otro tipo de anticonceptivo solo presentan a una paciente diagnosticadas con cáncer invasor de 40 y 43 años respectivamente.

El uso de ningún anticonceptivo y la OTB agrupan los dos diagnósticos, sin embargo, para los casos de in situ solo tiene una paciente, para ningún anticonceptivo es de 46 años y con OTB 68 años. Los grupos de edad para el diagnóstico cáncer invasor las edades para las que no utilizaron anticonceptivo van de 42 a 70 años y las que se practicaron OTB de 30 a 53 años.

4.3.5 Análisis Factor Oficio del Esposo

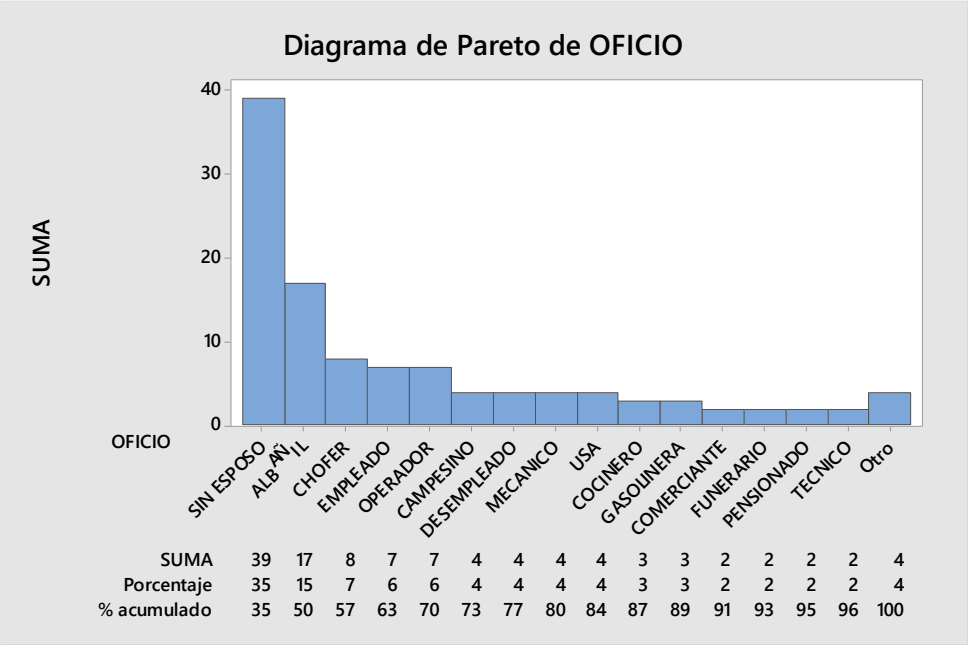


Figura 4.40 Diagrama de Pareto para Oficio del Esposo

Dentro de los factores registrados, en torno a la vida sexual de las pacientes, están la cantidad de parejas sexuales, la edad de inicio de la vida sexual y de forma adyacente referente a las parejas de ésta y cuya información se puede obtener de la fuente primaria, es el oficio de la pareja de la paciente.

Dentro de la población de información analizada, se tiene el estado civil de las pacientes enfermas, catalogado de la siguiente manera: soltera, casada, viuda, divorciada y unión libre, y esto relacionarlo con la actividad de la pareja de la paciente, sin embargo, la declaración de las mujeres con estatus soltera, viuda, o divorciada es sin esposo, ya que ellas mismas así lo expresan, sin tomar en cuenta el que hayan estado casadas anteriormente, en el caso de las viudas y divorciadas.

El gráfico de la figura 4.40 indica los diferentes oficios de las parejas de las mujeres, afectadas con CaCu en el período del 2013 al 2017, destaca el porcentaje de las mujeres que no tienen esposo impactando con un 35% del total, luego el oficio albañil y chofer que juntos representan el 50%.

Dados los resultados de este análisis es menester considerar como factor importante el estado civil de las mujeres afectadas, y mencionar también que las campañas de prevención y protección no solo se haga hincapié en evitar embarazos, si no evitar enfermedades de transmisión sexual.

4.3.6 Análisis de Regresión Logística para pacientes con CaCu vs. diagnóstico NIC I

Considerando como paciente saludable a una mujer diagnosticada con NIC I, con historia clínica documentada con mismos factores que una paciente no saludable, es decir con diagnóstico de CaCu, que es el caso de estudio; se presenta un análisis de regresión logística considerando como pacientes saludables, con categoría 1, diagnosticadas con NIC I y pacientes diagnosticadas con CaCu como paciente no saludables, con categoría 0; tomando los factores de edad de la menarca, IVSA, cantidad de parejas sexuales, prevención citológica y tabaquismo, este último también asumido como variable categórica.

Tabla de Desviaciones para Análisis de Regresión Logística

Tabla de desviaciones

Fuente	GL	Desv. ajust.	Media ajust.	Chi-cuadrada	Valor p
Regresión	5	3.4393	0.68785	3.44	0.633
MENARCA	1	0.1695	0.16950	0.17	0.681
IVSA	1	0.0078	0.00780	0.01	0.930
PAREJAS	1	0.0440	0.04402	0.04	0.834
PREV.CITOLOGIA	1	2.8489	2.84893	2.85	0.091
TABAQUISMO	1	0.9068	0.90677	0.91	0.341
Error	52	74.4650	1.43202		
Total	57	77.9043			

En la tabla se muestra que el valor correspondiente a la prevención citológica, es el único significante en este análisis, lo que nos indica que es determinante el que las pacientes se realicen la prueba anual de Papanicolaou como prevención al cáncer.

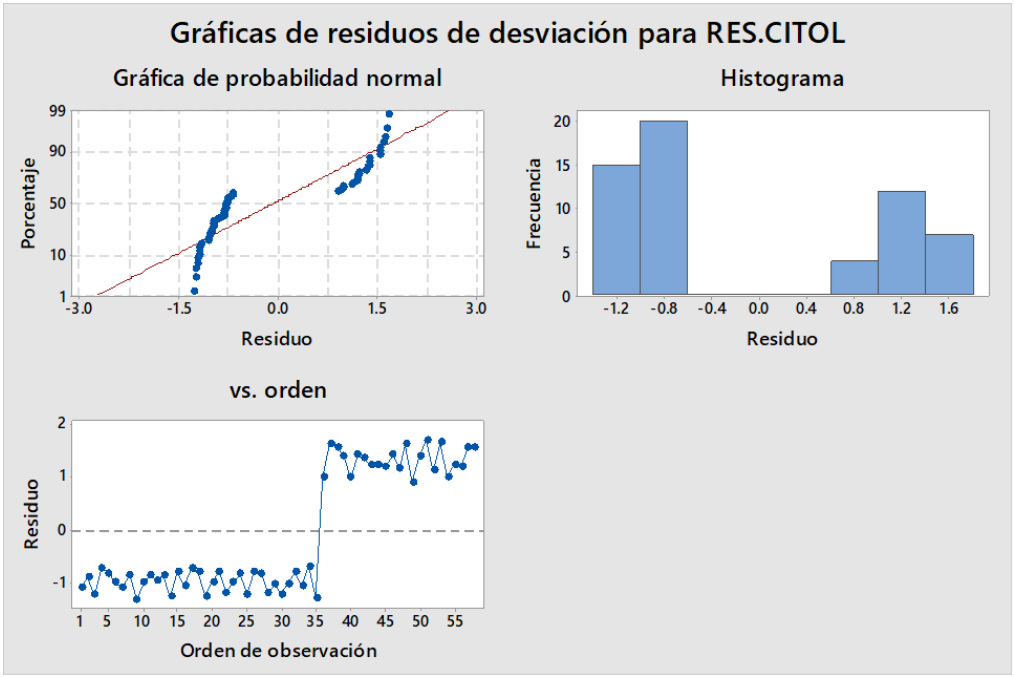


Figura 4.41 Gráfica de Residuos para Resultado de Citología

4.3.7 Análisis de los Factores para Modelar la Supervivencia de CaCu

En este apartado se estiman los parámetros de la regresión de Cox y se determina su significancia, así mismo se prueba si se cumple con el supuesto de riesgos proporcionales.

4.3.7.1 Residuos de Schoenfeld

Los residuos de Schoenfeld se definen como el valor de la covariable para los individuos que no sobrevivieron, menos su valor esperado. Este estadístico se usa para probar la independencia entre los residuos y el tiempo, y para probar el supuesto de riesgos proporcionales en el modelo de Cox.

Este procedimiento es análogo a probar si la pendiente de los residuos escalados en el tiempo es cero, hay proporcionalidad. Si la pendiente no es cero se ha violado el supuesto o si en la gráfica de los residuos de Schoenfeld contra el tiempo se muestra un patrón no aleatorio, se ha violado el supuesto de riesgos proporcionales.(https://www.mbaskool.com/business-concepts/statistics/8766-schoenfeld-residuals-test.html rescatado 8 de Mayo 2019)

Boj del Val en 2017 publica que para la realización de los residuos se debe: calcular los residuos de Schoenfeld del modelo de Cox; ordenar los tiempos de falla etiquetando el orden a cada tiempo; calcular la correlación entre los residuos y la variable de orden creada. Y se realiza el contraste de si $Ho: p = 0$ para cada covariable por separado. En el caso de aceptar la hipótesis nula de que la correlación es cero, se cumplirá la hipótesis de riesgos proporcionales para el predictor correspondiente. Con lo que, tal y como se ha mencionado, interesa que

los p-valores del contraste sean elevados para aceptar la hipótesis de proporcionalidad.

Por otro lado, Ayala en 2007 confirma que si se cumple la hipótesis de riesgos proporcionales los residuos debieran agruparse de forma aleatoria a ambos lados del valor de 0 del eje Y, y la curva ajustada debería ser próxima a una línea recta.

Boj del Val también recomienda utilizar conjuntamente los métodos gráficos y las pruebas de proporcionalidad, ya que por un lado estas son más objetivas y, por otro los gráficos pueden mostrar desviaciones concretas e información acerca del no cumplimiento de la hipótesis de riesgos proporcionales.

4.3.7.2 El contraste de la razón de verosimilitud

En este contraste se utiliza el valor de la función de verosimilitud parcial evaluada en $\widehat{\boldsymbol{\beta}}$, $L(\widehat{\boldsymbol{\beta}})$, y evaluada en $\boldsymbol{\beta_0}$, $\boldsymbol{L(\beta_0)}$:

$$X_{LR} = 2\left(\log L(\widehat{\boldsymbol{\beta}}) - \log L(\boldsymbol{\beta_0})\right) \qquad (\ )$$

Que bajo la hipótesis nula sigue una distribución X^2 con p grados de libertad.

En otras palabras, permite verificar si los factores son independientes uno de otro.

4.3.7.3 El contraste de Wald

Este contraste se basa en que los coeficientes $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \boldsymbol{K}, \hat{\beta}_p)$ siguen una distribución aproximadamente normal de media $(\hat{\beta}_1, \boldsymbol{K}, \hat{\beta}_p)$

Y matriz de varianzas y covarianzas $\widehat{\sum} = \mathbf{I}^{-1}(\widehat{\boldsymbol{\beta}})$. El estadístico se define como

$$X_W = (\widehat{\boldsymbol{\beta}} - \beta_0)^T \ \mathbf{I}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \qquad (\ )$$

Que bajo la hipótesis nula sigue una distribución X^2 con p grados de libertad .

Es un contraste de hipótesis donde se trata de ver la coherencia de afirmar un valor concreto de un parámetro de un modelo probabilístico, una vez que se tiene ya un modelo previamente seleccionado y ajustado

Se utiliza para contrastar si es cero o no un coeficiente que multiplica a una variable independiente en una regresión. $p<0.05$ se rechaza la $H_0: \beta = 0$

Si $p>0.05$ el valor del coeficiente podría ser cero, por lo que esa variable no influye a la hora de determinar la variable dependiente.

El contraste de Wald tiene una interpretación más directa que el contraste de verosimilitud y el del score, sin embargo, no es invariante ante diferentes parametrizaciones y los otros dos sí. Con el contraste del score sólo hace falta maximizar bajo la hipótesis nula, con lo que si hay que realizar el test para varios parámetros es más rápido computacionalmente. Sin embargo, el test de la máxima verosimilitud converge más rápido hacia la distribución normal. Ante la duda de cuál utilizar, es recomendable decantarse por el test de la máxima verosimilitud. (Boj del Val, 2017).

4.3.7.4 Logrank test

La prueba Log-Rank se utiliza para comparar las curvas de supervivencia de dos o más grupos. En esta prueba, la hipótesis nula es que no hay diferencia en la supervivencia de los grupos. Es una prueba no paramétrica ya que no hace suposiciones acerca de las distribuciones. La distribución del estadístico de prueba es aproximadamente chi-cuadrada. La prueba compara el número observado de eventos en cada grupo contra lo que se esperaría si la hipótesis nula fuera cierta (es decir, si las curvas de supervivencia fueran idénticas. (Reyes,2019). En el punto 3.6.4.7 se expone de forma teórica este contraste.

4.3.7.5 Análisis Kaplan-Meier para pacientes con CACU de la Jurisdicción Sanitaria II

A continuación, se presenta el cálculo de la curva de supervivencia con el modelo Kaplan-Meier utilizando el software R.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
24	127	1	0.992	0.00784	0.977	1.000
31	126	1	0.984	0.01105	0.963	1.000
40	125	1	0.976	0.01348	0.950	1.000
71	124	1	0.969	0.01550	0.939	0.999
87	123	1	0.961	0.01726	0.927	0.995
123	122	1	0.953	0.01883	0.917	0.990
168	121	1	0.945	0.02025	0.906	0.985
187	120	1	0.937	0.02156	0.896	0.980
198	119	1	0.929	0.02277	0.886	0.975
227	118	1	0.921	0.02390	0.876	0.969
296	117	1	0.913	0.02496	0.866	0.964
306	116	1	0.906	0.02596	0.856	0.958
309	115	1	0.898	0.02690	0.846	0.952
311	114	1	0.890	0.02779	0.837	0.946
339	112	1	0.882	0.02866	0.827	0.940
345	111	1	0.874	0.02948	0.818	0.934
393	108	1	0.866	0.03030	0.808	0.927
435	106	1	0.858	0.03109	0.799	0.921
447	104	1	0.849	0.03187	0.789	0.914
448	103	1	0.841	0.03261	0.780	0.908
463	101	1	0.833	0.03333	0.770	0.901
490	98	1	0.824	0.03406	0.760	0.894
497	95	1	0.816	0.03479	0.750	0.887
510	94	1	0.807	0.03548	0.740	0.880
525	93	1	0.798	0.03615	0.730	0.872
533	92	2	0.781	0.03738	0.711	0.858
562	89	1	0.772	0.03798	0.701	0.850
563	88	1	0.763	0.03855	0.691	0.843
749	84	1	0.754	0.03915	0.681	0.835
1052	73	1	0.744	0.03995	0.670	0.827
1324	57	1	0.731	0.04133	0.654	0.817

Figura 4.42 Estimación de Superviviencia Modelo Kaplan- Meier para mujeres con CACU de la Jurisdicción Sanitaria II.

La figura 4.42 se describe a continuación en donde “time” son los días en que sucedió el evento de interés, que en este caso es la muerte de las pacientes o paciente, “n.risk”, corresponde a las mujeres que quedan en riesgo una vez sucedido el evento y “survival” es la probabilidad de sobrevivencia calculada, “std.err” el error estándar y “lower 95% CI y 95% CI” los intervalos de confianza.

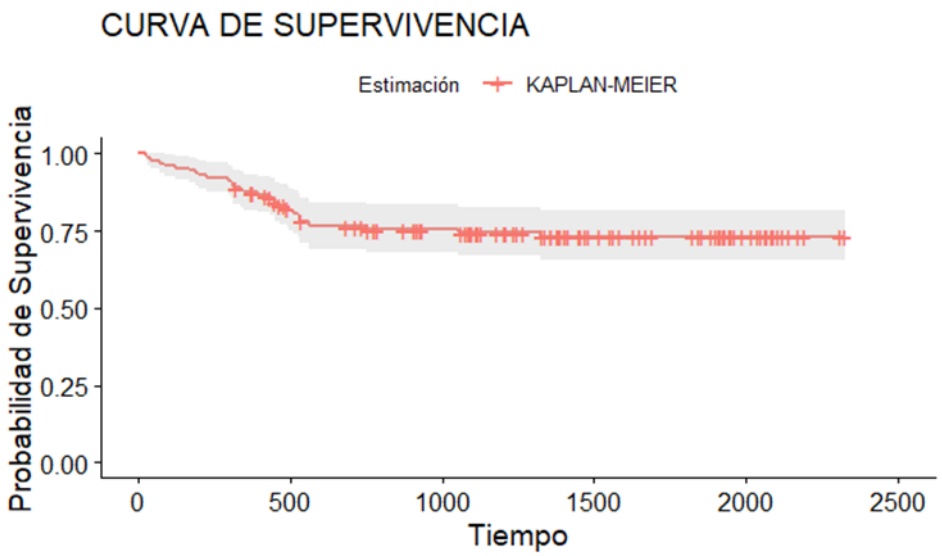


Figura 4.43 Curva de Supervivencia Estimación Kaplan-Meier para mujeres con CACU de la Jurisdicción Sanitaria II

4.3.7.6 Análisis Modelo de Riesgos Proporcionales de Cox para pacientes con CACU de la Jurisdicción Sanitaria

Se presenta la programación en R de los datos de las mujeres con CACU registradas en la Jurisdicción Sanitaria II

Al primer análisis realizado le llamamos cox1, en este análisis están involucradas las covariables diabetes, siendo esta dicotómica catalogada con 1 cuando no la padece y 0 cuando si la padece.

```
survdiff(Surv(datos20$tiempos,datos20$evento)~datos20$diabetes)
cox1<-
coxph(Surv(datos20$tiempos,datos20$evento)~datos20$diabetes+datos20$edad)
datos20<-read.csv("CANCER2.CSV",header = T)
> cox1<-coxph(Surv(datos20$tiempos,datos20$evento)~datos20$diabetes+datos
20$edad)
> cox1
Call:
coxph(formula = Surv(datos20$tiempos, datos20$evento) ~ datos20$diabetes
+
    datos20$edad)
                      coef exp(coef) se(coef)      z      p
datos20$diabetes -0.91453   0.40070  0.47019 -1.945 0.0518
datos20$edad      0.02646   1.02681  0.01315  2.012 0.0442

Likelihood ratio test=8.99  on 2 df, p=0.01118
n= 127, number of events= 32
```

Figura 4.44 Cox1 para nivel de significancia

En la figura 4.44 se muestra una salida de R program en donde la significancia del modelo puede verificarse solo a través de la razón de verosimilitud.

```
> cox1<-coxph(Surv(datos20$tiempos,datos20$evento)~datos20$diabetes+dat
os20$edad)
> summary(cox1)
Call:
coxph(formula = Surv(datos20$tiempos, datos20$evento) ~ datos20$diabete
s +
    datos20$edad)

  n= 127, number of events= 32

                      coef exp(coef) se(coef)      z Pr(>|z|)
datos20$diabetes -0.91453   0.40070  0.47019 -1.945   0.0518 .
datos20$edad      0.02646   1.02681  0.01315  2.012   0.0442 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
datos20$diabetes    0.4007     2.4956    0.1594     1.007
datos20$edad        1.0268     0.9739    1.0007     1.054

Concordance= 0.637  (se = 0.05 )
Likelihood ratio test= 8.99  on 2 df,   p=0.01
Wald test            = 10.41  on 2 df,   p=0.005
Score (logrank) test = 11.33  on 2 df,   p=0.003
```

Figura 4.45 Cox1 Estimación de Riesgos Relativos

La estimación obtenida directamente a través de la salida presentada en la figura 4.45 es la estimación de los riesgos relativos a partir de los exp(coef) con lo que podemos decir que las dos variables resultaron significativas, pudiendo no considerar la diabetes, sin embargo, los resultados del análisis de Cox permiten tomarla en cuenta.

La presencia de no padecer diabetes reduce el riesgo en un factor del .40 o 60%, la taza de riesgo para esta covariable es 0.407; el coeficiente β para diabetes es -.0.91453 por lo que su signo negativo indica una disminución en el riesgo, también así podemos verificar un valor p=.05.

De manera similar la edad con un coeficiente β=0.026 y un valor p=.04 representa una relación de riesgo HR=1.02, indica una relación significante a

mayor riesgo de muerte. Manteniendo constante la otra covariable. Una mayor edad se asocia con mala supervivencia.

Respecto al contraste de la razón de verosimilitud p=0.01<.05, indica que los factores son independientes uno de otro.

El contraste de Wald con p=.005<.05 indica que el valor de los coeficientes son diferentes de cero, y el contraste del “score” Logrank concluye que las curvas de supervivencia de los grupos de pacientes con diabetes difieren significativamente con p=.003<.05.

La figura 4.46 presenta la verificación de los supuestos del modelo de Cox que se resume para el análisis Cox1

```
> cox.zph(cox1)
                     rho  chisq     p
datos20$diabetes -0.0222 0.0149 0.903
datos20$edad      0.1218 0.4177 0.518
GLOBAL                NA 0.5025 0.778
```

Figura 4.46 Verificación de los Supuestos del Modelo De Cox1

No existe evidencia significativa al 5% de que se viole el supuesto de que, el coeficiente entre el riesgo para dos sujetos con el mismo vector de covariables es constante en el tiempo ni desde el punto de vista global, ni para ca da covariable.

Dado que los valores p>.05 para los respectivos factores significa que las covariables analizadas son independientes entre si.

```
plot(survfit(cox1))
```

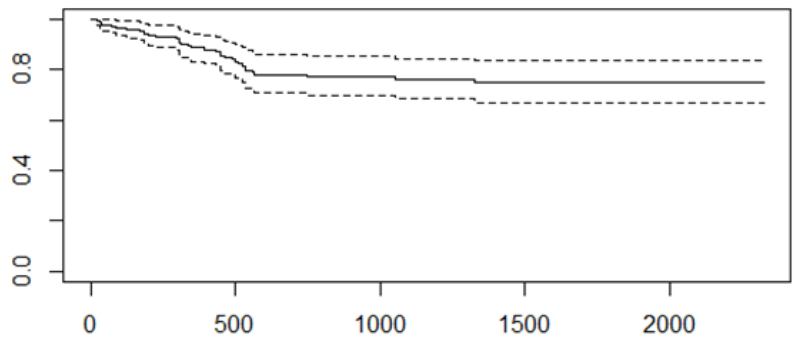


Figura 4.47 Ajuste de Supervivencia de Modelo de Cox1

```
summary(survfit(cox1))
Call: survfit(formula = cox1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   24    127       1    0.993 0.00688        0.980        1.000
   31    126       1    0.986 0.00972        0.967        1.000
   40    125       1    0.979 0.01192        0.956        1.000
   71    124       1    0.972 0.01380        0.946        1.000
   87    123       1    0.966 0.01545        0.936        0.996
  123    122       1    0.959 0.01696        0.926        0.992
  168    121       1    0.951 0.01838        0.916        0.988
  187    120       1    0.944 0.01970        0.906        0.984
  198    119       1    0.937 0.02094        0.897        0.979
  227    118       1    0.930 0.02210        0.888        0.974
  296    117       1    0.923 0.02320        0.879        0.970
  306    116       1    0.916 0.02424        0.870        0.965
  309    115       1    0.909 0.02523        0.861        0.960
  311    114       1    0.902 0.02617        0.852        0.954
  339    112       1    0.894 0.02711        0.843        0.949
  345    111       1    0.887 0.02802        0.834        0.944
  393    108       1    0.880 0.02893        0.825        0.938
  435    106       1    0.872 0.02985        0.816        0.933
  447    104       1    0.864 0.03074        0.806        0.927
  448    103       1    0.857 0.03160        0.797        0.921
  463    101       1    0.849 0.03245        0.788        0.915
  490     98       1    0.841 0.03330        0.778        0.909
  497     95       1    0.833 0.03417        0.769        0.903
  510     94       1    0.825 0.03501        0.759        0.896
  525     93       1    0.817 0.03581        0.749        0.890
  533     92       2    0.799 0.03733        0.729        0.876
  562     89       1    0.790 0.03807        0.719        0.869
  563     88       1    0.781 0.03877        0.709        0.861
  749     84       1    0.772 0.03950        0.698        0.854
 1052     73       1    0.761 0.04051        0.686        0.845
 1324     57       1    0.747 0.04212        0.669        0.835
```

Figura 4.48 Análisis de Supervivencia para Cox1

Modelo Cox1

$$h(t, diabetes, edad) = h_0 . e^{-0.914 diabetes + 0.264\ edad}$$

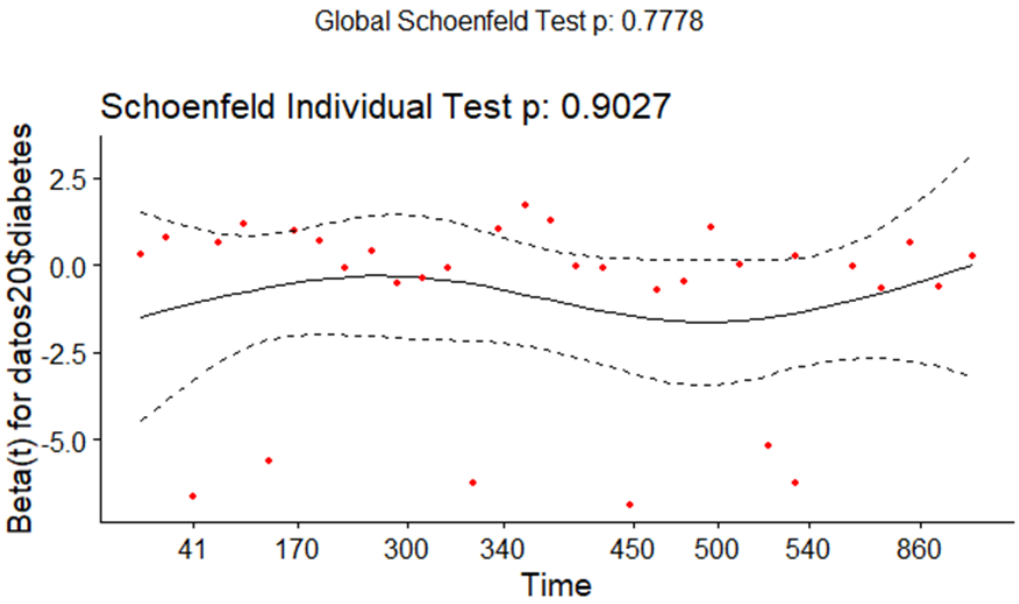


Figura 4.49 Residuos de Schoenfeld para Diabetes de Cox1

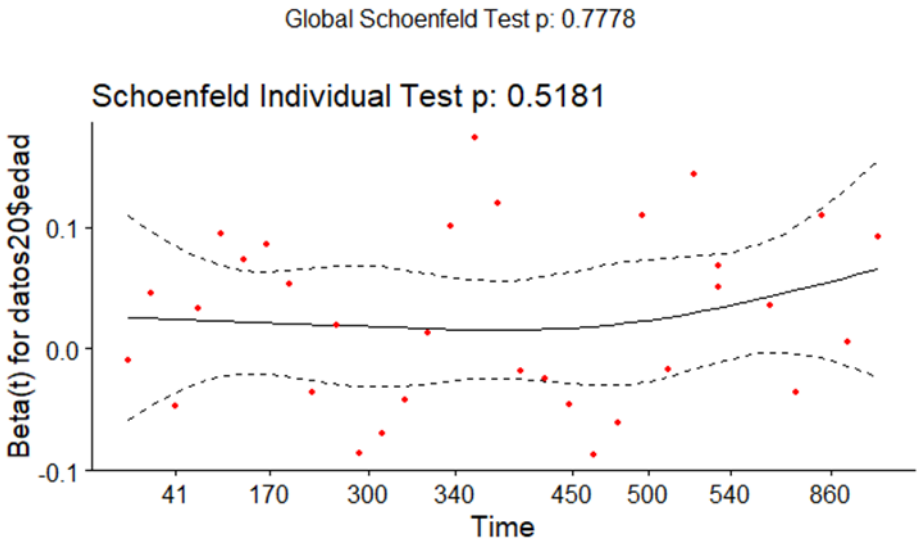


Figura 4.50 Residuos de Schoenfeld para Edad de Cox1

El siguiente análisis al cual nos referiremos como Cox2, considera como covariables dicotómicas la escolaridad, 1 considerando si la paciente cursó primaria o secundaria y 0 si tiene preparatoria o profesional, la variable estado civil catalogada con 0 si es soltera y 1 no soltera, así también las covariables edad y cantidad de parejas.

```
#PLANTEAMINETO ANALISIS DE COX2 CON ESCOL, EDAD, PAREJAS,EDOCIVIL
> cox2<-coxph(Surv(datos20$tiempos,datos20$evento)~datos20$escolaridad+
datos20$edad+datos20$parejas+datos20$ecivil)
> cox2
Call:
coxph(formula = Surv(datos20$tiempos, datos20$evento) ~ datos20$escolar
idad +
    datos20$edad + datos20$parejas + datos20$ecivil)

                       coef exp(coef) se(coef)     z      p
datos20$escolaridad 1.30243   3.67822  1.03420 1.259 0.2079
datos20$edad        0.02982   1.03027  0.01454 2.051 0.0403
datos20$parejas     0.04398   1.04496  0.12549 0.350 0.7260
datos20$ecivil      0.03863   1.03939  0.39254 0.098 0.9216

Likelihood ratio test=8.27  on 4 df, p=0.0821
n= 127, number of events= 32
```

Figura 4.51 Cox2 para nivel de Significancia

En la figura 4.51 se muestra una salida de R program en donde la significancia del modelo puede verificarse solo a través de la razón de verosimilitud.

También muestra con el comando coxph la estimación de riesgos relativos para el análisis Cox2

```
> summary(cox2)
Call:
coxph(formula = Surv(datos20$tiempos, datos20$evento) ~ datos20$escolar
idad +
    datos20$edad + datos20$parejas + datos20$ecivil)

  n= 127, number of events= 32
                       coef exp(coef) se(coef)     z Pr(>|z|)
datos20$escolaridad 1.30243   3.67822  1.03420 1.259   0.2079
datos20$edad        0.02982   1.03027  0.01454 2.051   0.0403 *
datos20$parejas     0.04398   1.04496  0.12549 0.350   0.7260
datos20$ecivil      0.03863   1.03939  0.39254 0.098   0.9216
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                    exp(coef) exp(-coef) lower .95 upper .95
datos20$escolaridad     3.678     0.2719    0.4845    27.922
datos20$edad            1.030     0.9706    1.0013     1.060
datos20$parejas         1.045     0.9570    0.8171     1.336
datos20$ecivil          1.039     0.9621    0.4816     2.243

Concordance= 0.65  (se = 0.046 )
Likelihood ratio test= 8.27  on 4 df,   p=0.08
Wald test            = 7.04  on 4 df,   p=0.1
Score (logrank) test = 7.73  on 4 df,   p=0.1
```

Figura 4.52 Cox2 Estimación de Riesgos Relativos

La estimación obtenida directamente a través de la salida presentada en la figura 4.52 es la estimación de los riesgos relativos de donde se explica a continuación.

La variable edad es la única que se presenta significante con un valor p=0.04, a diferencia de escolaridad, cantidad de parejas, y estado civil que no se muestran significantes.

El valor p=0.08 de la razón de máxima verosimilitud, indica que las influencias de dichos factores analizados no son de gran influencia en este modelo La escolaridad con un β=1.30 representa una relación de riesgo HR=3.67, aunque esta no sea significante con p=0.207 indica una relación significante a un mayor riesgo de muerte.

La edad con un β=0.02 con una razón de riesgo HR=1.03, un año adicional de edad induce un riesgo diario de muerte, aunque no con una contribución significativa dado su coeficiente β, así manteniendo constantes las otras variables.

Respecto a la cantidad de parejas, no representa significancia en el modelo, dado que p=0.72 con un coeficiente β=0.04. Al incrementar en 1 la cantidad de parejas induce un riesgo diario de muerte a la razón de riesgo de 1.044.

El estado civil no tiene significancia dado que su valor p=0.92 así lo demuestra; el coeficiente β=0.03 y su HR=1.03 indica un mayor riesgo de muerte.

Respecto al contraste de razón de máxima verosimilitud el valor p=0.08>0.05 expone que los valores analizados pudieran no ser independientes uno de otro en forma general, sin olvidar que la edad si lo es, así mismo el contraste de Wald presenta un valor p=0.1>0.05 lo que nos indica que los valores de los coeficientes pueden ser cero, por lo que podrían no influir en el modelo.

El contraste del score con p=0.1>0.05 indica que las curvas de supervivencia de los grupos con escolaridad y sin ella; no difieren de manera significativa.

La figura 4.53 presenta la verificación de los supuestos del modelo de Cox que se resume para el análisis Cox2, el comando cox.zph de R program arroja este resumen.

```
cox.zph(cox2)
                          rho    chisq     p
datos20$escolaridad  0.136308 0.611511 0.434
datos20$edad         0.156627 0.612584 0.434
datos20$parejas      0.155342 0.655322 0.418
datos20$ecivil      -0.000863 0.000024 0.996
GLOBAL                     NA 1.518387 0.823
```

Figura 4.53 Verificación de los Supuestos del Modelo De Cox2

Dado que los valores p>.05 para los respectivos factores significa que las covariables analizadas son independientes entre si. No se rechaza la hipótesis nula por lo que no se viola el supuesto de riesgos proporcionales , ni desde el punto de vista global ni para cada covariable.

```
plot(survfit(cox2))
```

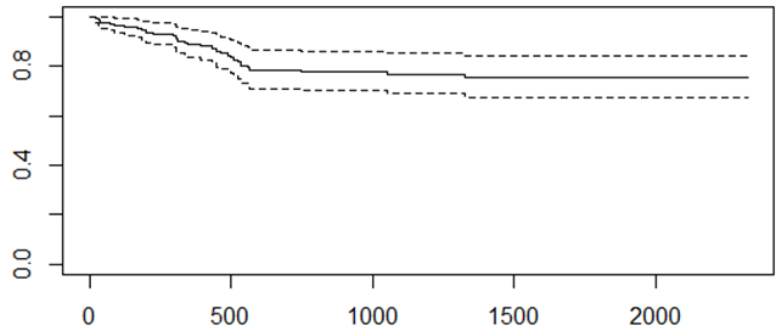


Figura 4.54 Ajuste de Supervivencia de Modelo de Cox2

```
summary(survfit(cox2))
Call: survfit(formula = cox2)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   24    127       1    0.993 0.00680        0.980        1.000
   31    126       1    0.986 0.00964        0.968        1.000
   40    125       1    0.980 0.01184        0.957        1.000
   71    124       1    0.973 0.01370        0.946        1.000
```

87	123	1	0.966	0.01536	0.936	0.997
123	122	1	0.959	0.01688	0.927	0.993
168	121	1	0.952	0.01830	0.917	0.989
187	120	1	0.945	0.01964	0.908	0.985
198	119	1	0.938	0.02090	0.898	0.980
227	118	1	0.931	0.02209	0.889	0.976
296	117	1	0.924	0.02322	0.880	0.971
306	116	1	0.917	0.02430	0.871	0.966
309	115	1	0.910	0.02533	0.862	0.961
311	114	1	0.904	0.02633	0.853	0.957
339	112	1	0.897	0.02731	0.845	0.952
345	111	1	0.889	0.02827	0.836	0.947
393	108	1	0.882	0.02924	0.827	0.941
435	106	1	0.875	0.03020	0.817	0.936
447	104	1	0.867	0.03115	0.808	0.930
448	103	1	0.860	0.03207	0.799	0.925
463	101	1	0.852	0.03298	0.790	0.919
490	98	1	0.844	0.03389	0.780	0.913
497	95	1	0.836	0.03483	0.771	0.907
510	94	1	0.828	0.03574	0.761	0.901
525	93	1	0.820	0.03662	0.751	0.895
533	92	2	0.803	0.03831	0.732	0.882
562	89	1	0.795	0.03914	0.722	0.875
563	88	1	0.786	0.03994	0.712	0.869
749	84	1	0.777	0.04078	0.701	0.862
1052	73	1	0.767	0.04187	0.689	0.854
1324	57	1	0.754	0.04359	0.673	0.844

Figura 4.55 “Análisis de Supervivencia para Cox2”

Modelo Cox2

$$h(t, escolaridad, edad, parejas, ecivil) = h_0 . e^{1.3escol+0.029edad+.0.043parejas+0038ecivil.}$$

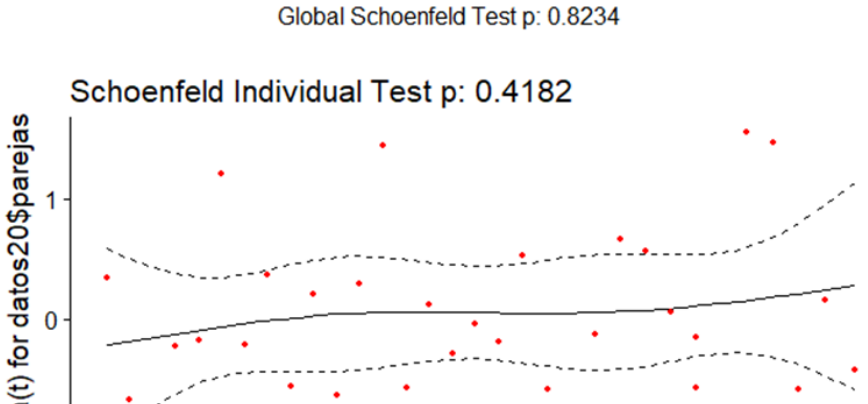


Figura 4.56 Residuos de Schoenfeld para Parejas de Cox2

Figura 4.57 Residuos de Schoenfeld para Estado Civil de Cox2

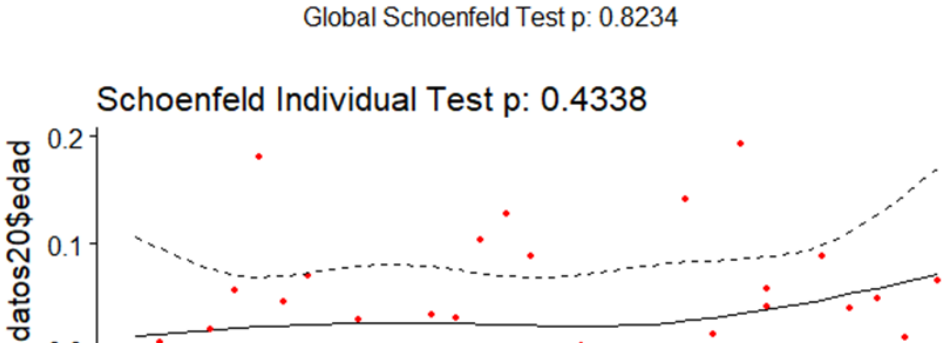


Figura 4.58 Residuos de Schoenfeld para Edad de Cox2

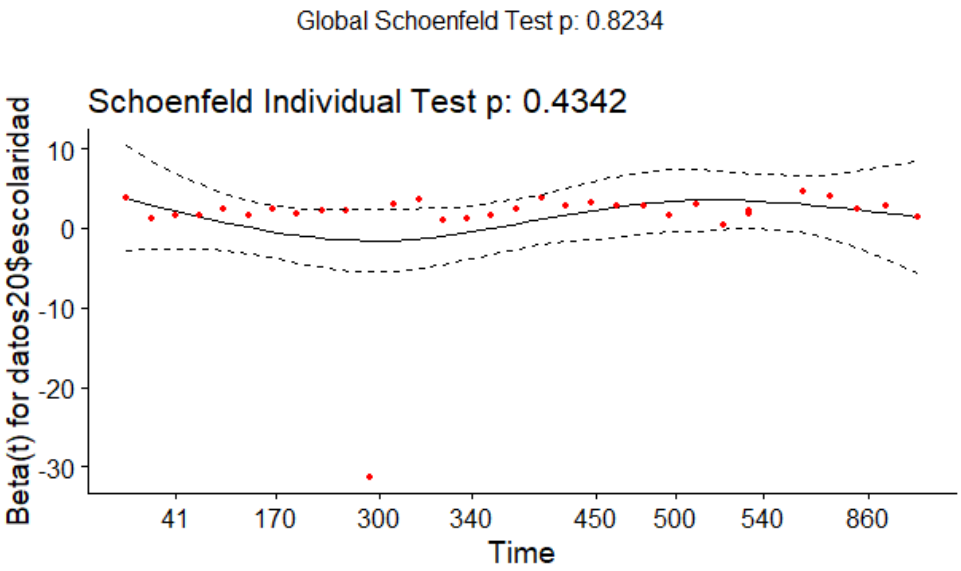


Figura 4.59 Residuos de Schoenfeld para Escolaridad de Cox2

Ahora considerando como covariables dicotómicas la escolaridad, 1 considerando si la paciente curso primaria o secundaria y 0 si tiene preparatoria o profesional, y edad, a este análisis lo llamamos Cox3 para su referencia.

```
#PLANTEAMIENTO ANÁLISIS COX3 CON ESCOLARIDAD Y EDAD
> cox3<-coxph(Surv(datos20$tiempos,datos20$evento)~datos20$escolaridad+
datos20$edad)
> summary(cox3)
Call:
coxph(formula = Surv(datos20$tiempos, datos20$evento) ~ datos20$escolar
idad +
    datos20$edad)

  n= 127, number of events= 32

                         coef exp(coef) se(coef)     z Pr(>|z|)
datos20$escolaridad 1.28008   3.59694  1.02418 1.250   0.2113
datos20$edad        0.02747   1.02785  0.01275 2.155   0.0311 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                    exp(coef) exp(-coef) lower .95 upper .95
datos20$escolaridad     3.597     0.2780    0.4832    26.774
datos20$edad            1.028     0.9729    1.0025     1.054

Concordance= 0.647  (se = 0.047 )
Likelihood ratio test= 8.15  on 2 df,   p=0.02
Wald test            = 6.97  on 2 df,   p=0.03
Score (logrank) test = 7.64  on 2 df,   p=0.02
```

Figura 4.60 Cox3 Estimación de Riesgos Relativos

La figura 4.60 muestra una salida de R program con el comando coxph la estimación de riesgos relativos para el análisis Cox3 a partir de los exp(coef), razón de riesgo. (HR).

El valor p=0.21 para escolaridad, la presenta como no significativa, a diferencia de la edad, con un valor p=0.03 siendo esta nuevamente significativa como lo vimos en los dos análisis anteriores en Cox1 y Cox2.

El hecho de tener escolaridad mínima, presenta un coeficiente β=1.28 lo que indica que un bajo nivel de estudios tiene una HR=3.59, que representa una relación de riesgo fuerte, es decir se asocia con una mala supervivencia.

Siendo la edad significante como mencionamos anteriormente, se presenta consistente en el valor del coeficiente β=0.02 con una relación de riesgo HR=1.02, manteniendo constante la otra covariable. Una mayor edad se asocia con mala supervivencia.https://estadisticaorquestainstrumento.wordpress.com/2013/04/30/test-de-wald/(rescatado 8 mayo 2019)

Ahora evaluando los contrastes se verifica que el de máxima verosimilitud con p=0.02< que 0.05 indica que los factores son independientes uno de otro, ahora el contraste de Wald con p=0.03<0.05 nos dice que los coeficientes son diferentes de cero, y por último el logrank test con p=0.02<0.05, nos indica que las curvas de supervivencia de los grupos con escolaridad y sin ella difieren significativamente.

La figura 4.61 presenta la verificación de los supuestos del modelo de Cox que se resume para el análisis Cox3, el comando cox.zph de R program arroja este resumen.

```
> #VERIFICACIÓN DE LOS SUPUESTOS DEL MODELO DE COX
> cox.zph(cox3)
                      rho chisq     p
datos20$escolaridad 0.1270 0.498 0.480
datos20$edad        0.0962 0.253 0.615
GLOBAL                 NA 0.851 0.653
```

Figura 4.61 Verificación de los Supuestos del Modelo de Cox3

```
> plot(survfi
```

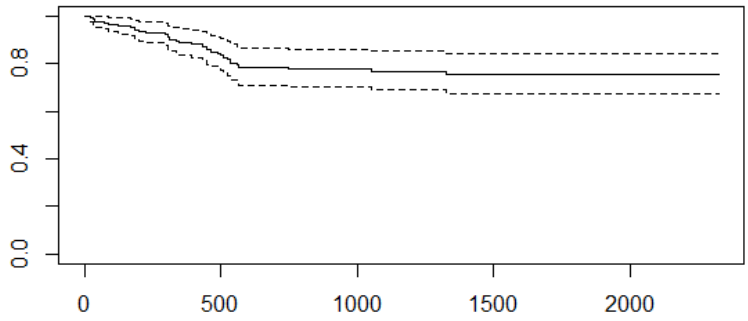


Figura 4.62 Ajuste de Supervivencia de Modelo de Cox3

```
> #FUNCIÓN DE SOBREVIVENCIA AJUSTADA MEDIANTE COX
> summary(survfit(cox3))
Call: survfit(formula = cox3)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   24    127       1    0.993 0.00681        0.980        1.000
   31    126       1    0.986 0.00966        0.968        1.000
   40    125       1    0.980 0.01187        0.957        1.000
   71    124       1    0.973 0.01374        0.946        1.000
   87    123       1    0.966 0.01540        0.936        0.997
  123    122       1    0.959 0.01692        0.926        0.993
  168    121       1    0.952 0.01834        0.917        0.989
  187    120       1    0.945 0.01968        0.907        0.984
  198    119       1    0.938 0.02094        0.898        0.980
  227    118       1    0.931 0.02213        0.889        0.976
  296    117       1    0.924 0.02326        0.880        0.971
  306    116       1    0.917 0.02434        0.871        0.966
  309    115       1    0.910 0.02537        0.862        0.961
  311    114       1    0.903 0.02636        0.853        0.956
  339    112       1    0.896 0.02734        0.844        0.951
  345    111       1    0.889 0.02829        0.835        0.946
  393    108       1    0.882 0.02926        0.826        0.941
  435    106       1    0.874 0.03022        0.817        0.936
  447    104       1    0.867 0.03117        0.808        0.930
  448    103       1    0.859 0.03209        0.799        0.925
  463    101       1    0.852 0.03299        0.790        0.919
  490     98       1    0.844 0.03390        0.780        0.913
  497     95       1    0.836 0.03483        0.770        0.907
  510     94       1    0.828 0.03574        0.761        0.901
  525     93       1    0.820 0.03661        0.751        0.895
  533     92       2    0.803 0.03829        0.732        0.882
  562     89       1    0.795 0.03912        0.722        0.875
  563     88       1    0.786 0.03992        0.712        0.868
  749     84       1    0.777 0.04075        0.701        0.861
 1052     73       1    0.767 0.04184        0.689        0.853
 1324     57       1    0.754 0.04354        0.673        0.844
```

Figura 4.63 Análisis de Supervivencia para Cox3

Modelo Cox3

$$h(t, escolaridad, edad) = h_0 . e^{1.28 escol + 0.027 edad.}$$

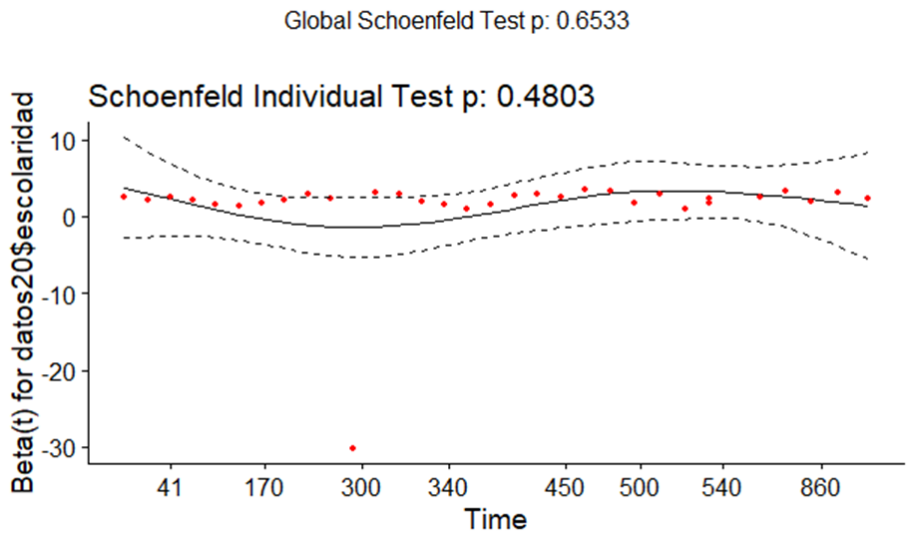


Figura 4.64 Residuos de Schoenfeld para Escolaridad de Cox3

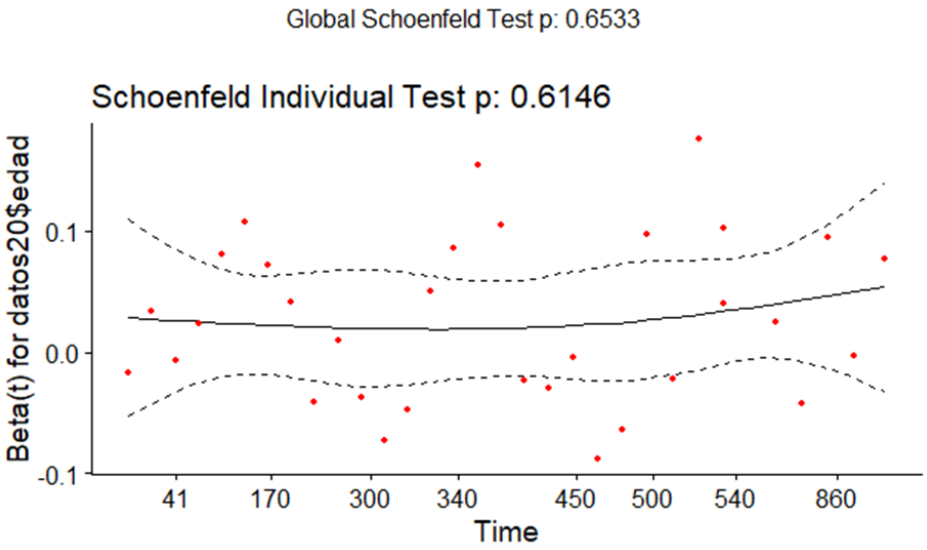


Figura 4.65 Residuos de Schoenfeld para Edad de Cox3

4.3.8 Análisis para Identificar Factores Determinantes en Pacientes con CaCu y Grupo de Control

Con el propósito de determinar los factores de mayor significancia que describen a las mujeres que padecen cáncer se presentan análisis con el sistema MTS, Componentes Principales y Modelo de riesgos proporcionales de Cox con los registros de la historia clínica de 128 pacientes diagnosticadas con CaCu mediante una biopsia y con el reporte para la prueba de Papanicolaou en la Clínica de colposcopía de la Jurisdicción Sanitaria II de Cd. Juárez Chihuahua México, durante el período 2013 a 2017. Para crear el grupo de control, se aplicó un cuestionario para registrar la historia clínica con los aspectos destacados en las pacientes con CaCu a 120 mujeres que se realizaron la prueba de Papanicolau en el Hospital de la Mujer de esta misma entidad, y que resultaron sin ninguna neoplasia.

4.3.8.1 Análisis por Componentes Principales para Pacientes con CaCu y Grupo de Control

Para Sánchez Mangas (2012), la técnica del análisis de Componentes Principales consiste en analizar un conjunto de datos de entrada, el cual contiene diferentes observaciones descritas por múltiples variables independientes o dependientes y cuyas relaciones entre sí no tienen por qué conocerse. Como ya se ha dicho anteriormente el objetivo principal es reducir la dimensión del conjunto de datos de entrada intentado mantener la mayor cantidad de información posible para poder analizarlos de forma más fácil y que en etapas posteriores, como clasificadores o regresores, se puedan simplificar los criterios de decisión.

En este nuevo sistema de coordenadas las componentes principales están ordenadas automáticamente según la varianza de la proyección de datos, es decir,

según la cantidad de información que contengan. Finalmente, se puede reducir la dimensión de los datos resultantes en el nuevo espacio eliminando las componentes principales que presenten una menor varianza, es decir, que aporten menos información. La base matemática que se utiliza para desarrollar el PCA es el álgebra lineal. Dicho en forma resumida este modelo busca: concentrar la mayor parte de varianza en un número reducido de variables del nuevo espacio, o dicho también; busca que la mayor cantidad de información quede contenida en el menor número posible de variables.

Mediante esta técnica se pueden procesar un extenso conjunto de datos y reducir su dimensionalidad con una pérdida mínima de información.(video)

La principal desventaja está en el hecho que cada componente principal sea una combinación lineal de todas las variables originales, lo que hace muy difícil poder analizar la importancia de cada variable original en el nuevo espacio. Este método no ayuda a reducir la dimensión en términos de variables originales ya que para el cálculo de los componentes principales se necesitan todas las variables originales.

El propósito de la aplicación de este modelo, es identificar de entre las 18 variables que describen los datos, obtener la mayor información de ellas en un número reducido de variables.

En la tabla 4.66 se muestran los resultados de los componentes principales, las primeras siete columnas muestran los valores propios mayores que 1 y explican el 60.3% de la variación de los datos.

Tabla 4.66 “Análisis de Valores y Vectores Propios de la Matriz de Correlación

Análisis de los valores y vectores propios de la matriz de correlación

Valor propio	3.8208	1.9316	1.7087	1.3420	1.1843	1.0507	1.0302	0.9661	0.9334	0.8976
Proporción	0.191	0.097	0.085	0.067	0.059	0.053	0.052	0.048	0.047	0.045
Acumulada	0.191	0.288	0.373	0.440	0.499	0.552	0.603	0.652	0.698	0.743

La figura 4.66 muestra la gráfica de sedimentación, compara visualmente el tamaño de los valores propios y ayuda a determinar el número de componentes con base en el tamaño de los valores propios, y aquí podemos notar el punto de inflexión que marca del primero al segundo componente y de ahí al resto. Debemos resaltar los siguientes tres componentes los cuales forman la mayor parte de la variabilidad, por lo que la línea comienza a enderezarse. El resto de estos representan una porción muy pequeña de ésta y probablemente son poco importantes.

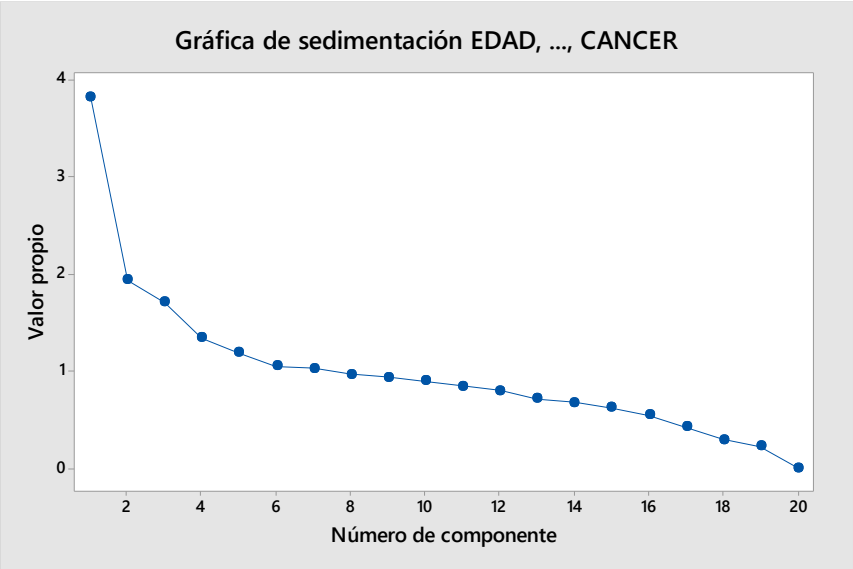


Figura 4.66 “Gráfica de Sedimentación”

La tabla 4.67 corresponde al cálculo de los componentes principales, donde se examina la magnitud y dirección de los coeficientes de las variables originales.

Cuanto mayor es el valor absoluto del coeficiente más importante es la variable correspondiente en el cálculo del componente.

Tabla 4.67 “Cálculo de Componentes Principales”

Variable	PC1	PC2	PC3	PC4	PC5
EDAD	0.111	-0.574	-0.026	0.076	-0.213
EDO.CIVIL	0.133	-0.162	-0.144	-0.147	-0.166
ESCOLARIDAD	-0.312	0.241	0.177	0.121	-0.121
RELIGIÓN	-0.079	0.189	-0.338	0.329	0.200
FAM.ANTEC	0.089	0.023	0.003	-0.478	0.032
DIABETES	-0.131	0.133	-0.208	-0.089	0.278
MENARCA	0.065	-0.105	0.205	0.265	0.081
RITMO MENSTRUAL	-0.148	0.076	0.014	-0.130	0.099
IVSA	-0.131	-0.127	0.504	0.175	-0.141
PAREJAS	-0.043	0.308	-0.209	0.352	-0.146
GESTAS	0.255	-0.271	-0.337	0.083	-0.033
ANTICONCEPTIVO	0.097	0.018	0.441	-0.155	0.042
VPH	-0.469	-0.146	-0.061	-0.033	0.016
PREVCITOL	-0.121	-0.008	-0.209	-0.277	-0.586
SITGINEC	0.082	0.411	0.222	-0.147	-0.219
EXPLORACIÓN	0.073	-0.139	0.147	0.434	-0.018
ETS	-0.256	-0.266	0.111	-0.006	0.051
CERVICITIS	0.062	-0.125	0.063	-0.216	0.582
CITOLOGÍA	0.430	0.118	0.058	0.067	-0.031
CANCER	-0.469	-0.146	-0.061	-0.033	0.016

En el Componente Principal 1 (PC1) está relacionado con la salud del paciente, el resultado de la citología y el número de gestas afectan de manera positiva al primer componente; la escolaridad, el padecer VPH y padecer cáncer afectan de manera negativa al componente. La relación de factores se concentra en el padecimiento de la enfermedad que corresponde al cáncer y VPH, dado el

resultado de la citología. Para el PC2, resalta la relación entre la edad de la paciente, la cual influye de manera negativa, sin embargo, la situación ginecobstétrica y la cantidad de parejas influyen de forma positiva.

La relación en el PC3 está concentrada en el inicio de la vida sexual y el tipo de anticonceptivo que influyen de manera positiva, por otro lado, la religión influye de manera negativa en el componente.

Para el PC4, la influencia positiva para la puntuación de este componente está conformado principalmente por la religión, cantidad de parejas y la exploración, y de forma negativa y destacada los antecedentes familiares .

En el PC5 la relación de factores se concentra en la prevención citológica que influye de manera negativa y la cervicitis, la cual su influencia e positiva

En la figura 4.67 podemos observar influencias negativas grandes en el primer componente como lo son cáncer, VPH, y la escolaridad y contrapunto la citología y gestas. Con lo que podemos decir que el primer componente mide principalmente la salud de las pacientes y los factores negativos la afectan.

Las influencias positivas en el segundo componente como son la situación ginecobstétrica y la cantidad de parejas sobre este y en contrapunto está la edad que de acuerdo a la gráfica está en relación con la cantidad de parejas.

Cada componente aporta las características con mayor variabilidad y no se pierde información.

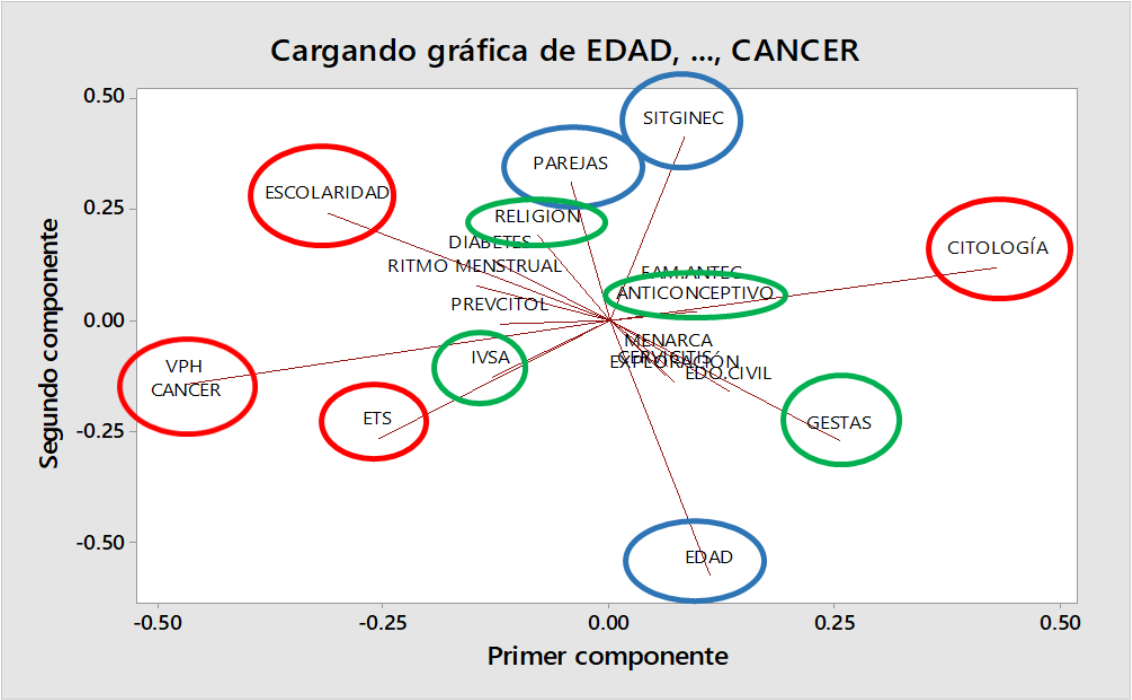


Figura 4.67 “Gráfico de Influencias por Componente”

La figura 4.68 muestra la gráfica de puntuación, en donde podemos destacar que las influencias que se aproximan a -1 ó 1 indican que la variable afecta considerablemente al componente. Las influencias cercanas a 0 indican que la variable tiene poca influencia. De tal manera que entonces esta gráfica divide en este caso al Componente 1 en cuanto si padecen cáncer y VPH en las pacientes, es decir si pertenecen al grupo Saludable ó No Saludable. Está gráfica detecta conglomerados. Los dos primeros componentes explican la mayor parte de la varianza.

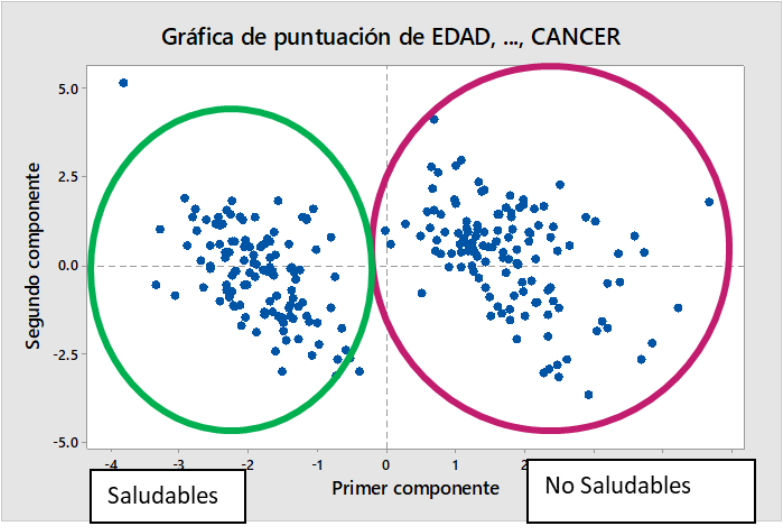


Figura 4.68 “Gráfica de Puntuación Grupo Saludable y Grupo No Saludable”

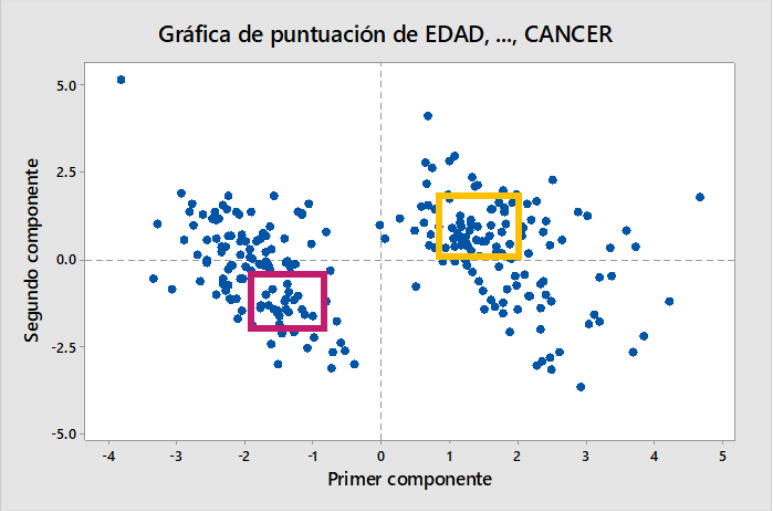


Figura 4.69 “Gráfica de Puntuación conglomerados Componente 2”

En la figura 4.69 que se detectan conglomerados, comparten características si se acercan a 1 o -1 tienen las mismas o resaltan las significantes del segundo componente. Por ejemplo, el recuadro amarillo son pacientes no saludables en las que tienen en común pocas parejas, en el recuadro rosa, tiene en común la cantidad de parejas y la edad.

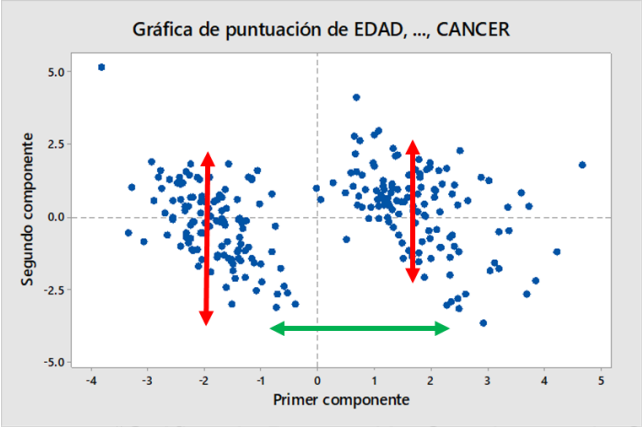


Figura 4.70 “Gráfica de Puntuación Conglomerado Cruzado”

En figura 4.70 podemos verificar que las diferencias entre el primer componente son más importantes que las diferencias a lo largo del segundo componente

4.3.8.1 Metodología para Desarrollar el MTS pacientes con Cáncer Cérvico Uterino

El MTS (Mahalanobis-Taguchi System) es usado para encontrar las variables significativas de un sistema multidimensional fusionando la distancia Mahalanobis, que ayuda a encontrar las condiciones anormales y la correlación de las variables, con herramientas de la metodología Taguchi que ayudan a reducir el número de variables que se van a analizar, como los arreglos ortogonales y la relación señal a ruido obtenidas de las condiciones anormales.

Una de las ventajas del MTS es introducir una escala basada en todas las características de las variables para medir el grado de anormalidad**,** minimiza el número de variables requeridas para un diagnóstico efectivo; el cual es el objetivo principal de este análisis; predice el desarrollo de un sistema multidimensional bajo varias condiciones; establece zonas de tratamiento de un producto o paciente basado en la severidad y costo (Cudney,2007).

El análisis de los datos se realizó con Minitab™ y Microsoft Excel®

La figura 4.7 representa la metodología para desarrollar el modelo aplicado para destacar los factores en las mujeres con CaCu de Cd. Juárez Chih. Mex. Registrados en la Clínica de Colposcopía de la Jurisdicción Sanitaria II.

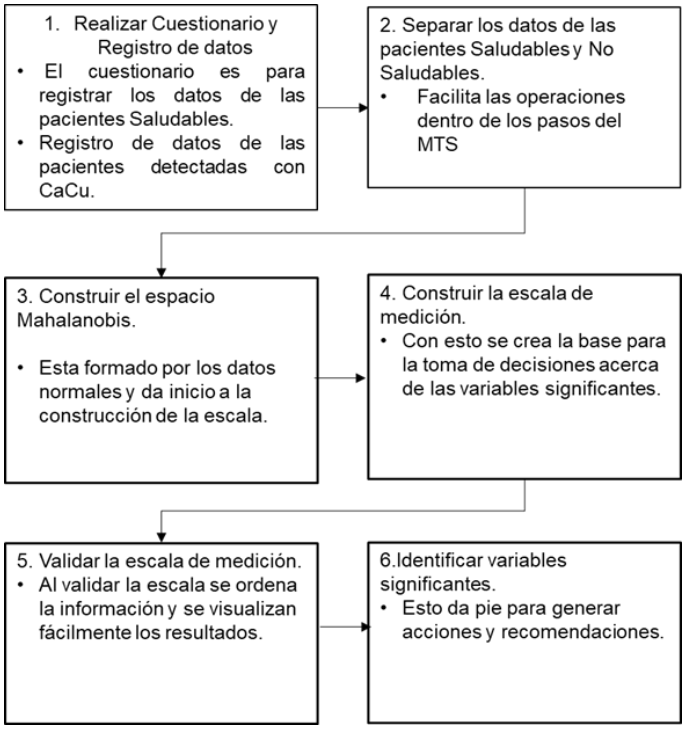


Figura 4.71 **“**Diagrama de Bloques para Metodología”

a) Se registran los factores asentados en la historia clínica de los expedientes de 128 pacientes diagnosticadas con CaCu mediante una biopsia, en el período comprendido de 2013 a 2017 en la Jurisdicción Sanitaria II de Cd. Juárez, Chihuahua. Este conjunto representa el grupo de las No Saludables.

b) Se realizó un cuestionario enfocado a la obtención de datos sobre los factores referentes a la historia clínica de mujeres que se practicaron el examen Papanicolaou en el Hospital de la Mujer de Cd. Juárez, Chihuahua durante los meses de enero a mayo del 2019 y el cual hubiese arrojado resultado negativo al CaCu y/o cualquier alteración. La encuesta se aplicó a 40 mujeres. La Tabla 1. muestra los factores a analizar.

Tabla 4.68 “Descripción de Variables”

Variables a Considerar		Descripción	Referencia S/N
x_1	Edad	Años	A
x_2	Estado Civil	Sotera, casada, viuda, divorciada, unión libre	B
x_3	Escolaridad	Primaria, secundaria, preparatoria, profesional, analfabeta	C
x_4	Religión	Católica, cristiana, Testigo de Jehová, otras	D
x_5	Antecedentes familiares de cáncer	Si /No	E
x_6	Menarca	Edad de la primera menstruación	F
x_7	Ritmo menstrual	Regular, irregular, abundante, escaso	G
x_8	IVSA	Edad inicio de vida sexual activa	H
x_9	Parejas	Cantidad de parejas sexuales	J
x_{10}	Gestas	Cantidad de gestas	K
x_{11}	Partos	Número de partos vaginales	L
x_{12}	Tipo anticonceptivo.	Hormonal oral, hormonal inyectable, implante, OTB, DIU, Ninguno	M
x_{13}	Prevención citológica	Primera vez en la vida, primera vez después de 3 años, subsecuente	N
x_{14}	Situación ginecobstétrica	Puerperio o post aborto, postmenopausia, uso de hormonales, DIU, Histerectomía, Tratamiento farmacológico, Embarazo actual, Tratamiento colposcópico previo, Ninguno	O
x_{15}	Exploración	Cuello aparentemente sano, cuello anormal, lesión de cuello, cervicitis, leucorrea, sangrado anormal, no se observa cuello	P
x_{16}	Antecedentes de infección de transmisión sexual	Si/No	Q
x_{17}	Cervicitis	Si/No	R
x_{18}	Resultado de la citología	Negativa para lesión intraepitelial o malignidad, dentro del límite normal, cambios celulares benignos reactivos inflamación, cambios celulares benignos infección, lesión escamosa intraepitelial carcinoma IN SITU, carcinoma epidermoide, adenocarcinoma endocervical in situ, adenocarcinoma(endocervical, endometrial, o extrauterino)	S

c) Separar los datos normales (condición saludable) de los anormales (condición no saludable). Las mujeres que presentan la encuesta son aquellas que se realizaron la citología vaginal para la valoración de cáncer d cérvico uterino y/o cualquier otra alteración y que su valoración fue libre de CaCu y sin alguna

alteración. Esta porción de mujeres sin enfermedad son parte del grupo normal, y la porción que sí presenta la enfermedad corresponde a los datos anormales.

d) Construir el espacio Mahalanobis. Este está formado por la fracción de individuos que componen el grupo saludable.

e) Construir escala de medición. Una vez establecido el espacio Mahalanobis se construye la escala de medición basándose en la porción de datos normales. Es importante estandarizar los datos y trabajar de acuerdo a una curva normal. Después se hace una matriz de correlación y la matriz inversa de esta, ya que con esta información se obtiene la matriz de vectores estandarizados que es necesaria para calcular las distancias Mahalanobis del grupo saludable y construir la escala.

f) Para validar la escala de medición se estandarizan los datos anormales (No Saludables), se calcula la matriz de correlación y la inversa de esta para calcular la distancia Mahalanobis y comparar las distancias de los datos anormales contra los datos saludables. La escala ha quedado validada cuando los datos anormales sean más altos que los datos normales.

g) La identificación de variables significantes se realiza por medio de arreglos ortogonales y relación señal a ruido *(S/N).* Con esto se prueba la importancia de cada variable analizada. Se realizan las sumas y diferencias *(S/N)* de ambos grupos de variables para identificar las variables útiles, es decir, aquellas que presentan mayores valores en dichas diferencias.

Usando del sistema Mahalanobis-Taguchi (MTS), destacamos cuáles variables aumentan la probabilidad de obtener CaCu. Por lo que las distancias Mahalanobis

calculadas con la Ecuación 3 se presentan en tabla 4.69 y corresponde a la presencia de todos los factores. Debemos considerar que las distancias Mahalanobis del grupo no saludable deben ser mayores que las del grupo saludable, así mismo considerar hacer referencia que las distancias mostradas corresponden a la primera corrida del OA, donde incluye todos los factores.

La figura 4.71 expone una notoria diferencia entre las distancias mostradas entre los dos grupos de referencia, las pacientes saludables, y las pacientes no saludables, las diagnosticadas con CaCu.

Después de calcular la distancia Mahalanobis, es necesario usar la señal a ruido para calcular las variables significantes. El diseño Tagushi representado en la Ecuación 1 correspondiente a este análisis es $L_{32}\ 2^{18}$, un arreglo de 32 corridas para 18 factores con 2 niveles.

La figura 4.72 muestra efectos principales para relaciones *S/N*, revelan que los recuadros con las letras B, O, Q, R y S correspondientes a los factores Estado civil, Situación ginecobstétrica, Antecedentes de enfermedades de transmisión sexual, Cervicitis y Resultado de la citología son los más destacados dentro de los factores ya establecidos. Los factores que tienen mayor pendiente y por lo tanto mayor significancia.

En la tabla 4.70 se verifica que, si el valor de la T calculada que se origina de las dos muestras es desmesurado, entonces se rechaza la hipótesis nula de que no hay diferencia en la media de las dos poblaciones (Sánchez Turcios, 2015).

Los factores destacados en este análisis indican valores $p < 0\ .05$ por lo que podemos decir que existe diferencia en las medias de estas dos muestras.

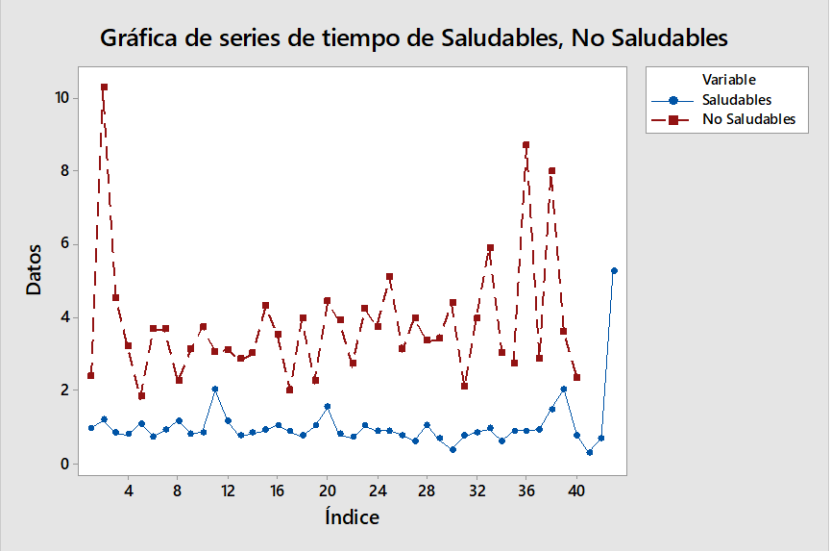


Figura 4.71 “Gráfica comparativa de Distancias de Mahalanobis”

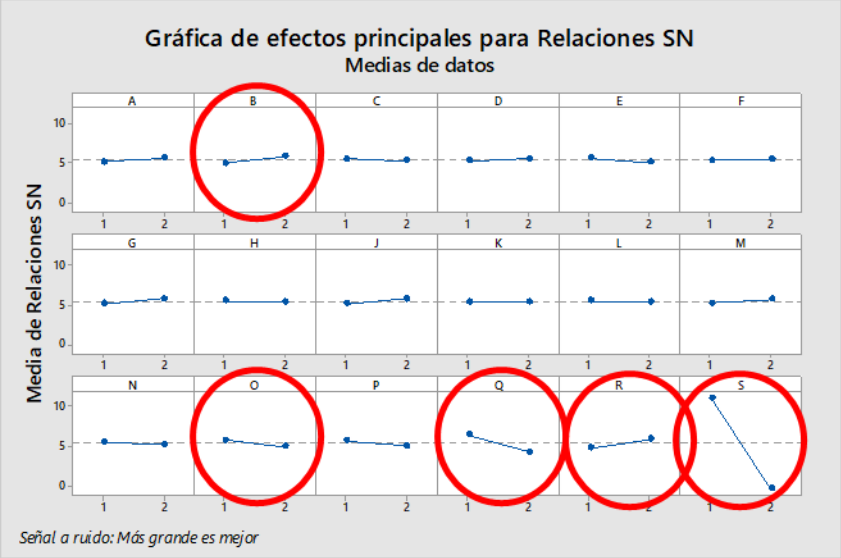


Figura 4.72 “Gráfica de Efectos Principales para Relaciones *S/N*”

Tabla 4.69 “Distancias de Mahalanobis”

No.	MD Grupo SIN Cáncer Cervico Uterino	MD Grupo Con Cáncer Cérvico Uterino
1	.9587	2.3568
2	1.1755	10.2643
3	.8246	4.5155
4	0.7707	3.2170
5	1.0676	1.8132
6	0.7121	3.6679
7	0.9017	3.6646
8	1.1455	2.2349
9	0.7905	3.0990
10	0.8302	3.7264
11	1.9947	3.0476
12	1.1266	3.0850
13	0.7291	2.8390
14	0.818	2.9846
15	0.8999	4.2843
16	1.0283	3.5322
17	0.864	1.9860
18	0.7379	3.9340
19	1.0049	2.2572
20	1.5275	4.4129
21	0.7872	3.8948
22	0.706	2.7195
23	1.005	4.2206
24	0.8744	3.7115
25	0.8742	5.0928
26	0.7485	3.1115
27	0.5955	3.9522
28	1.0252	3.3401
29	0.6631	3.3891
30	0.3646	4.3724
31	0.7301	2.0831
32	0.8406	3.9377
33	0.9300	5.8973
34	0.6079	3.0050
35	0.8782	2.7332
36	0.8772	8.6702
37	0.9086	2.8486
38	1.4617	7.9687
39	2.0211	3.5750
40	0.7645	2.3289

Tabla 4.70 “Coeficientes Relación *S/N”*

Término	***S/N***	**T**	**P**
Constante	5.43251	23.388	0.000
A Edad	-0.19517	-0.840	0.416
B Estado Civil	-0.41830	-1.801	0.095
C Escolaridad	0.16739	0.721	0.484
D Religión	-0.13099	-0.564	0.582
E Antecedentes familiares de cáncer	0.27912	1.202	0.251
F Menarca	-0.02192	-0.094	0.926
G Ritmo Menstrual	-0.33302	-1.434	0.175
H IVSA	0.07459	0.321	0.753
J Cantidad de Parejas	-0.33321	-1.435	0.175
K Embarazos	0.00303	0.013	0.990
L Partos Vaginales	0.08179	0.352	0.730
M Tipo de Anticonceptivo	-0.24018	-1.034	0.320
N Prevención citológica	0.16481	0.710	0.491
O Situación Ginecobstétrica	0.39107	1.684	0.116
P Exploración	0.34017	1.464	0.167
Q Antecedentes de Enfermedades De transmisión sexual	1.07491	4.628	0.000
R Cervicitis	-0.56390	-2.428	0.030
S Resultado citología	5.85376	25.201	0.000

Como se ha venido destacando a lo largo del texto, la utilidad del sistema MTS radica en que es una forma de determinar la similitud entre dos variables aleatorias multidimensionales (Escobedo 2008). Esto debido a que, en los sistemas multivariados la existencia de multicolinealidad (incidencia de fuertes correlaciones) dificulta el análisis para verificar diferencias o patrones de comportamiento.

La diferencia entre la MD y la distancia Euclidiana radica en que la primera tiene en cuenta la correlación entre las variables aleatorias. Las distancias de Mahalanobis nos ayuda a distinguir si un determinado conjunto de condiciones similares pertenece al conjunto de condiciones ideales.

Los factores registrados como relevantes para considerar en la historia clínica y análisis citológico de las mujeres para seguimiento en la salud y prevención al CaCu ya están cualificados por diferentes normas de salud a nivel mundial. El MTS nos ayuda a destacar las variables más significantes describiendo así a las mujeres de Ciudad Juárez Chihuahua que acuden a recibir atención en el Hospital de la Mujer para prevención de esta enfermedad. Si bien los factores registrados en esta investigación pueden considerarse como los usuales destacados en las investigaciones mencionadas, la posibilidad de analizar factores como la química sanguínea o perfiles hormonales de las pacientes con CaCu se ve inaccesible al menos en esta comunidad debido al elevado costo que esto representa para el Sector Salud, sin embargo, aquí se demostró que el modelo MTS jerarquiza, clasifica y determina significancia.

Este tipo de estudios puede ser utilizado por las autoridades pertinentes y especialistas en el área para la toma de decisiones sobre salud pública, ya que

destaca las principales características que afectan a las mujeres de determinadas comunidades.

4.3.8.3 Análisis Modelo de Riesgos proporcionales de Cox para pacientes con CaCu y Grupo de Control

El siguiente análisis al cual nos referiremos como Cox30, considera como covariables dicotómicas el padecer cancer, 1 si la paciente no tiene el padecimioneto y 0 si lo padece, y la misma identificación para la covariable diabetes. También muestra con el comando coxph la estimación de riesgos relativos para el análisis Cox30

La estimación obtenida directamente a través de la salida presentada en la figura 4.72 podemos observar como significantes en el modelo los siguientes factores los cuales serán explicados uno a uno.

```
> cox30<-coxph(Surv(datos30$tiempos,datos30$evento)~datos30$cancer+datos30$ci
tologia+datos30$cervicitis+datos30$ets+datos30$ecivil+datos30$sitginec+datos3
0$exploracion+datos30$parejas+datos30$ivsa+datos30$edad+datos30$menarca+datos
30$gestas+datos30$diabetes)
> summary(cox30)
Call:
coxph(formula = Surv(datos30$tiempos, datos30$evento) ~ datos30$cancer +
    datos30$citologia + datos30$cervicitis + datos30$ets + datos30$ecivil +
    datos30$sitginec + datos30$exploracion + datos30$parejas +
    datos30$ivsa + datos30$edad + datos30$menarca + datos30$gestas +
    datos30$diabetes)

  n= 127, number of events= 32
   (118 observations deleted due to missingness)

                        coef exp(coef) se(coef)      z Pr(>|z|)
datos30$cancer            NA        NA  0.00000     NA       NA
datos30$citologia    0.14159   1.15210  0.08237  1.719  0.08562 .
datos30$cervicitis   0.12826   1.13685  0.54875  0.234  0.81519
datos30$ets         -0.27022   0.76321  0.41601 -0.650  0.51599
datos30$ecivil       0.27004   1.31002  0.42055  0.642  0.52080
datos30$sitginec     0.02516   1.02548  0.07400  0.340  0.73388
datos30$exploracion  0.39953   1.49112  0.11546  3.460  0.00054 ***
datos30$parejas     -0.01644   0.98369  0.12820 -0.128  0.89793
datos30$ivsa        -0.33475   0.71551  0.12635 -2.649  0.00806 **
datos30$edad         0.03862   1.03938  0.01782  2.167  0.03021 *
datos30$menarca     -0.08625   0.91737  0.11106 -0.777  0.43742
datos30$gestas      -0.08135   0.92187  0.08537 -0.953  0.34062
datos30$diabetes    -1.44772   0.23511  0.55746 -2.597  0.00940 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 4.72 “Análisis Supervivencia Cox30”

Respecto a la citología, que corresponde a las categorías generales del diagnóstico citológico, que van desde negativas para lesión intraepitelial o malignidad hasta adenocarcinoma.

Un diagnóstico citológico con características atípicas o con carcinoma epidermoide representa una relación de riesgo HR = 1.15210 indica una relación significante a un mayor riesgo de muerte. Esta covariable tiene un coeficiente β = 0.14159 y valor p = 0.08562.

La covariable denominada “Exploración” describe a lo observado en el cuello uterino al ser revisado, si es un cuello aparentemente sano, anormal, si presenta lesión; (la tabla 4.68 describe las características de las covariables) de las cuales se destaca que al no observarse aparentemente sano representa un riesgo de muerte con una relación de riesgo HR = 1.49112, coeficiente β = 0.39953 y valor p=0.00054, lo cual lo posiciona con gran significancia.

La edad inicio de la vida sexual (IVSA), con un coeficiente β = -0.33475 lo que indica una disminución en el riesgo, y HR = 0.7155 señala que el inicio de la vida sexual después de la mayoría de edad reduce el riesgo el riesgo en un 28.45%, teniendo un valor p = 0.0086.

Por otro lado, la edad, con un coeficiente β = 0.03862 y una razón de riesgo HR = 1.03938 y un valor p = 0.03021, indica una relación significante a mayor riesgo de muerte. Una mayor edad se asocia con mala supervivencia.

La presencia de no padecer diabetes reduce la razón de riesgo en un factor HR = 0.2351 o 76.49%, con un coeficiente β = - 1.4477, el signo negativo en el

coeficiente indica una disminución en el riesgo, la significancia de esta covariables destacada ya que su valor p = 0.0094, por lo que el no padecer diabetes se asocia con una buena supervivencia.

	exp(coef)	exp(-coef)	lower .95	upper .95
datos30$cancer	NA	NA	NA	NA
datos30$citologia	1.1521	0.8680	0.98035	1.3539
datos30$cervicitis	1.1368	0.8796	0.38780	3.3327
datos30$ets	0.7632	1.3103	0.33770	1.7249
datos30$ecivil	1.3100	0.7633	0.57451	2.9871
datos30$sitginec	1.0255	0.9752	0.88703	1.1855
datos30$exploracion	1.4911	0.6706	1.18914	1.8698
datos30$parejas	0.9837	1.0166	0.76513	1.2647
datos30$ivsa	0.7155	1.3976	0.55856	0.9166
datos30$edad	1.0394	0.9621	1.00370	1.0763
datos30$menarca	0.9174	1.0901	0.73791	1.1405
datos30$gestas	0.9219	1.0848	0.77983	1.0898
datos30$diabetes	0.2351	4.2534	0.07884	0.7011

```
Concordance= 0.764  (se = 0.044 )
Likelihood ratio test= 35.32  on 12 df,   p=4e-04
Wald test            = 29.53  on 12 df,   p=0.003
Score (logrank) test = 33.6  on 12 df,   p=8e-04
```

Figura 4.73 “Evaluación de Contrastes”

Ahora evaluando los contrastes, como lo vemos en la figura 4.73, se verifica que el de razón de máxima verosimilitud con p = 0.0004 < que 0.05 indica que los factores son independientes uno de otro, ahora el contraste de Wald con p=0.0003<0.05 nos dice que los coeficientes son diferentes de cero, y por último el contraste del score (logrank tes)t con p=0.0008 <0.05, nos indica que las curvas de supervivencia de los grupos con cáncer y sin cáncer difieren significativamente.

```
> #VERIFICACIÓN DE LOS SUPUESTOS DE COX
> cox.zph(cox30)
                        chisq df      p
datos30$citologia    5.63e+00  1 0.0177
datos30$cervicitis   4.91e-01  1 0.4836
datos30$ets          3.91e-01  1 0.5319
datos30$ecivil       7.95e-04  1 0.9775
datos30$sitginec     1.38e-02  1 0.9065
datos30$exploracion 7.91e-01  1 0.3737
datos30$parejas      1.52e-01  1 0.6965
datos30$ivsa         2.44e-01  1 0.6212
datos30$edad         5.56e-01  1 0.4559
datos30$menarca      9.86e+00  1 0.0017
datos30$gestas       1.67e-01  1 0.6824
datos30$diabetes     8.74e-01  1 0.3498h
GLOBAL               2.54e+01 12 0.0129
```

Figura4.74 “Verificación de los Supuestos de Cox”

Dado que los valores de $p > 0.05$ para los diferentes factores significa que las covariables analizadas son independientes entre si.

No se rechaza la hipótesis nula, por lo que no se viola el supuesto de riesgos proporcionales, ni desde el punto de vista global ni para cada covariable.