UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Modelos de supervivencia para empresas de la Zona Metropolitana del Valle de México: Análisis comparativo de métodos paramétricos y no-paramétricos

TESIS

QUE PARA OBTENER EL TÍTULO DE:

Actuario

PRESENTA:
Diego Adonai Camarillo Garrido

DIRECTOR DE TESIS:
Dr. Omar de la Riva Torres

Ciudad Universitaria, Ciudad de México, 2018

1. Datos del alumno
Camarillo
Garrido
Diego Adonai
59131170
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
310021401

2. Datos del tutor
Dr
Omar
De La Riva
Torres

3. Datos del sinodal
M en C
Antonio
Soriano
Flores

4. Datos del sinodal
Act
Francisco
Sánchez
Villarreal

5. Datos del sinodal
Act
Harim
García
Lamont

6. Datos del sinodal
M en C
Fernando Daniel
Pérez
Arriaga

7. Datos del trabajo escrito
Modelos de supervivencia para empresas de la Zona Metropolitana del Valle de México: Análisis comparativo de métodos paramétricos y no paramétricos
86p
2018

Índice General

Introducción **8**

Capítulo I Marco referencial **10**

1.1 Introducción 10
1.2 Planteamiento del estudio 11
1.2.1 Objetivos 11
1.2.2 Campo de análisis 12
1.2.2.1 Zona Metropolitana del Valle de México 12
1.2.2.2 MiPyMES en México 14
1.2.3 Estratificación de los negocios 15
1.2.4 Clasificación Sectorial 17
1.2.5 Algunos Resultados de los Censos Económicos 2014 18
1.2.6 Definición del problema 21
1.3 Fuentes de información 22
1.3.1 Esperanza de vida de los negocios en México 22
1.3.1.1 Resultados Generales 22
1.3.1.2 Tablas supervivencia 24
1.3.2 Directorio Estadístico Nacional de Unidades Económicas 26
1.3.3 Variables fuera del DENUE 27
1.3.3.1 Producto Interno Bruto per cápita 27
1.3.3.2 Índice de Marginación 28
1.3.3.3 Remuneración anual promedio por persona 29
1.4 Generación base de datos con indicador de cierre 29
1.4.1 Ajuste a los estratos de personal ocupado del DENUE 29
1.4.2 Calculo probabilidades de cierre conjuntas 30
1.4.3 Asignación de probabilidad de cierre 31
1.4.4 Selección usando un muestreo estratificado 32

Capítulo II Análisis de supervivencia **33**

2.1 Definiciones 33
2.2 Introducción a los datos de supervivencia 33
2.3 Funciones para el modelado de la supervivencia 34
2.3.1 Función de supervivencia 35
2.3.2 Función de densidad 35
2.3.3 Función de riesgo 36
2.4 Relaciones entre las funciones de supervivencia 36
2.5 Casos discretos 37

2.6 Distribuciones de los tiempos de falla 38
2.6.1 Exponencial 38
2.6.2 Weibull 39
2.6.3 Log- Logística 39
2.6.4 Log-normal 40
2.6.5 Otros modelos paramétricos 41
2.7 Procedimientos no-paramétricos 41
2.7.1 Estimador de Kaplan-Meier 42
2.7.2 Comparación de dos funciones de supervivencia 43
2.7.2.1 Prueba de Log-Rangos 43
2.7.2.2 Prueba de Wilcoxon 45
2.7.3 Comparación de tres o más grupos de supervivencia 45
2.8 Modelo de Riesgos Proporcionales 47
2.8.1 Inclusión de variables y factores en el modelo 47
2.8.2 Estimación del modelo de riesgos proporcionales 49
2.8.3 Validación del supuesto de riesgos proporcionales 50
2.8.4 Comparación de modelos alternativos 51
2.9 Modelo de vida acelerada 53
2.9.1 Representación log-lineal del modelo de vida acelerada 55
2.9.2 Modelos paramétricos de vida acelerada 56
2.9.3 Ajuste y comparación de modelos de vida acelerada 57

Capítulo III Ajuste y análisis de los modelos de supervivencia 59

3.1 Introducción 59
3.2 Análisis Exploratorio 59
3.3 Estimación no paramétrica (Kaplan Meier) 65
3.3.1 Comparación de la supervivencia entre grupos 67
3.4 Ajuste de un modelo paramétrico 70
3.4.1 Inclusión de variables explicativas 72
3.5 Análisis comparativo 74
3.6 Evolución geográfica de la supervivencia 75

Conclusiones 77

Anexos 78

Bibliografía 86

Introducción

El siguiente documento presentará el análisis para el ajuste de modelos de supervivencia no paramétricos y paramétricos a empresas de la Zona Metropolitana del Valle México con el fin de obtener información más precisa de los factores que alteran el tiempo de cierre de las mismas.

La esperanza de vida al nacer de las empresas a nivel nacional en México es de 7.8 años[1]; esta cifra varía dependiendo la entidad federativa o el sector económico en que se clasifique a la empresa; en las entidades del Estado de México y la Ciudad de México se concentra el 22.4 % [2] de las unidades económicas del país, siendo estos negocios fuentes de empleo para millones de personas, se plantea la necesidad de analizar los factores que ocasionen su cierre.

El análisis de supervivencia es una rama de la estadística aplicable en áreas como biología, demografía, ingeniería, medicina y economía, su propósito es modelar el *tiempo de falla o el tiempo cuando se observa un evento.* Existen diferentes técnicas para analizar la supervivencia basadas en estadística paramétrica y no-paramétrica, apoyándose ambas en el modelado de funciones que aportan información al *tiempo de falla* como la *función de supervivencia* y la *función de riesgo.* Una de las grandes ventajas del análisis de supervivencia es su manera de afrontar la pérdida de información del evento de estudio, fenómeno conocido como *censura,* la cual, por la naturaleza de la información, es muy común en este tipo de análisis.

Actualmente existen paquetes estadísticos que son de gran utilidad para modelar este tipo de eventos, esto ha ayudado a que diversos profesionistas, principalmente en el campo de la medicina, usen estas herramientas para obtener conclusiones propias aplicables a sus áreas de desempeño. Gracias a estas ventajas es posible analizar el problema desde las distintas técnicas para el modelado de la supervivencia y así poder elegir cuál se ajusta mejor a la realidad.

En el capítulo I de este documento se analizará algunos resultados de estudios previos en el ámbito de demografía económica en México realizados por el Instituto Nacional de Estadística y Geografía (INEGI) y cómo estos han sido la base para la generación de la información usada para el estudio presentado, posteriormente se realizará un análisis exploratorio y descripción de las variables que conforman nuestra población de estudio.

Los concepto básicos utilizados en el análisis de supervivencia, así como el sustento estadístico de los distintos modelos; las ventajas y desventajas de cada uno de ellos, se describen en el capítulo II.

[1] *INEGI, Esperanza de vida de los Negocios en México, 2016.*

[2] *INEGI, Censos Económicos 2014.*

En el capítulo III se presentarán los resultados y la metodología seguida para el ajuste de diversos modelos con el método de Kaplan Meier para distintos grupos de empresas. También se describirá el proceso para la selección de un modelo paramétrico para los tiempos de falla y finalmente se analizará el impacto de las variables que afectan dicho fenómeno.

Capítulo I

Marco referencial

1.1 Introducción

Actualmente, en la zona conformada por la Ciudad de México y el Estado de México, según información de los *Censos Económicos* realizados por el INEGI en 2014, existen alrededor de 1 millón de unidades económicas[3], las cuáles son fuente de empleo para más de 5.5 millones de personas.

En México se observa un comportamiento de muertes y nacimientos de negocios con gran frecuencia, por lo que su movilidad demográfica es considerada elevada, esta movilidad es generalmente observada en los establecimientos micro, pequeños y medianos. Del periodo 2009-2012 se observó que para las micro, pequeñas y medianas empresas (MiPyMES) en el Estado de México y en la Ciudad de México un crecimiento neto de 5 % y 0.3 %[4] respectivamente (INEGI, 2012).

La demografía económica en México hace énfasis en el estudio de los fenómenos ocurridos durante la vida de los negocios. En esta línea de investigación, el INEGI, con estudios como el de *Esperanza de vida de los negocios* y *Análisis de la demografía de los establecimientos* ha proporcionado una base para analizar más a fondo el problema de la supervivencia de los negocios.

La supervivencia de los negocios en México depende de diversos factores, entre otros, los que son más fáciles de caracterizar, debido a la información que se encuentra disponible, son:

- El tamaño del negocio
- Tipo de actividad que realiza o sector
- Ubicación geográfica

[3] Las unidades económicas comprenden los establecimientos y las empresas que comparten la misma razón social en actividades como Construcción; Transportes; Servicios financieros y de seguros; Electricidad, agua y gas; y Telecomunicaciones.

[4] *INEGI, Análisis de la demografía de los establecimientos 2012.*

1.2 Planteamiento del estudio

1.2.1 Objetivos

El objetivo general es la aplicación de los modelos estadísticos de supervivencia en una población no humana, en este caso, empresas de la Zona Metropolitana del Valle de México.

Se plantea la generación y ajuste de distintos modelos para negocios de la ZMVM, la cual incluye las 16 delegaciones de la Ciudad de México, 59 municipios del Estado de México y 1 municipio del Estado de Hidalgo; éste último no se considerará en el presente estudio debido a que su inclusión implicaba realizar un análisis completo de las tasas de supervivencia en empresas de todos los municipios del estado de Hidalgo.

La población de estudio serán los negocios de 0 a 250 personas que actualmente se encuentran registrados en el Directorio Estadístico Nacional de Unidades Económicas (DENUE).

Entre los principales objetivos de este estudio se destacan:

- Realizar una simulación del tiempo de falla de los negocios registrados en el DENUE con base en la experiencia de la esperanza de vida de los negocios por sector y tamaño, esto con el fin de generar una base de datos que sirva para ajustar los modelos de supervivencia.

- Ajuste de distintos modelos No-Paramétricos para las 2 entidades, 4 sectores y 6 grupos de tamaños de las empresas por personal ocupado, con el propósito de evaluar si existen diferencias significativas entre las variables que afectan las tasas de supervivencia entre distintos grupos.

- Verificar si es factible el ajuste de un modelo paramétrico con el fin de conocer si la utilización de una distribución de probabilidad concreta es adecuada, además de evaluar si es factible la inclusión de covariables predictivas.

- Ubicar geográficamente la función de riesgo a través del tiempo mediante mapas regionales por municipio y delegación con base en la simulación efectuada.

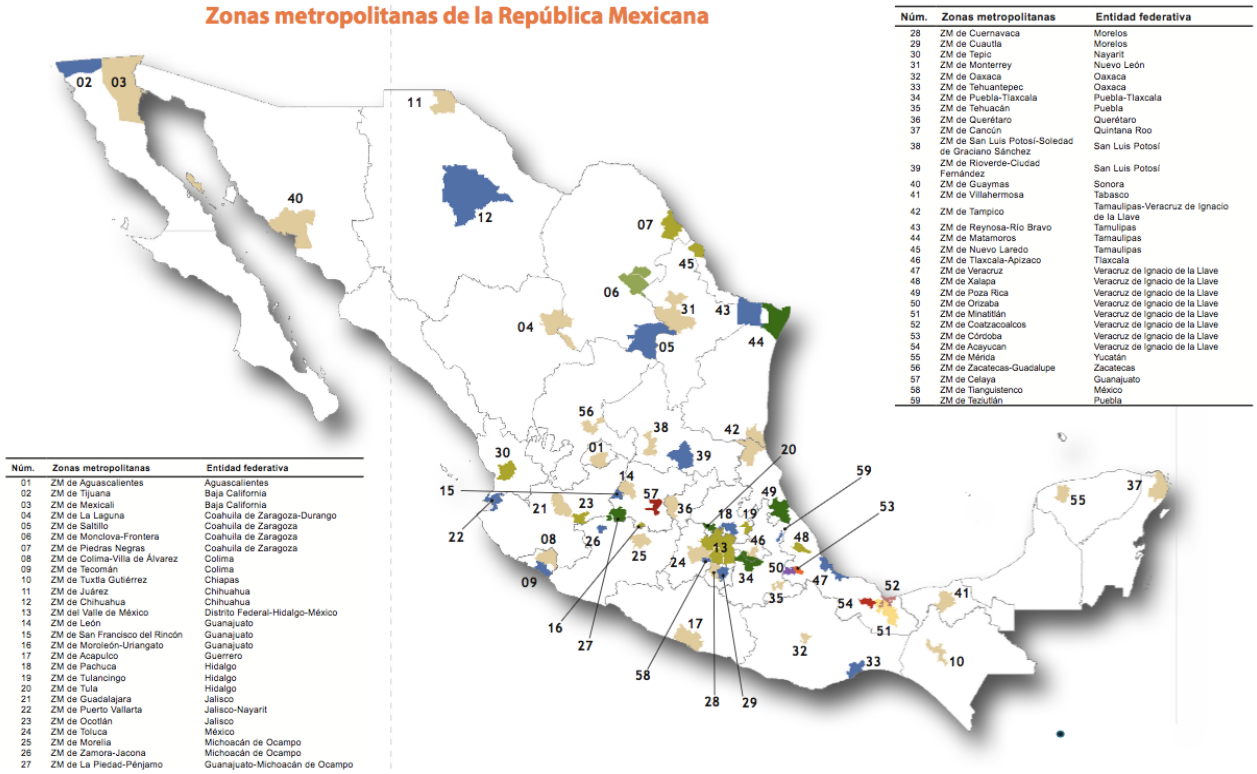
1.2.2 Campo de análisis

Se eligió el universo de MiPyMES debido a que, cómo veremos más adelante, son los establecimientos están más expuestos a detener sus actividades; el estudio abarca los municipios y delegaciones del Estado de México y Ciudad de México que forman parte de la ZMVM, esto debido a su importancia en el ámbito económico, financiero y cultural en México.

1.2.2.1 Zona Metropolitana del Valle de México

Según la definición del Consejo Nacional de Población (CONAPO): *"una zona metropolitana es el conjunto de dos o más municipios donde se localiza una ciudad de 50 mil o más habitantes, cuya área urbana, funciones y actividades rebasan el límite del municipio que originalmente la contenía, incorporando como parte de sí misma o de su área de influencia directa a municipios vecinos, predominantemente urbanos, con los que mantiene un alto grado de integración socioeconómica*". En el mapa 1.1 se muestran las 59 zonas metropolitanas de México.

Mapa 1.1 : Zonas Metropolitanas de México



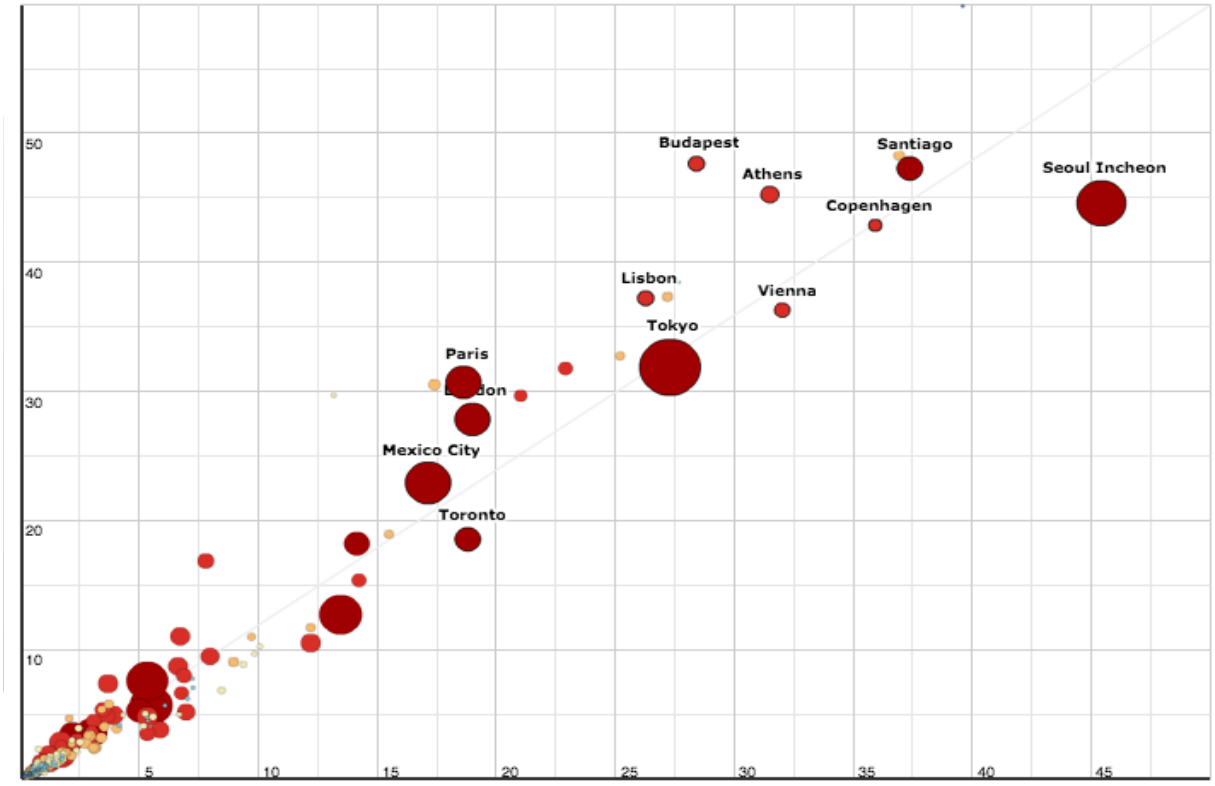
Fuente: Minimonografía. Las zonas metropolitanas en México. Censos Económicos 2014.

En México, actualmente existen 59 zonas metropolitanas. La Zona Metropolitana del Valle de México es la de mayor relevancia en el país, debido principalmente a su población; ésta cuenta con más de 20 millones de habitantes, lo que es equivalente al 17 % [5] de la población nacional.

La importancia económica de la ZMVM radica en que en ella se encuentran el mayor número de negocios y actividades económicas del país, generando cerca del 18 % de los empleos en México y 23 %[6] del Producto Interno Bruto PIB.

A continuación en la gráfica 1.1 se muestra una comparación de las principales zonas metropolitanas de la OCDE y su aportación en el PIB del país, como podemos observar, la ZMVM se encuentra por debajo de la aportación en el PIB comparado con otras con características poblacionales similares:

Gráfica 1.1: Proporción de la población nacional vs proporción del PIB de las ciudades más grandes de la OCDE, 2010



Fuente: Regions at a Glance, Metropolitan areas. OECD Regional Database

[5] *INEGI, Censo de Población y Vivienda 2010.*

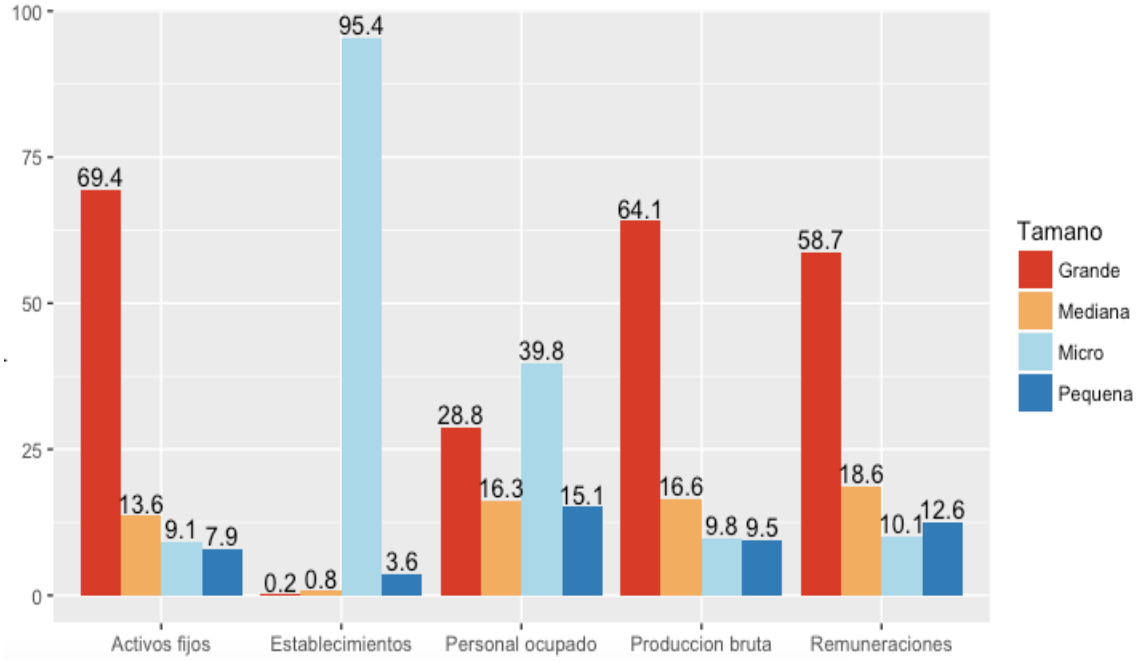
[6] *Estudios Territoriales de la OCDE , Valle de México, México.*

1.2.2.2MiPyMES en México

Conocer la dinámica de las micro, pequeñas y medianas empresas es importante es debido a que éstas representan más del 90% de las empresas alrededor del mundo, en América Latina entre el 95 – 99 % [7].

Para México, según resultados de los *Censos Económicos* de 2014, el 99.8 % de los establecimientos eran MiPyMES, teniendo un total del 71.2 % del personal ocupado, como se muestra en la Gráfica 1.2.

Gráfica 1.2: Características económicas según variables seleccionadas por tamaño de los establecimientos, 2013



Fuente: INGEI, Censos Económicos 2014.

Es común relacionar eventos demográficos, generalmente asociados a poblaciones humanas (nacimiento, muerte) a otros ámbitos como el de este tipo de análisis ya que tienen una interpretación similar.

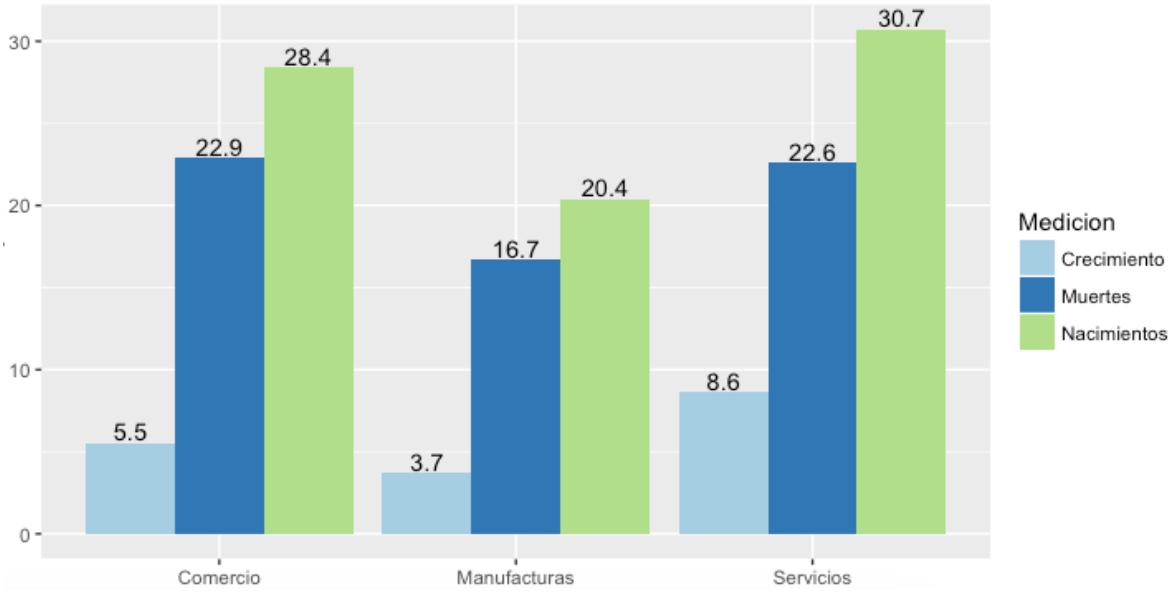
Es normal suponer que la vida de los negocios está ligado al tamaño de éstos, por lo tanto, una empresa de menor tamaño está más expuesta a finalizar sus actividades que una empresa más grande, el INEGI ha realizado estudios para determinar el grado de movilidad de los establecimientos, principalmente enfocados en los de tamaño de 0 a 100 empleados.

[7] *INEGI. Micro, pequeña, mediana y gran empresa. Estratificación de los establecimientos Censos Económicos 2104. 2015*

Según datos del estudio *Análisis de la demografía de los establecimientos 2012*, en el cual se realizó un seguimiento a una muestra de establecimientos del subconjunto de MiPyMES de 0 a 100 empleados de los sectores manufacturero, comercio y servicios no financieros; a nivel nacional, durante el periodo de abril 2009 – mayo 2012, hubo un 28.3 % de nacimientos de este tipo de establecimientos, la proporción de muertes fue del 22 % , al restar esas proporciones se obtiene un crecimiento del 6.2 % durante 37 meses.

Éstos resultados varían dependiendo del sector, siendo el de servicios no financieros el que mayor dinamismo tuvo al registrar un 8 % [8] de crecimiento neto. (Ver Gráfica 1.3)

Gráfica 1.3: Porcentaje de nacimientos y muertes de establecimientos por sector de actividad en un periodo de 37 meses



Fuente: INEGI, Análisis de la demografía de los establecimientos 2012

1.2.3 Estratificación de los negocios

Los parámetros que existen para medir el tamaño de un negocio se basan principalmente en el número de personal ocupado y el monto de ventas o ingresos anuales. El criterio utilizado actualmente en México para definir a la micro, pequeña, medianas empresas se muestra en la Tabla 1.1:

[8] *INEGI, Análisis de la demografía de los establecimientos 2012.*

Tabla 1.1: Estratificación de las empresas por sector y tamaño				
Tamaño	**Sector**	**Rango de número de trabajadores**	**Rango de monto de ventas anuales (mdp)**	**Tope máximo combinado***
Micro	Todos	Hasta 10	Hasta $4	4.6
Pequeña	Comercio	Desde 11 hasta 30	Desde $4.01 hasta $100	93
	Industria y Servicios	Desde 11 hasta50	Desde $4.01 hasta $100	95
Mediana	Comercio	Desde 31 hasta 100	Desde $100.01 hasta $250	235
	Servicios	Desde 51 hasta 100		
	Industria	Desde 51 hasta 250	Desde $100.01 hasta $250	250

Fuente: Diario Oficial de la Federación, 30 Junio 2009

Siendo Tope Máximo Combinado = (Trabajadores) X 10% + (Ventas Anuales) X 90%. El tamaño de una empresa está basado en el puntaje obtenido de la fórmula del Tope Máximo Obtenido y no deberá ser mayor a éste, considerándose como grandes empresas las que no se encuentren dentro de ésta clasificación.

Para efectos de los más recientes *Censos Económicos* realizados por el INEGI, se realiza una estratificación basada en el personal ocupado, aunque también existen otras estratificaciones basadas en ingresos, producción bruta y activos fijos. (Ver Tabla 1.2).

Tabla 1.2: Estratificación usada en *Censos Económicos*
0 a 2
3 a 5
6 a 10
11 a 15
16 a 20
21 a 30
31 a 50
51 a 100
101 a 250
251 a 500
501 a 1000
1001 y más

Fuente: INEGI, Censos Económicos 2014

La identificación de las unidades económicas conocidas como MiPyMES cuenta con criterios reconocidos por organismos como la Unión Europea y la OCDE. Estos criterios dependen del fin para el que se realice el estudio; siendo, personal ocupado, ventas anuales y el balance anual los ocupados para fines legales y administrativos. Para fines estadísticos el único criterio utilizado es el de personal ocupado[9].

[9] *INEGI. Micro, pequeña, mediana y gran empresa. Estratificación de los establecimientos Censos Económicos 2104. 2015*

1.2.4 Clasificación Sectorial

En la información recabada por los *Censos Económicos* se obtienen datos de todas las actividades económicas del país, exceptuando las agrícolas, ganaderas y forestales; las cuales son descritas en el Censo Agropecuario. Las actividades económicas que cubren los *Censos Económicos* se describen en la Tabla 1.3:

Tabla 1.3: Clasificación sectorial usada en el SCIAN 2013

Clave	Actividad
11	Pesca y agricultura
21	Minería
22	Electricidad, agua y gas
23	Construcción
31-33	Manufacturas
43	Comercio al por mayor
46	Comercio al por menor
48-49	Transportes, correos y almacenamiento
51	Información en medios masivos
52	Servicios financieros y de seguros
53	Servicios inmobiliarios y de alquiler de bienes muebles e intangibles
54	Servicios profesionales, científicos y técnicos
55	Corporativos
56	Servicios de apoyo a los negocios y manejo de servicios y desechos, y servicios de remediación
61	Servicios educativos
62	Servicios de salud y de asistencia social
71	Servicios de esparcimiento culturales y deportivos, y otros servicios recreativos
72	Servicios de alojamiento temporal y de preparación de alimentos y bebidas
81	Otros servicios excepto actividades gubernamentales
93	Actividades legislativas, gubernamentales, de impartición de justicia y de organismos internacionales y extraterritoriales

Fuente: INEGI, Censos Económicos 2014

Las claves de clasificación está basado en el Sistema De Clasificación Industrial de América del Norte (SCIAN), la versión más reciente es la de 2013. Este sistema se utiliza para analizar y difundir información estadística de una forma estandarizada para América del Norte. En la clasificación actual ocupada en el SCIAN, existen 20 sectores de actividad económica, los cuales se dividen en 94 subsectores, 303 ramas, 614 subramas y 1 059 clases de actividad, de las cuales 981 fueron objeto de los *Censos Económicos* 2014.

Es común que en los *Censos Económicos* se consideren 4 grupos de actividades económicas los cuales de ahora en adelante consideraremos sólo como sectores; Manufacturero, Comercio, Servicios y Otras actividades económicas.

Las actividades del sector manufacturero se dedican a la transformación de materias primas en nuevos productos para su distribución y consumo. La manufactura es la actividad del sector secundario de la economía, también es conocido cómo sector industrial. Este tipo de unidades económicas usan maquinaria accionada por energía y equipo manual.

El sector comercio incluye las actividades de comercio al por mayor y comercio al por menor. Comprende unidades económicas dedicadas principalmente a la compra-venta; actúan principalmente como intermediarias entre negocios.

El sector servicios es el que concentra más actividades; incluye Información en medios masivos; Servicios financieros y de seguros; Servicios inmobiliarios y de alquiler; Servicios profesionales, científicos y técnicos; Corporativos; Servicios de apoyo a los negocios y manejo de desechos; Servicios educativos; Servicios de salud y de asistencia social; Servicios de esparcimiento, culturales y deportivos; Servicios de alojamiento y de preparación de alimentos; Otros servicios, excepto gobierno.

En el Resto de los Sectores económicas se concentraron las actividades de Pesca y acuicultura; Minería; Electricidad, agua y gas; Construcción; y Transportes, correos y almacenamiento.

1.2.5 Algunos Resultados de los Censos Económicos 2014

Se denomina establecimiento a la unidad económica que en una sola ubicación física, asentada en un lugar de manera permanente y delimitada por construcciones o instalaciones fijas, combina acciones y recursos bajo la dirección de una sola entidad propietaria o controladora, para realizar actividades de producción de bienes, comercialización de mercancías o prestación de servicios, sea con fines de lucro o no.

El *Sector privado y paraestatal*[10] por unidades económicas representa el objeto de estudio sobre el panorama de las micro, pequeñas, medianas y grandes empresas, debido a que comparten una temática censal en común, la cual comprende una amplia gama de variables, a diferencia del resto de las unidades de observación.

La distribución del total de establecimientos a nivel nacional y la importancia del Sector privado y paraestatal se muestra en la Tabla 1.4.

[10] *El sector privado y paraestatal comprende a los productores de bienes o de servicios que realizan actividades económicas como personas físicas o sociedades constituidas como empresas, incluidas aquellas con participación estatal y las empresas productivas del estado cuya finalidad es la producción de bienes y servicios de mercado.*

Tabla 1.4: Distribución del total de establecimientos y personal ocupado

	Establecimientos	Personal ocupado	Personal ocupado/Establecimientos
Total de establecimientos (universo total)	**5,654,014**	**29,642,421**	**5.24**
Establecimientos que iniciaron actividades en 2014	403,728	691,663	1.71
Establecimientos con actividades en 2013	5,250,286	28,950,758	5.51
Captación por muestreo en área rural	705,135	1,839,885	2.61
Captación por recorrido total	4,545,151	27,110,873	5.96
Servicios públicos y asociaciones religiosas	283,109	5,534,515	19.55
Sector privado y paraestatal por establecimientos	4,262,042	21,576,358	5.06
Sector privado y paraestatal por unidades económicas	**4,230,745**	**21,576,358**	**5.10**

Fuente: INEGI, Censos Económicos 2014

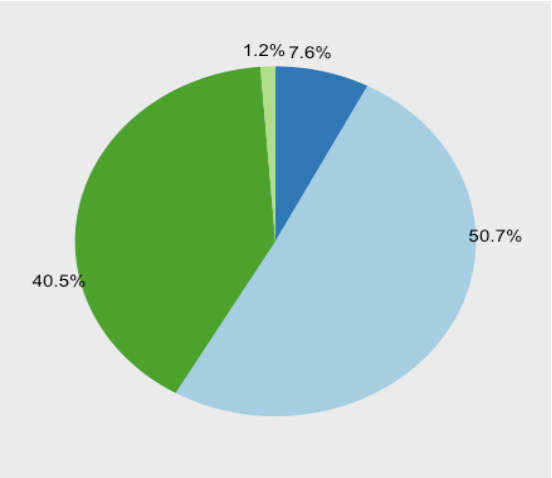
Las unidades económicas del país han presentado un crecimiento en cuanto a número de establecimientos y personal ocupado; de 2003 a 2008 se observó un crecimiento del 23.9 % tanto para el número de establecimientos como en el personal ocupado y para el periodo de 2008 a 2013 el crecimiento fue de 13.6 % y 7.3 % para establecimientos y personal ocupado respectivamente.

En el último periodo los cambios más significativos se dieron en el estrato de 0 a 2 personas se registró un 26.9 % de crecimiento en el número de establecimientos y 21.6 % de crecimiento en el personal ocupado.

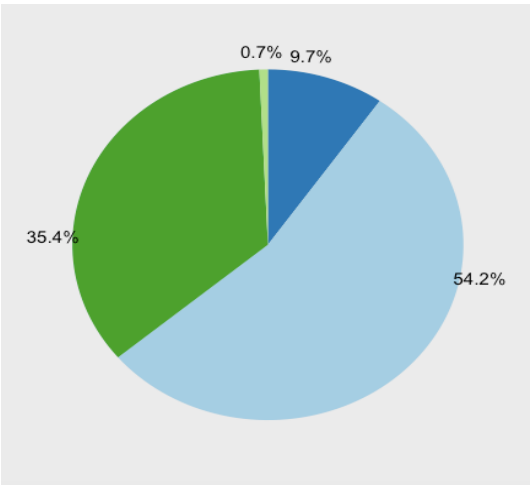
Del total del *Sector privado y paraestatal* 98.6 % de las unidades económicas y 91.1 % del personal ocupado correspondían a los sectores Manufacturero, Comercio y Servicios en conjunto; mientras que lo restante pertenece a otras actividades económicas.

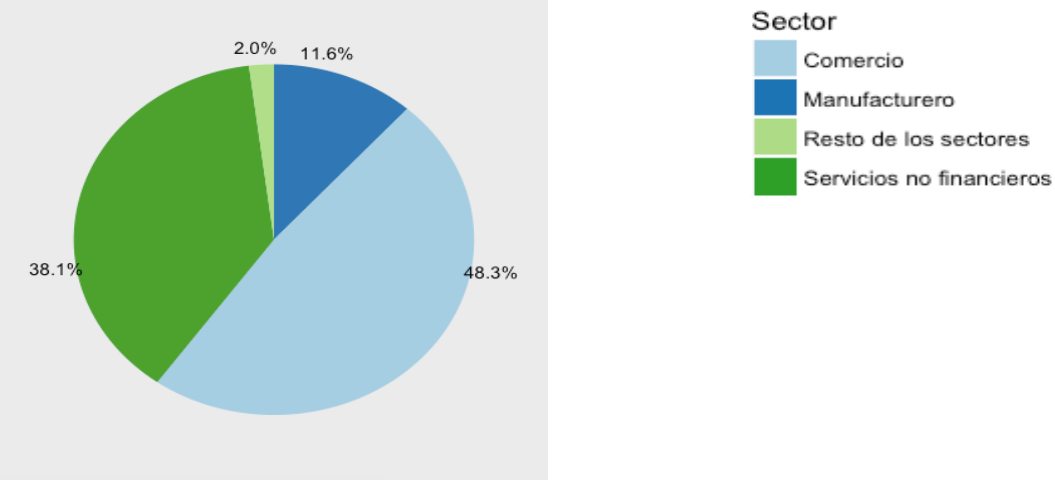
Gráfica 1.4: Estructura Sectorial (porcentaje) de las unidades económicas a nivel Nacional, Estado de México y Ciudad de México

Nacional



Estado de México





Fuente: INEGI, Censos Económicos 2014

Como se puede observar en la Gráfica 1.4 la estructura sectorial en el Estado de México y en la Ciudad de México es similar a la nacional, eso es claro si tomamos en cuenta que en estas entidades se concentra el 22.4 % del total de las unidades económicas del país.

En la sección 1.5 se detallará la distribución de las empresas por sector en la ZMVM con base en los registros del DENUE.

El Estado de México y la Ciudad de México aportan el 29.03 % de la Producción Bruta Total[11] a nivel nacional y emplean al 26.08 % del personal ocupado del país.

Otro dato interesante es el nivel de remuneraciones; es decir, el pago de retribuciones al personal, el cual representa el 34.19 % del total nacional como se muestra en la Tabla 1.5; lo que nos indica que las percepciones, son en promedio, mayores al de otras entidades federativas; cabe destacar que ésta característica sólo se observa en la Ciudad de México, que es, por debajo de Tabasco la entidad Federativa en que más remuneraciones se le pagan al personal.

[11] *Es el valor de todos los bienes y servicios emanados de la actividad económica como resultado de las operaciones realizadas por las unidades económicas, incluido el margen de comercialización de las mercancías revendidas de las firmas. Incluye: la producción realizada que no salió al mercado porque se encontraba en proceso de producción o en espera de clientes y la producción de activos fijos para uso propio. Valoración a precios productor. Se define como el monto a cobrar por el productor al comprador, menos el impuesto al valor agregado (IVA), facturado al comprador.*

Tabla 1.5: Características principales de las unidades económicas del sector privado y paraestatal que realizaron actividades en 2013, según entidad federativa.

Entidad	Unidades económicas	Personal Ocupado	Remuneraciones	Producción bruta total	Valor Agregado Censal Bruto	Activos fijos	Rentabilidad
Nacional	4,230,745	21,576,358	1,394,342.70	13,984,313.20	5,984,586.39	8,072,726.50	0.569
Cd México	415,481	3,603,572	380,315.00	2,943,782.80	1,535,020.71	2,863,865.20	0.403
México	534,838	2,023,837	96,443.60	1,116,235.40	392,363.67	431,921.10	0.685
Proporción	**22.46%**	**26.08%**	**34.19%**	**29.03%**	**29.03%**	**40.83%**	

Fuente: INEGI, Censos Económicos 2014

El Índice de rentabilidad se obtiene al restar del Valor Agregado Censal Bruto[12] las Remuneraciones y el resultado dividirlo entre los Activos Fijos, éste mide la ganancia generada por cada peso de activos fijos; estando la Ciudad de México por debajo del promedio nacional y el Estado de México arriba. La entidad federativa que mayor índice de rentabilidad tiene es Campeche generando 1.15 pesos por cada peso de activos, seguida de Tabasco (0.92 pesos) y Sonora (0.87 pesos).

1.2.6 Definición del problema

Una vez conocida la relevancia de las MiPyMES y la ZMVM en el país, es de interés conocer cuáles son los factores que afectan al cierre de este grupo de establecimientos, ya que, cómo se mencionó anteriormente tienen una repercusión directa en la economía del país y la mayoría de sus habitantes.

Para efectuar un análisis de supervivencia es necesario contar una base de datos con características particulares, la principal es conocer los tiempos de falla de los integrantes del estudio. Generalmente éste tipo de análisis se enfocan en estudios clínicos con poblaciones controladas y es información a la que el público en general no tiene acceso.

Por lo tanto, el primer reto afrontado es la generación de una base de datos que cumpla con las características requeridas para el ajuste de los distintos modelos de supervivencia con base en información pública disponible; se utilizará como punto de partida los registros del DENUE en conjunto con el estudio *Esperanza de vida de los negocios* realizado por el INEGI en su versión actualizada a 2016 para la generación de dicha base de datos.

[12] *Es el valor de la producción que se añade durante el proceso de trabajo por la actividad creadora y de transformación del personal ocupado, el capital y la organización (factores de la producción), ejercida sobre los materiales que se consumen en la realización de la actividad económica.*

Después se espera conocer si las variables disponibles en esta base puede servir para explicar la supervivencia de las empresas, o en caso contrario, buscar si existen otras que sí tengan efectos significativos.

1.3 Fuentes de información

El principal sustento de este documento se basa en el estudio realizado por el INEGI, *Esperanza de vida de los negocios en México*; ya que en éste estudio se generaron las tablas de supervivencia que posteriormente se usaron para simular las muertes de los establecimientos registrados en el DENUE y así poder obtener la base con los indicadores del tiempo en que ocurre el evento de cierre.

En la sección 1.4.2 explicará cómo se generaron las tablas de vida con las probabilidades de cierre dependiendo del sector y tamaño de la empresa.

Adicional a las variables tamaño y sector que se encuentran en el DENUE se agregaron al estudio otras a nivel municipal cómo: el PIB per cápita, el nivel de marginación (CONAPO 2010) y los niveles de ingreso promedio (INEGI 2010).

1.3.1 Esperanza de vida de los negocios en México

En este estudio se realizó el seguimiento de 21 generaciones de negocios durante los *Censos Económicos* de 1989 a 2014; cuyos resultados expresan el comportamiento de indicadores como las probabilidades de muerte, número de supervivientes y cierres, así como la esperanza de vida por edad de los negocios a nivel nacional.

La información se clasifica de la siguiente manera:

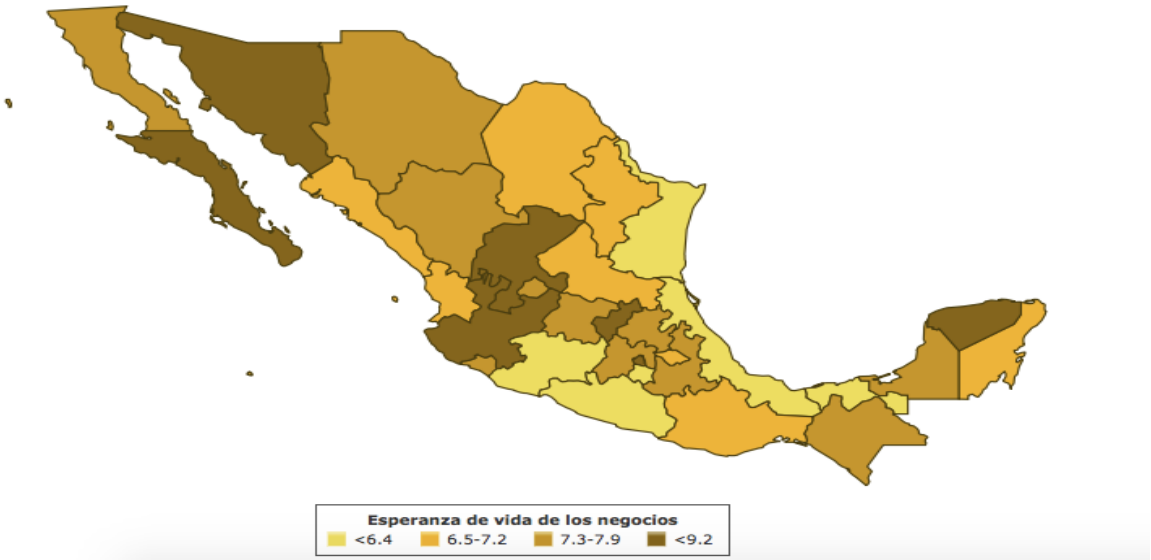
- Geográfica: Nacional y por entidad federativa.
- Sectores: Manufacturas, comercio, servicios privados no financieros y resto de los sectores.
- Personal ocupado: 0 a 2, 3 a 5, 6 a 10, 11 a 15, 16 a 20, 21 a 30, 31 a 50 , 51 a 100 y 101 a 250.

1.3.1.1 Resultados Generales

Entre los resultados a destacar del estudio se encuentra que la esperanza de vida al nacer de los negocios a nivel nacional es de 7.8 años, en el Estado de México se repite el mismo número,

mientras que en la Ciudad de México se obtuvo una esperanza de vida al nacer de 8.1 años. En el Mapa 1.2 se describen los rangos de esperanza de vida de las MiPyMES de las 32 entidades federativas de la República Mexicana.

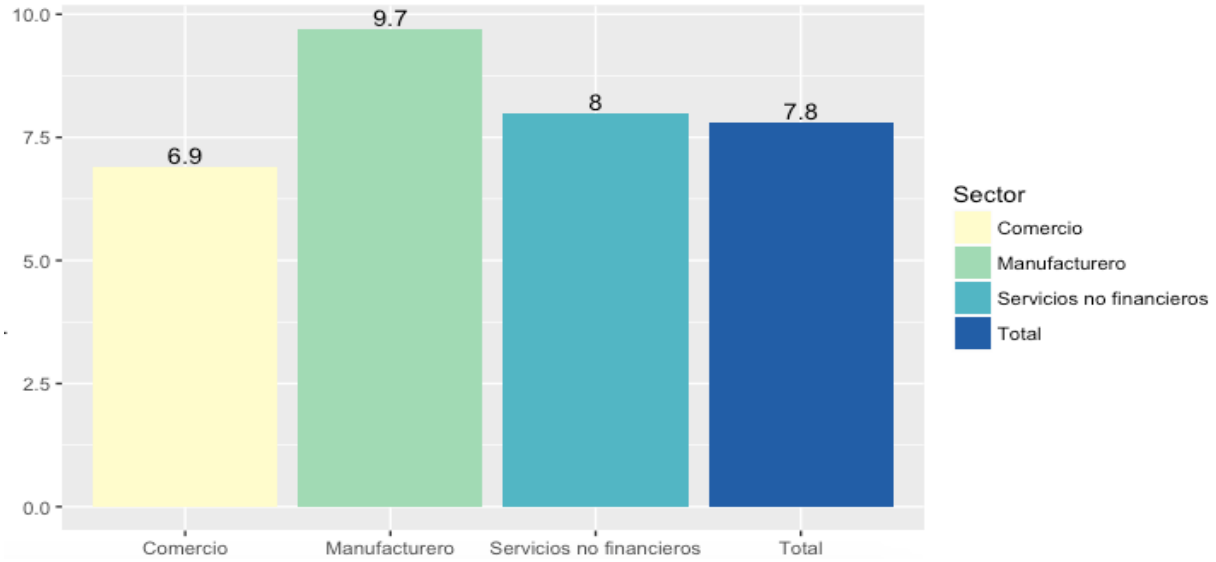
Mapa 1.2 : Esperanza de vida al nacer de los negocios por entidad federativa



Fuente: INEGI, Esperanza de vida de los Negocios en México, 2016.

Mientras que por sector, a su nacimiento, el manufacturero es el que más expectativa de vida tiene con 9.7 años, le siguen los servicios no financieros con 8 años, y por último el sector comercio con 6.9 años.(ver Gráfica 1.4).

Gráfica 1.4: Esperanza de vida al nacer por sector económico



Fuente: INEGI, Esperanza de vida de los Negocios en México, 2016.

1.3.1.2 Tablas supervivencia

Las tablas de supervivencia generadas en el estudio de la *Esperanza de vida de los negocios en México* se encuentran disponibles por entidad federativa, sector y tamaño.

En total se descargaron 28 tablas, las cuales incluyen :

- 2 tablas para conocer la probabilidad de que la empresa cierre sin considerar el sector ni el tamaño de la empresa, una para la Ciudad de México y otra para el Estado de México.

- 8 tablas para conocer la probabilidad de que la empresa cierre dado que pertenece a un sector, 4 para cada entidad federativa.

- 18 tablas para conocer la probabilidad de que la empresa cierre dado que tiene un tamaño específico, 9 para cada entidad federativa.

Posteriormente, basándose en éstas tablas, se calcularon las probabilidades conjuntas de supervivencia de cada estrato[13]. En la sección 1.4 se describirá la metodología usada para calcular estas probabilidades.

Para calcular el número de supervivientes en las tablas del estudio de la *Esperanza de vida de los negocios*, se realizó un seguimiento generacional de los últimos 6 *Censos Económicos* (1989 – 2014), debido a que estos se realizan cada 5 años, se calculó la probabilidad de sobrevivir entre censo y censo, esto es, la probabilidad de alcanzar la edad $x+5$, siendo que en el censo anterior el negocio tenía la edad x.

De esta manera se estudiaron 21 generaciones desde 1983 hasta 2003 y una vez concentradas las probabilidades de supervivencia cada 5 años, se obtuvo una estimación lineal de la supervivencia anual para cada generación.

Conociendo las probabilidades anuales de supervivencia y partiendo de un radix[14] de 100,000 negocios, se estimó el número de supervivientes de cada generación. Posteriormente se promedió el número de negocios supervivientes de cada generación y se obtuvo una estimación única de la supervivencia para cada categoría.

A partir del número de supervivientes obtenidos de las tablas del estudio realizado por el INEGI, es fácil calcular el número cierres y la probabilidad de cierre anual de los negocios. Se muestra como ejemplo en la Tabla 1.6 los negocios del sector comercio de la Ciudad de México.

---

[13] *Subconjunto de la población definido por las combinaciones de sectores y tamaño del personal.*

[14] *Número de negocios supervivientes a edad cero.*

Tabla 1.6: Tabla de supervivencia de los negocios del Sector Comercio de la Ciudad de México, obtenida a través de seis *Censos Económicos* de 1989 al 2014.

Edad	S_x	d_x	q_x
0	100 000	32 797	0.328
1	67 203	18 066	0.269
2	49 138	6 182	0.126
3	42 956	4 386	0.102
4	38 569	3 402	0.088
5	35 167	2 780	0.079
6	32 387	2 350	0.073
7	30 037	2 036	0.068
8	28 001	1 796	0.064
9	26 205	1 606	0.061
10	24 599	1 453	0.059
11	23 146	1 327	0.057
12	21 819	1 220	0.056
13	20 599	1 130	0.055
14	19 469	1 052	0.054
15	18 417	984	0.053
16	17 433	924	0.053
17	16 509	871	0.053
18	15 637	824	0.053
19	14 813	782	0.053
20	14 031	744	0.053
21	13 287	709	0.053
22	12 578	678	0.054
23	11 900	649	0.055
24	11 251	622	0.055
25	10 629	598	0.056
26	10 031	0	0.000
27	10 031	0	0.000
28	10 031	0	0.000
29	10 031	0	0.000
30	10 031	0	0.000
31	10 031	0	0.000
32	10 031	0	0.000
33	10 031	0	0.000
34	10 031	0	0.000
35	10 031	0	0.000

Fuente: INEGI, Censos Económicos 1989, 1994, 1999, 2004, 2009 y 2014.

Dónde:

S_x = Negocios supervivientes al final de la edad x

d_x = Número de negocios que cierran entre las edades x *y* $x+1$

q_x = Probabilidad de que el negocio cierre entre las edades x *y* $x+1$

Formulación de los indicadores:

$d_x = S_x - S_{x+1}$
$q_x = d_x / S_x$

El resto de tablas utilizadas se encuentra disponible en los anexos.

1.3.2 Directorio Estadístico Nacional de Unidades Económicas

El Directorio Estadístico Nacional de Unidades Económicas es una herramienta facilitada por el INEGI dónde se tienen datos de identificación, ubicación, actividad económica y tamaño de los negocios activos en el territorio nacional. Su última actualización fue realizada en el segundo semestre de 2016.

A nivel nacional cuenta con 5,032,503 negocios registrados. Para la Ciudad de México existen datos de 464,578 establecimientos mientras que para el Estado de México se encuentran registrados 613,120.

Según datos de los *Censos Económicos* de 2014, en la Ciudad de México se encontraban operando 452,939 establecimientos mientras que el Estado de México la cifra era de 664,786. Por lo tanto es claro que volumen de información disponible en el DENUE es muy cercano al real.

En general, la información de la que se dispone para los establecimientos en el DENUE es: nombre y razón social, clave y descripción de actividades del SCIAN, estrato de personal ocupado[15], referencias de ubicación, clave y nombre de la entidad federativa, clave y nombre del municipio/delegación, datos de contacto cómo: teléfono, página web y correo electrónico, además de coordenadas para conocer su ubicación exacta.

En el análisis exploratorio de la base de datos se describirán las variables seleccionadas que servirán para realizar el análisis de supervivencia de los negocios.

[15] *NOTA: La estratificación de personal ocupado disponible en la base de datos del DENUE no es la misma de las tablas de supervivencia elaboradas con base en el estudio Esperanza de vida de los negocios en México. En el DENUE se clasifica al personal ocupado de la siguiente manera: 0 a 5, 6 a 10, 11 a 30, 31 a 50, 51 a 100, 101 a 250 y más de 251 personas. Por ese motivo se realizó un ajuste a algunas de las tablas de supervivencia descargadas para que coincidieran con la clasificación del DENUE.*

1.3.3 Variables fuera del DENUE

Además de las variables entidad federativa, sector económico y personal ocupado disponibles en el DENUE, se agregaron otras a nivel municipal para hacer una mejor identificación de las características de los negocios en la ZMVM, esto con el fin de ver si son útiles cómo variables explicativas para modelar la supervivencia en el ajuste de un modelo paramétrico.

1.3.3.1 Producto Interno Bruto per cápita

Según la definición del Instituto Nacional para la Evaluación de la Educación (INEE); *El PIB per cápita es la relación entre el valor total de mercado de todos los bienes y servicios finales generados por la economía de una nación, durante un año, y el número de habitantes de ese año.* El PIB per cápita representa el valor monetario de los bienes y servicios finales generados en una región, generalmente un país, que le correspondería a cada habitante en un año dado si esa riqueza se repartiera de manera proporcional.

Este indicador resulta al dividir el Producto Interno Bruto entre la población total estimada a mitad del año evaluado. Es un buen indicador macroeconómico de un país; sin embargo no toma en cuenta desigualdades en la distribución de la riqueza.

Se puede expresar en pesos corrientes, pesos a precios de un año base; con el fin de descontar el efecto inflacionario y en dólares.

Con fines de agregar información que sea útil para modelar la supervivencia de los negocios se calculó una aproximación al PIB municipal per cápita[16]:

Primero se utilizó cómo base la estimación a pesos corrientes del Producto Interno Bruto por entidad federativa para 2015. Para la Ciudad de México de 2,866,254 millones de pesos y para el Estado de México de 1,622,186[17].

Para la población por entidad, la Ciudad de México en 2015 registraba un total de 8,918,653 habitantes y el Estado de México de 16,187,608[18]. Esta información se encuentra desagregada municipalmente.

Por último; para calcular el PIB municipal, se distribuyó el PIB cada entidad entre la participación de cada municipio en el Valor Agregado Censal Bruto (VACB), disponible de manera municipal en los *Censos Económicos* de 2014.

---

[16] *NOTA: Esta aproximación no es una cifra oficial; sin embargo, para fines de esta tesis, cumple con su propósito de caracterización a nivel municipal.*

[17] *Sistema Nacional de Cuentas Nacionales de México (SNCM), cifras preliminares.*

[18] *INEGI, Encuesta Intercensal 2015.*

Con esta información se estimó PIB per cápita de manera municipal para la Ciudad de México y para el Estado de México.

1.3.3.2 Índice de Marginación

El Índice marginación es una medida que permite diferenciar entidades y municipios con respecto a las carencias de su población a nivel educativo, residencia en viviendas inadecuadas, percepciones insuficientes y residencia en localidades pequeñas.

Se utilizó la información del índice de marginación proporcionada por CONAPO a nivel municipal de 2015, ya que es la información disponible más reciente. El índice está calculado con base en 10 indicadores socioeconómicos, que se describen en la Tabla 1.6

Tabla 1.6: Indicadores socioeconómicos medidos en el Índice de marginación por entidad federativa y municipio.

Dimensión socioeconómica	Indicador
Educación	% Población de 15 años o más analfabeta
	% Población de 15 años o más sin primaria completa
Vivienda	% Ocupantes en viviendas sin drenaje ni excusado
	% Ocupantes en viviendas sin energía eléctrica
	% Ocupantes en viviendas sin agua entubada
	% Viviendas con algún nivel de hacinamiento
	% Ocupantes en viviendas con piso de tierra
	% Ocupantes en viviendas sin drenaje ni servicio sanitario exclusivo
Distribución de la población	% Población en localidades con menos de 5 000 habitantes
Ingresos monetarios	% Población ocupada con ingreso de hasta 2 salarios mínimos

Fuente: CONAPO, Índice de marginación por entidad federativa y municipio 2010

El cálculo del índice de marginación está basado en un análisis de componentes principales[19] y se clasifica en 5 niveles: muy alto, alto, bajo, medio y muy bajo.

Dentro del Estado de México y la Ciudad de México el municipio que mayor índice de marginación tiene es Donato Guerra, mientras que el menor es el de la delegación Benito Juárez.

[19] *CONAPO, Índices de marginación 2000 e Índices de marginación 2010, Anexo metodológico.*

1.3.3.3 Remuneración anual promedio por persona

La remuneración anual promedio por persona remunerada se obtiene al dividir el monto total de remuneraciones anuales entre el número del personal ocupado remunerado.

Las remuneraciones, aparte de representar el poder adquisitivo de los trabajadores también son un indicador del costo de la mano de obra de los trabajadores de los establecimientos, ya que se encuentran integradas por sueldos y salarios, prestaciones sociales, y utilidades repartidas a los trabajadores[20].

Se construyó este indicador con la información de los *Censos Económicos 2014* para los municipios y delegaciones del Estado de México y la Ciudad de México.

La remuneración anual promedio por persona remunerada en la Ciudad de México es de 167,130 pesos, mientras que en el Estado de México es de 103,177 pesos.

1.4 Generación base de datos con indicador de cierre

A continuación se enlistan los pasos seguidos para la generación de la base de datos con el indicador de cierre:

1.4.1 Ajuste a los estratos de personal ocupado del DENUE

Fue necesario ajustar la experiencia de la supervivencia por personal ocupado obtenida en las tablas del estudio de la *Esperanza de vida de los negocios en México* debido a que la estratificación ocupada no es la misma que la disponible en los registros del DENUE.

Para realizar este ajuste se sumaron los supervivientes de los grupos que coincidían y a partir del número sumado de supervivientes:

Para las tablas de supervivencia de los negocios de 0 a 2 y 3 a 5 personas ocupadas; se formó un nuevo grupo de 0 a 5 personas ocupadas y de los negocios de 11 a 15, 16 a 20 y 21 a 30 personas ocupadas; se formó un nuevo grupo de 11 a 30 personas ocupadas, esto es:

$Sx_1 =$ Negocios de 0 a 2 personas ocupadas supervivientes al final de la edad x
$Sx_2 =$ Negocios de 3 a 5 personas ocupadas supervivientes al final de la edad x

[20] *INEGI, Censos Económicos 2014.*

$Sx = Sx_1 + Sx_2 =$ Negocios de 0 a 5 personas ocupadas supervivientes al final de la edad x

De manera análoga se calcularon los negocios supervivientes de 11 a 30 personas. Esto se hizo para cada entidad federativa. A partir de este nuevo número de supervivientes se calcularon los indicadores d_x y q_x como se definieron previamente.

1.4.2 Calculo probabilidades de cierre conjuntas

Para el cálculo de las probabilidades conjuntas de cierre por estrato definido por los cruces de sector económico y personal ocupado, es necesario definir los siguientes eventos:

$X =$ El establecimiento cierra
$A_i =$ El establecimiento pertenece al sector económico i con $i=1,\ldots,4$

1= Comercio
2= Manufacturero
3= Servicios no financieros
4= Resto de los sectores

$B_j =$ El establecimiento pertenece al estrato de personal ocupado j con $i=1,\ldots,5$

1= 0 a 5 personas
2= 6 a 10 personas
3= 11 a 30 personas
4= 31 a 50 personas
5= 51 a 100 personas
6=101 a 250 personas

La probabilidad de interés es que el establecimiento cierre dado que pertenece al sector económico i y al estrado de personal j, es decir:

$$P\left[X|A_iB_j\right]$$

Basándose en las tablas de supervivencia previamente descritas se obtuvieron las siguientes probabilidades:

$P(X/A_i)=$ Probabilidad de cierre dado que el establecimiento pertenece al sector i
$P(X/B_j)=$ Probabilidad de cierre dado que el establecimiento pertenece al estrato de personal j

Las probabilidades de interés se calcularon de la siguiente forma:

$$P[X \mid A_i B_j] \propto \frac{P[X \mid A_i] P[X \mid B_i]}{P[X]}$$

donde *P(X)* Es la probabilidad de que el establecimiento cierre sin considerar sector económico ni personal ocupado; en las tablas de supervivencia está indicada cómo la probabilidad de cierre de la entidad federativa.

El resultado anterior se justifica de la siguiente forma:

Se supone que se conocen las distribuciones condicionales de *P(X/Y)* y *P(X/Z)*, usando el teorema de Bayes

$$P(X|Y,Z) = \frac{P(X,Y,Z)}{P(Y,Z)} = \frac{P(Y,Z|X)P(X)}{P(Y,Z)} \tag{1}$$

Asumiendo independencia condicional de Y y Z dado X

$$P(X|Y,Z) = \frac{P(Y|X)P(X)P(Z|X)P(X)}{P(Y,Z)P(X)}$$

dado que

$$P(X|Y)P(Y) = P(Y|X)P(X)$$

y

$$P(X|Z)P(Z) = P(Z|X)P(X)$$

Por lo que, por la Ecuación (1) se tiene

$$P(X|Y,Z) = \frac{P(Y)P(Z)}{P(Y,Z)} \times \frac{P(X|Y)P(X|Z)}{P(X)} \propto \frac{P(X|Y)P(X|Z)}{P(X)}$$

1.4.3 Asignación de probabilidad de cierre

Es necesario definir una edad inicial de cero a todos los establecimientos dentro de la base de datos para asignar una probabilidad de cierre inicial basándose en las probabilidades previamente calculadas.

1.4.4 Selección usando un muestreo estratificado

Una vez que se tiene asociada la probabilidad de cierre para todos los registros de la base de datos se calcula el número de eventos de cierre para cada estrato definido por sector económico, personal ocupado y edad con base en la probabilidad de cierre asociada.

Posteriormente se hace la selección de las unidades que cierran a partir de un muestreo estratificado, considerando como tamaño de la muestra o número de empresas seleccionadas para cerrar de cada estrato el número de eventos de cierre previamente calculado.

Los establecimientos que fueron seleccionados se acumulan en una nueva base de datos y se les asigna un indicador (1) del evento mientras los establecimientos que no fueron seleccionados aumentan su edad en 1 año.

Este proceso se realiza hasta la edad 25, ya que, a partir de esa iteración no todos lo estratos encuentran eventos. Al finalizar la última iteración a los eventos que no experimentaron el evento de cierre se les asigna un indicador (0) de censura.

Posteriormente se juntan las bases donde se acumularon los establecimientos que cerraron y los censurados para obtener base de datos con el indicador del evento de cierre y la edad a la que se experimentó.

Este mismo procedimiento se realiza por separado para la Ciudad de México y para el Estado de México.

Capítulo II

Análisis de supervivencia

2.1 Definiciones

Evento: suceso de interés en el estudio, generalmente asociado a una falla (muerte, recaída, descompostura).

Inicio del estudio: fecha de diagnóstico, de inicio del tratamiento o de remisión completa.

Final del periodo: fecha de culminación del estudio.

Seguimiento: es la observación de los individuos de un grupo a partir de la fecha inicial, para conocer su estado.

Período de seguimiento: el tiempo transcurrido entre la fecha de inicio y la fecha de corte del estudio.

Tiempo de supervivencia: es el intervalo de tiempo transcurrido entre el inicio y la fecha de última noticia.

Censura: fenómeno que ocurre cuando el valor de una observación sólo se conoce parcialmente.

Función de riesgo: función que mide la probabilidad de que a un individuo le ocurra cierto suceso a lo largo del tiempo.

Función de supervivencia: la función que mide la probabilidad de que un sujeto sobreviva más allá de un periodo de tiempo dado.

2.2 Introducción a los datos de supervivencia

En diversos campos de estudio es necesario conocer el momento de ocurrencia de un evento de interés; en el análisis de supervivencia se pretende inferir el tiempo en que tarda en ocurrir dicho evento.

El análisis de supervivencia es una forma de modelar el tiempo de falla utilizando información de sucesos que han ocurrido en circunstancias similares.

Una de las principales características de los datos del análisis de supervivencia es la frecuencia de los datos censurados:

Decimos que un dato es censurado cuando no podemos conocer el estado final para el individuo en estudio; esto puede ocurrir, en el caso de analizar la supervivencia, cuando el individuo aún se encuentra con vida al final del estudio o porque se le ha dejado de dar seguimiento por distintas razones, por ejemplo cuando se realiza un estudio clínico y un paciente cambia de residencia.

Existen tres tipos de censura:

Censura por la derecha: un ejemplo de la censura por la derecha se da en cada uno de los ejemplos anteriores. Un individuo entró al estudio al tiempo t_0 y muere en el tiempo $t_0 + t$; sin embargo, el tiempo t es desconocido, ya sea porque el individuo aún se encuentra con vida o porque se le ha dejado de dar seguimiento. Es el tipo más común de cesura.

Censura por la izquierda: Suceden cuando el evento de interés ha ocurrido antes del tiempo t , un ejemplo de este tipo de censura es en un estudio de pacientes con cáncer; el interés es conocer el tiempo de recurrencia después de una primer cirugía. Después de cierto tiempo de la operación se examina a los pacientes para determinar si ha reaparecido la enfermedad; en este caso el tiempo de recurrencia es menor al tiempo en que se realizó la primer examinación, por lo tanto es desconocido.

Censura de intervalo: En este tipo de censura sólo se conoce que la falla ocurre dentro de cierto intervalo de tiempo, se puede ilustrar con el ejemplo anterior: un paciente puede estar libre de la enfermedad pasada la primer examinación; sin embargo, al tiempo que se realiza la segunda ya cuenta con la enfermedad; en este caso se conoce que el tiempo de falla ocurrió dentro de la primer y segunda examinación.

2.3 Funciones para el modelado de la supervivencia

Es necesario definir una variable aleatoria T la cual denotará el tiempo de falla. Existen 3 funciones principalmente útiles para modelar el tiempo de falla. Estas funciones pueden ser utilizadas para ilustrar diferentes aspectos de los datos, además, conociendo una de ellas se pueden derivar las otras.

2.3.1 Función de supervivencia

El tiempo de supervivencia de un individuo, *t,* es un valor que puede tomar la variable aleatoria *T*. Los diferentes valores que puede tomar *T* siguen una distribución de probabilidad la cual tiene asignada una función de densidad denotada por *f(t)* . Entonces, la función de distribución de T está dada por:

$$F(t) = P[T < t] = \int_o^t f(u)du,$$

y representa la probabilidad de que el tiempo de supervivencia sea menor que *t*.

S(t) es definida como la probabilidad de que un individuo sobreviva más allá del tiempo *t, por lo tanto:*

$$S(t) = P[T \geq t] = 1 - F(t)$$

La *función de supervivencia S(t)* puede entonces ser usada para representar la probabilidad de que un individuo sobreviva del tiempo origen al tiempo *t*.

Por tales características *S(t)* cumple con las siguientes propiedades:

- *S(t) es no creciente*
- *S(t)=1 cuando t=0*
- *S(t)=0 cuando* $t \to \infty$

Esto es, la probabilidad de supervivencia en el tiempo 0 es 1 y además si se pudiera observar durante un tiempo infinito nadie sobrevivirá.

Cuando *S(t)* es estimada a partir de datos reales, la representación gráfica corresponde a una función escalonada y se denota por $\widehat{S}(t)$.

2.3.2 Función de densidad

La *función de densidad f(t)* es la probabilidad no condicional de que un evento ocurra exactamente en el tiempo *t, es decir:*

$$f(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t]}{\Delta t}$$

Algunas propiedades de la función de densidad son:

- *f(t)≥ 0, para todo t ≥ 0*
- *f(t)= 0, para todo t < 0*

De la *función de densidad* puede ser encontrada la proporción de individuos que cae en un intervalo de tiempo.

2.3.3 Función de riesgo

La *función de riesgo h(t)* es obtenida de la probabilidad de que un individuo muera a tiempo t, condicionado a que haya sobrevivido a ese momento.

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t \mid T \geq t]}{\Delta t}$$

El comportamiento de esta función puede incrementar, disminuir o permanecer constante, su única restricción es que no sea negativa. Sirve para observar como las probabilidades de experimentar el evento cambian durante el transcurso del tiempo.

2.4 Relaciones entre las funciones de supervivencia

Como se mencionó anteriormente, a partir de alguna de las funciones de supervivencia básicas se pueden obtener las demás.

1. $h(t) = \dfrac{f(t)}{S(t)}$

$$h(t) = \lim_{\Delta t \to 0} = \frac{P[t \leq T < t + \Delta t \mid T \geq t]}{\Delta t} = \lim_{\Delta t \to 0} = \frac{P[t \leq T < t + \Delta t, T \geq t]}{\Delta t \cdot P[T \geq t]} =$$

$$= \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t]}{\Delta t} \cdot \frac{1}{P[T \geq t]} = f(t) \cdot \frac{1}{S(t)} = \frac{f(t)}{S(t)}$$

2. $f(t) = -\dfrac{dS(t)}{dt}$

$$f(t) = \frac{dF(t)}{dt} = \frac{d(1 - S(t))}{dt} = -\frac{dS(t)}{dt}$$

3. $h(t) = -\dfrac{d}{dt} lnS(t)$

$$h(t) = \frac{f(t)}{S(t)} = -\frac{dS(t)}{dt} \cdot \frac{1}{S(t)} = -\frac{d}{dt} lnS(t)$$

4. $S(t) = exp\left[-\int_o^t h(t)\right]$

$$h(t) = -\frac{d}{dt} lnS(t)$$

La función de riesgo acumulado se define cómo

$$H(t) = \int_0^t h(u)du = -ln(S(t))$$

entonces

$$S(t) = exp\left\{-H(t)\right\}$$

2.5 Casos discretos

En el caso de que la variable aleatoria T sea discreta, las *funciones de densidad, riesgo y supervivencia* están dadas por:

$$f(t_j) = P\left[T = t_j\right] = f(t_j) = S(t_{j-1}) - S(t_j)$$

$$h(t_j) = P\left[T = t_j | \, T \geq t_j\right] = \frac{f(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}$$

$$S(t_j) = \left[1 - h(t_j)\right] S(t_{j-1}) \; ; \qquad S(t) = P\left[T > t_j\right] = \sum_{t_j > t} f(t_j) = \prod_{t_j \leq t} \frac{S(t_{j-1})}{S(t_j)}$$

2.6 Distribuciones de los tiempos de falla

En ocasiones los tiempos de falla pueden ser descritos por familias de distribuciones de probabilidad dependientes de uno o más parámetros.

Para realizar la selección de un modelo paramétrico de supervivencia lo más común es observar el comportamiento de la función de riesgo; por ejemplo, si lo que se está analizando es la supervivencia de pacientes después de una cirugía, es común suponer que el riesgo de muerte sea más alto en las primeras horas y se estabilice y disminuya en tiempos posteriores; en este caso sería necesario encontrar una distribución cuya función de riesgo sea creciente hasta alcanzar un punto máximo y después decreciente. En otras ocasiones podremos encontrar tiempos de falla en las que el riesgo no dependa del tiempo y el evento de falla pueda ocurrir en cualquier momento; en estos casos lo normal es suponer que la función de riesgo es constante.

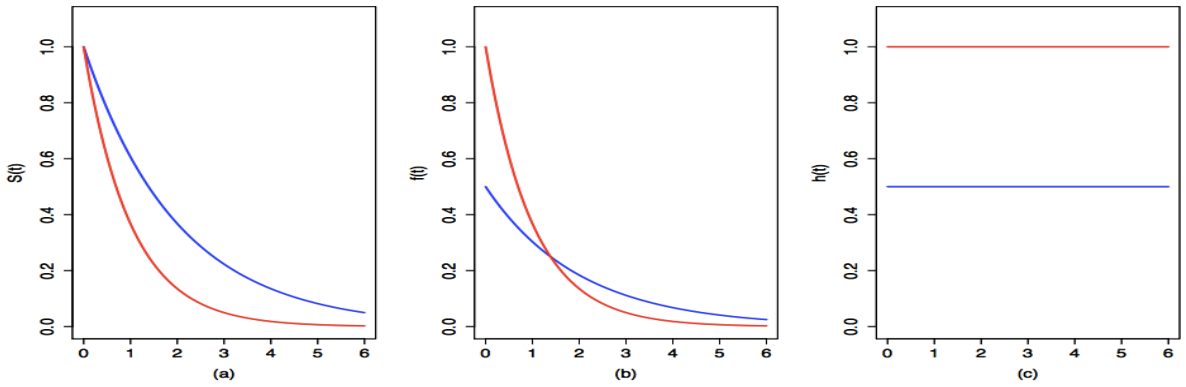
A continuación se enlistarán algunas de las más comunes distribuciones de tiempos de falla paramétricos y se describirán sus principales características.

2.6.1 Exponencial

Es la más simple de las distribuciones de supervivencia, la edad del individuo no afecta la probabilidad de supervivencia por lo que su función de riesgo es constante. A finales de 1940 investigadores la empezaron a utilizar para modelar el tiempo de falla de sistemas electrónicos. Tiene un único parámetro λ de escala.

Las *funciones de supervivencia, densidad y riesgo* están dadas por:

$$S(t) = e^{-\lambda t} \qquad f(t) = \lambda e^{-\lambda t} \qquad h(t) = \lambda$$

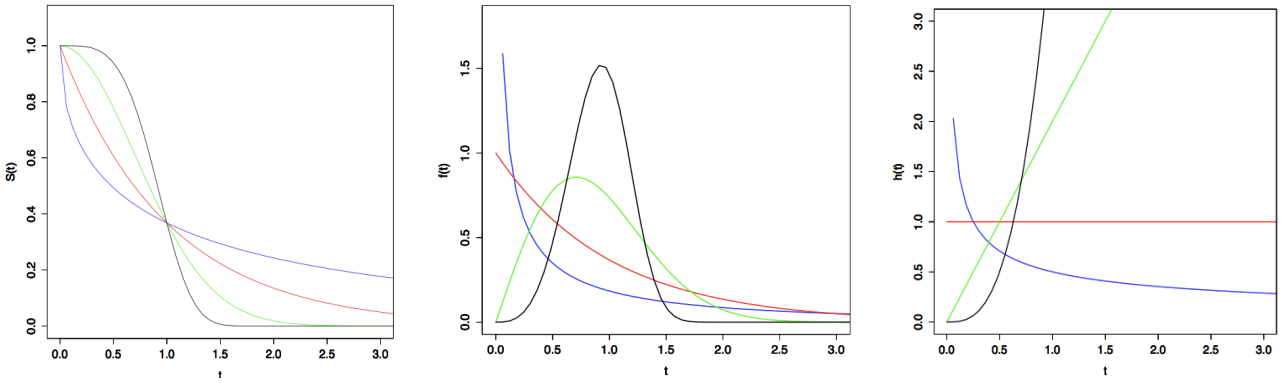


En las figuras anteriores se muestra la forma de las funciones características de la distribución exponencial con λ=1(rojo) y λ=0.5(azul).

2.6.2 Weibull

Es una generalización de la distribución exponencial, a diferencia de esta, no asume una comportamiento constante de la función de riesgo. Ha sido utilizada ampliamente en estudios de rehabilitación. La distribución está caracterizada por los parámetros de forma γ y de escala λ. Sus *funciones de densidad, supervivencia y riesgo son:*

$$S(t) = e^{-(\lambda t)^{\gamma}} \qquad f(t) = \lambda\gamma(\lambda t)^{\gamma-1}e^{-(\lambda t)^{\gamma}} \qquad h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$$

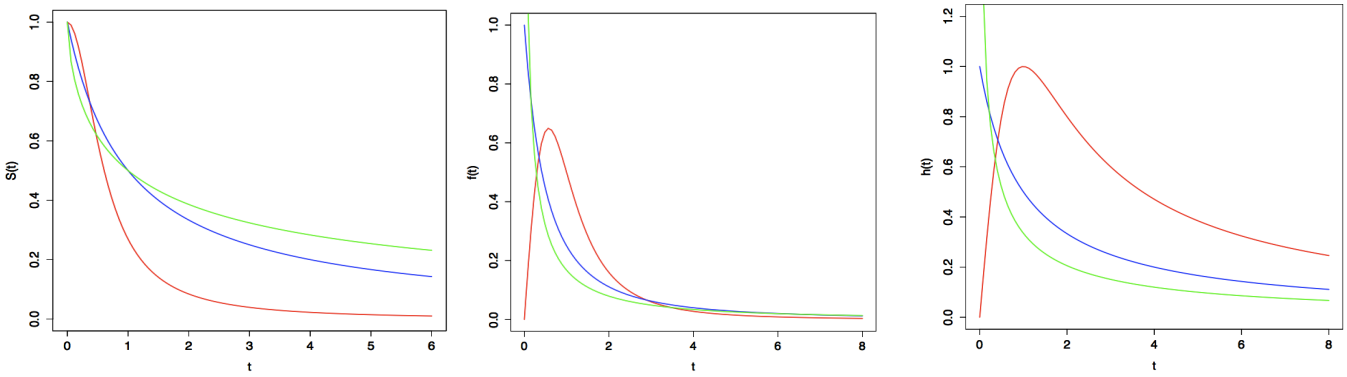


En las figuras anteriores se muestran las formas de las funciones características de la supervivencia con λ=1 y , γ=0.5 (azul), γ=1.0 (rojo), γ=2.0 (verde) y γ=4.0 (negro). Como se puede observar la distribución exponencial es un caso particular donde γ=1.

2.6.3 Log- Logística

El tiempo de supervivencia T se dice que tiene una distribución log-logística si $\log T$ tiene una distribución logística. Está definida por los parámetros de escala α y de forma γ. Es uno de los modelos paramétricos de supervivencia en el que la *función de riesgo* es decreciente para $\gamma \leq 1$ y tiene forma de 'joroba' para $\gamma > 1$. Sus funciones características son de las siguientes formas:

$$S(t) = \frac{1}{(1+\alpha t^{\gamma})} \qquad f(t) = \frac{\alpha\gamma t^{-1}}{(1+\alpha t^{\gamma})^2} \qquad h(t) = \frac{\alpha\gamma t^{\gamma-1}}{(1+\alpha t^{\gamma})}$$



Las figuras anteriores muestran las formas de las funciones de supervivencia, densidad y riesgo con parámetros α=1 y γ=2.0 (rojo), γ =1.0 (azul) y γ=0.67 (verde).

La distribución log-logística puede ser útil para para describir un incremento inicial y después un decremento del riesgo.

2.6.4 Log-normal

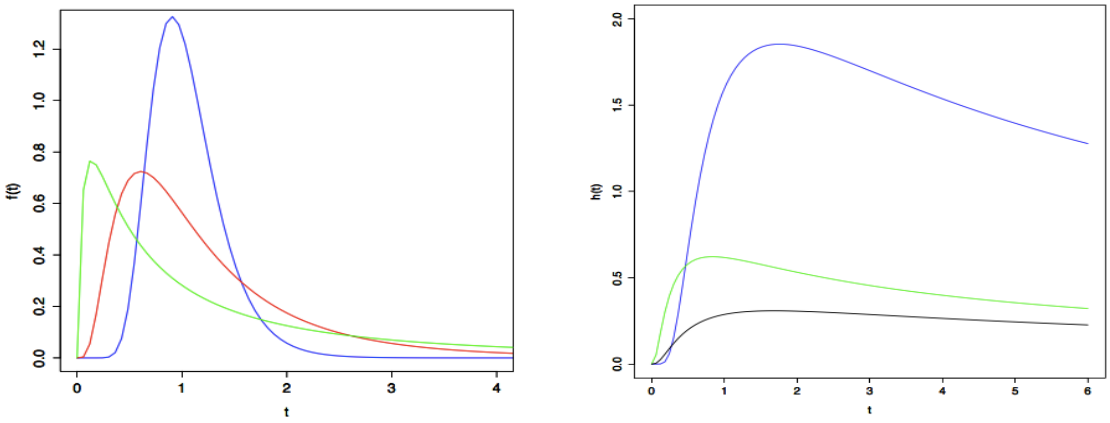
Si los logaritmos de las variables aleatorias tienen una distribución normal, éstas se distribuyen log-normal. Es decir, si el tiempo de supervivencia T es tal que $\log T$ se distribuye normal con media μ y varianza σ^2. La *función de riesgo* tiene la característica de crecer inicialmente hasta un máximo y después decrece (generalmente pasando la mediana). Ha sido ampliamente utilizada en el modelado de tiempos de reparación en sistemas de ingeniería.

Las funciones características de la distribución log-normal son:

$$S(t) = 1 - \Phi(\frac{logt - \mu}{\sigma}) \qquad f(t) = \frac{1}{\sqrt{2\pi}\sigma t} e^{\frac{-(logt-\mu)^2}{2\sigma^2}} \qquad h(t) = \frac{\frac{1}{\sigma t}\phi(\frac{logt-\mu}{\sigma})}{1 - \Phi(\frac{logt-\mu}{\sigma})}$$

dónde $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ denota la función de densidad de probabilidad y Φ(x) la función de distribución acumulativa de una normal estándar.

Recordemos que T no puede tomar valores de cero porque $logT$ no está definido cuando t=0.



En las figuras anteriores se muestra la *función de densidad* con parámetros μ = 0 y σ = 0.1(azul), σ = 0.5 (rojo) y σ= 2.0 (verde) y la *función de riesgo* con parámetros μ=*0,* σ = 0.5 (azul); μ=*0.3,* σ = 1 (verde) y μ=*1,* σ = 1 (negro).

2.6.5 Otros modelos paramétricos

Otros de los modelos que se pueden encontrar son el Gompertz, Pareto y el Gamma Generalizado cuyas funciones características de *riesgo, supervivencia* y *densidad* se describen en la Tabla 2.1.

Tabla 2.1: Otros modelos paramétricos de supervivencia

Distribución	Función de riesgo $h(t)$	Función de supervivencia $S(t)$	Función de densidad $f(t)$
Gompertz $\theta,\ \alpha > 0$ $t \geq 0$	$\theta e^{\alpha t}$	$\exp\left[\frac{\theta}{\alpha}\left(1 - e^{\alpha t}\right)\right]$	$\theta e^{\alpha t} \exp\left[\frac{\theta}{\alpha}\left(1 - e^{\alpha t}\right)\right]$
Pareto $\theta,\lambda > 0$ $t \geq \lambda$	$\frac{\theta}{t}$	$\frac{\lambda^{\theta}}{t^{\theta}}$	$\frac{\theta\lambda^{\theta}}{t^{\theta+1}}$
Gamma generalizada $\beta,\ \lambda > 0$ $\alpha > 0, t \geq 0$	$\frac{f(t)}{S(t)}$	$1 - I(\lambda t^{\alpha}, \beta)$	$\frac{\alpha\lambda^{\beta} t^{\alpha\beta-1} \exp(-\lambda t^{\alpha})}{\Gamma(\beta)}$

2.7 Procedimientos no-paramétricos

Una pregunta usual es conocer cuándo es conveniente usar métodos paramétricos y cuándo no-paramétricos. Una ventaja del uso de métodos no-paramétricos es su flexibilidad, una desventaja es que requieren de muestras mucho más grandes para obtener resultados razonables; otra desventaja del uso de modelos no-paramétricos es la dificultad de obtener una estimación de la función de riesgo la cual ofrece información relevante. En cambio los modelo paramétricos ofrecen formas cerradas de las funciones de *riesgo* y *supervivencia* a través de la estimación de pocos parámetros.

El proceso inicial en el análisis de supervivencia consiste en presentar resúmenes de los tiempos de supervivencia de los individuos en el grupo de análisis. Estos procedimientos pueden ser el resultado de la investigación o precursores de análisis más detallados.

La información de la supervivencia se resume principalmente a través de estimaciones de la *función de supervivencia* y *la función de riesgo.* Los métodos más usados para la estimación de estas funciones son denominados no-paramétricos ya que no requieren asumir una distribución concreta de los tiempos de supervivencia. Entre estos métodos se encuentran el estimador de *Kaplan-Meier* y el de *Nelson-Aalen.*

Cuando es de interés comparar los tiempos de supervivencia de grupos distintos de la población son utilizados también métodos no-paramétricos, los métodos de comparación más usados son el de *Wilcoxon* y el de *Log-Rangos.*

2.7.1 Estimador de Kaplan-Meier

También conocido como el estimador Producto-Límite, es un caso especial de la técnica de la tabla de vida, en el que una serie de intervalos de tiempo es formado de tal manera que sólo una falla ocurra en cada intervalo y ese fallo ocurre en el inicio de cada intervalo. Estima la probabilidad de sobrevivir más allá de determinado tiempo t, i.e. *S(t)*. El estimador es el producto de una serie de probabilidades condicionales estimadas. Por ejemplo, la probabilidad de sobrevivir k años es estimada como,

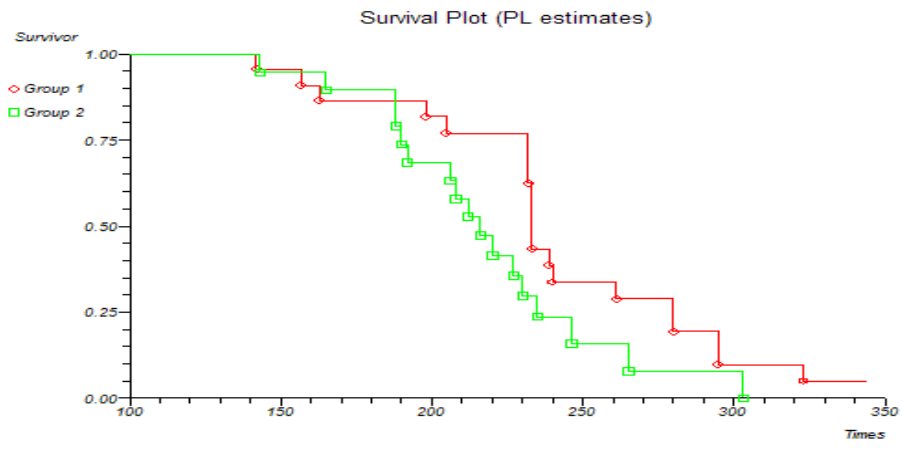
$$\hat{P}\left[T > k\right] = \hat{S}(k) = p_1 \cdot p_2 \cdot p_3 \cdots p_k,$$

Donde p_1 denota la proporción de individuos que sobrevivieron al menos un año, p_2 denota la proporción de individuos que sobrevivieron el segundo año dado que sobrevivieron el primer año, p_k denota la proporción de individuos que sobrevivieron al k-ésimo año habiendo sobrevivido *k-1* años. Sea t_i el i-ésimo tiempo de supervivencia, censurado o no censurado y δ_i una variable indicadora con valor 0 para observaciones censuradas y 1 para no censuradas. Ahora ordenamos los valores de t_i en orden ascendente de magnitud. Si r_i es la ordenación de t_i , el estimador de Kaplan-Meier puede ser escrito como

$$\hat{S}(t) = \prod_{t_i<t} \left(\frac{n - r_i}{n - r_i + 1} \right)^{\delta_i}, t \leq t_n$$

dónde n es el numero total de observaciones y t_n el tiempo de supervivencia más largo observado.

El estimador Producto Límite de Kaplan-Meier es útil para estimar distribución de supervivencia *S(t)*. Sin embargo, las estimaciones están limitadas al intervalo de tiempo en que las observaciones caen. Si la observación más larga es censurada, el estimador Producto Límite en ese tiempo es cero. Si la observación más larga es censurada, el estimador Producto Límite no puede ser nunca cero y no se encuentra definido más allá de la observación más grande. Adicionalmente si menos del 50 % de las observaciones son no-censuradas y la observación más grande es censurada, el tiempo medio de supervivencia no puede ser estimado. Por lo tanto, el método no es perfecto y existen razones suficientes para buscar un modelo paramétrico.



2.7.2 Comparación de dos funciones de supervivencia

En análisis de supervivencia es común la comparación de funciones entre distintos grupos de la población observada. Por ejemplo, en un estudio clínico de pacientes con diabetes puede ser de interés comparar la supervivencia de pacientes con hipertensión con la de pacientes con presión arterial normal para saber si la hipertensión está relacionada a la supervivencia de pacientes con diabetes.

Las diferencias de las funciones de supervivencia se pueden observar mediante gráficas de la función estimada, sin embargo, las gráficas no suelen asegurar si las diferencias son significativas o se deben a variaciones aleatorias. En estos casos es necesaria una prueba estadística.

Dos pruebas no paramétricas comúnmente utilizadas para la comparación de dos funciones de supervivencia son el test de *Wilcoxon* y el de *Log-Rangos.*

2.7.2.1 Prueba de Log-Rangos

La prueba de *Log-Rangos* require que los tiempos de supervivencia se consideren por separado para los dos grupos estudiados. Se considera que existen r tiempos de supervivencia distintos y ordenados de manera ascendente, $t_1<t_2<\ldots<t_r$ para los dos grupos, además en el tiempo t_j d_{1j} elementos del grupo I fallan y d_{2j} personas en el grupo II fallan para j=1,2,…,r.

A menos de que haya dos o más individuos con el mismo tiempo de supervivencia en el mismo grupo, los valores de d_{1j} y d_{2j} serán 0 ó 1. Suponemos que hay n_{1j} elementos en riesgo en el grupo I y n_{2j} en el grupo II justo antes del tiempo t_j. Por lo que en el tiempo t_j hay $d_j= d_{1j+} d_{2j}$ fallas de un total de $n_j=n_{1j} + n_{2j}$ en riesgo. Lo anterior se muestra en la Tabla 2.1

Tabla 2.1: Número de fallas en el tiempo tj en cada uno de los grupos

Grupo	Número de fallas al tiempo t_j	Número de supervivientes más allá de t_j	Número en riesgo antes de t_j
I	d_{1j}	n_{1j} - d_{1j}	n_{1j}
II	d_{2j}	n_{2j} - d_{2j}	n_{2j}
Total	d_j	n_j - d_j	n_j

La siguiente hipótesis nula a probar es la siguiente. H_0: No hay diferencia en las experiencias de supervivencia de ambos grupos.

Si además consideramos que las experiencias de supervivencia de los dos grupos son independientes.d_{1j} tiene una distribución hipergeométrica, la cual nos dice que la probabilidad de que la variable aleatoria del número de muertes en el primer grupo tome el valor d_{1j} es:

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j-d_j}{n_{1j}-d_{1j}}}{\binom{n_j}{n_{1j}}}$$

Por lo tanto el valor esperado de la variable aleatoria d_{1j} es

$$e_{1j} = \frac{n_{1j}d_j}{n_j}$$

Para observar una medida general de la desviación de d_{1j} de sus valores esperados se suman las diferencias d_{1j} - e_{1j} sobre el número total de tiempos de falla en ambos grupos, r. El estadístico resultante está dado por:

$$U_L = \sum_{j=1}^{r} d_{1j} - e_{1j}$$

Dado que los tiempos de falla son independientes, la varianza de U_L es la suma de varianzas de d_{1j} dadas por:

$$v_{ij} = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

Entonces la varianza de U_L está dada por:

$$V_L = \sum_{j=1}^{r} v_{1j}$$

Como la media de U_L entonces $U_L/\sqrt{V_L}$ se distribuye normal con media 0 y varianza 1, por lo tanto

$$W_L = \frac{U_L^2}{V_L} \sim \chi^2_{1g.l.}$$

El estadístico W_L resume la magnitud de las diferencias de los tiempos observados de supervivencia respecto a los valores esperados bajo la hipótesis nula de la no existencia de diferencias en los dos grupos.

2.7.2.2 Prueba de Wilcoxon

La prueba de *Wilcoxon*, es usada para probar la hipótesis nula de que no existen diferencias entre las funciones de supervivencia para dos grupos en un estudio de supervivencia. La prueba está basada en el estadístico

$$U_w = \sum_{j=1}^{r} n_j(d_{1j} - e_{1j})$$

donde d_{1j} es el número de fallas en el tiempo t_j en el primer grupo y e_{1j} el número esperado de fallecimientos en el grupo 1 en el tiempo t_j.

La varianza del estadístico de *Wilcoxon* es

$$V_w = \sum_{j=1}^{r} n_j^2 v_{1j}$$

donde

$$v_{1j} = \frac{n_{1j}n_{2j}(n_j - d_j)}{n_j^2(n_j - 1)}$$

Entonces el estadístico de prueba

$$W_w = \frac{U_w^2}{V_w}$$

Se distribuye ji-cuadrado con un grado de libertad cuando la hipótesis nula es cierta.

La prueba de *Log-Rangos* es apropiada para funciones de supervivencia cuyas funciones de riesgo son proporcionales a lo largo del tiempo, es decir, las dos curvas de supervivencia no se intersectan, para probar de otro tipo de hipótesis la prueba de *Wilcoxon* es más adecuada.

2.7.3 Comparación de tres o más grupos de supervivencia

Las pruebas de *Log-Rangos* y de *Wilcoxon* pueden ser extendidas a más de dos grupos de supervivencia. Definimos entonces formas análogas de los estadísticos U para comparar el número observado de fallas en los grupos 1,2,…g – 1 con sus valores esperados. En una extensión de la notación usada en los estadísticos previamente definidos obtenemos:

$$U_{Lk} = \sum_{j=1}^{r} \left(d_{kj} - \frac{n_{kj} d_j}{d_j} \right) \quad , \quad U_{Wk} = \sum_{j=1}^{r} n_j \left(d_{kj} - \frac{n_{kj} d_j}{d_j} \right)$$

Para k=1,2, . . . , g – 1. Estas cantidades son expresadas en forma de un vector con (g – 1) componentes, lo cual denotamos como U_L y U_W.

También necesitamos expresiones para las varianzas de U_{LK} y U_{WK} y para las covarianzas entre los pares de valores, en particular la covarianza entre U_{Lk} y $U_{Lk'}$ está dada por

$$V_{Lkk'} = \sum_{j=1}^{r} \left(\frac{n_{kj} d_j (n_j - d_j)}{n_j (n_j - 1)} \right) \left(\delta_{kk'} - \frac{n_{k'j}}{n_j} \right)$$

para k, $k' = 1, 2, \ldots, g - 1$, donde $\delta_{kk'}$ es tal que

$$\delta_{kk'} = \begin{cases} 1 \text{ si } k = k', \\ 0 \text{ e.o.c} \end{cases}$$

Estos términos son arreglados en la forma de una matriz de varianzas y covarianzas V_L, la cual es una matriz simétrica que tiene las varianzas de los estadísticos U_{Lk} en la diagonal. Por ejemplo en la comparación de tres grupos de supervivencia la matriz estaría dada por

$$V_L = \begin{pmatrix} V_{L11} & V_{L12} \\ V_{L21} & V_{L22} \end{pmatrix}$$

donde V_{L11} y V_{L22} son las varianzas de U_{L1} y U_{L1} respectivamente.

Similarmente, la matriz de varianzas y covarianzas para el estadístico de *Wilcoxon* es V_W, cuyo *(k, k')*-ésimo elemento es:

$$V_{Wkk'} = \sum_{j=1}^{r} n_j^2 \frac{n_{kj} d_j (n_j - d_j)}{n_j (n_j - 1)} \left(\delta_{kk'} - \frac{n_{k'j}}{n_j} \right)$$

para k, $k' = 1, 2, \ldots, g - 1$.

Finalmente, para probar la hipótesis nula de no diferencias entre los grupos de supervivencia, hacemos uso del resultado que nos dice que los estadísticos $U'_L V^{-1}_L U_L$ y $U'_W V^{-1}_W U_W$, tienen una distribución ji-cuadrada con (g – 1) g.l. cuando la hipótesis nula es verdadera.

2.8 Modelo de Riesgos Proporcionales

En el análisis de supervivencia es de interés conocer si ciertas características de un individuo en estudio están relacionadas a la ocurrencia de dicho evento. Por ejemplo, para eventos relacionados con la salud, si un paciente es fumador, sus niveles de colesterol, historial familiar de enfermedades del corazón son referidos como factores de riesgo o covariables.

El modelo de riesgos proporcionales también conocido como Modelo de Regresión de Cox ha sido el método más usado en análisis de supervivencia para conocer la relación de cada variable con el evento de interés, sin importar si el tiempo de supervivencia es discreto o continuo o si existe o no censura.

Dado un conjunto de p covariables $x_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$ la función de riesgo para un individuo es modelada por

$$h_i(t, x_i) = h_0(t) exp(b^T x_i)$$

Esto es, la función de riesgo de un individuo es el producto de una función de riesgo inicial $h_0(t)$ arbitraria y una función lineal exponencial de las p covariables. La función lineal $b^T x_i = b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi}$ es exponenciada para asegurar que la función de riesgo sea no-negativa. Dado que la distribución de los tiempos de supervivencia no es especificada el modelo de Cox es un ejemplo de un modelo semiparamétrico.

A pesar de que el modelo de Cox es introducido en el marco de referencia de *riesgos proporcionales*, el modelo puede ser extendido en casos de funciones de riesgo no-proporcionales.

Adicionalmente de ser un modelo de regresión el modelo de Cox también puede ser aplicado a problemas de dos muestras con y sin covariables. Consideremos el caso de comparar la supervivencia de dos grupos. Sea $p=1$ y x_1 el indicador que toma valor 0 si el individuo pertenece al grupo 1 y 1 si pertenece al grupo 2. Asumir la hipótesis de *riesgos proporcionales* implica que la función de riesgo para el grupo 2 es proporcional a la del grupo 1. Esto es escrito como $h_2(t) = e^b h_1(t)$. Si $e^b > 1$, la supervivencia en el grupo 1 es superior con respecto al grupo 2.

2.8.1 Inclusión de variables y factores en el modelo

Existen dos tipos de variables de las cuales una función de riesgo puede depender: *variables* y *factores*. Las *variables* pueden tomar valores numéricos y que frecuentemente están en una escala continua, tales como la edad. Un *factor* es una variable que puede tomar un limitado conjunto de valores, conocidos como niveles del *factor*.

Las *variables* pueden ser fácilmente incorporadas en un modelo de riesgos proporcionales. Cada variable aparece en el modelo con un coeficiente b. Consideremos como ejemplo una situación en que la función de riesgo depende de dos variables X_1 y X_2. Para el i-ésimo individuo estas variables serían x_{1i} y x_{2i} y el modelo de riesgos proporcionales estaría expresado como:

$$h_i(t) = exp(\beta_1 x_{1i} + \beta_2 x_{2i}) h_0(t)$$

En este tipo de modelos la función de riesgo inicial $h_0(t)$ es la función de riesgo para un individuo en el que todas las variables incluidas tomen el valor de cero.

En el caso de que la función de riesgo dependa de un solo *factor, A,* con a niveles. El modelo para un individuo para cuyo nivel de A es j ,entonces necesitaríamos incorporar el término α_j el cual representa el efecto debido al j-ésimo nivel del *factor*. Los términos $\alpha_1, \alpha_2, \ldots, \alpha_a$ son conocidos como los efectos principales del *factor A.* De acuerdo al modelo de riesgos proporcionales, la función de riesgo para un individuo con factor A en el nivel j es $exp(\alpha_j)\ h_0(t)$. La función de riesgo inicial $h_0(t)$ ha sido definida como el riesgo para un individuo con valores de todas las covariables explicativas como cero; entonces uno de los niveles α_j debe ser tomado igual a cero. Una posibilidad es tomar la restricción $\alpha_1 = 0$ que corresponde a tomar la función de riesgo inicial para quien A está en el primer nivel.

Los modelos que contienen términos que corresponden factores pueden ser expresados como combinaciones lineares de variables explicativas definiendo *variables indicadoras* para cada factor.

Si se toma la restricción $\alpha_1 = 0$, el término α_j puede ser incluido en el modelo definiendo α-*1* variables indicadoras, $X_2, X_3, \ldots, X_a$ que tomen los valores mostrados la siguiente tabla.

Nivel de A	X_2	X_3	...	X_a
1	0	0	...	0
2	1	0	...	0
3	0	1	...	0
⋮	⋮	⋮	⋮	⋮
a	0	0	...	1

El término α_j puede ser incorporado en la parte lineal del modelo de riesgos proporcionales incluyendo las α-*1* variables explicativas $X_2, X_3, \ldots, X_a$ con coeficientes $\alpha_2, \alpha_3, \ldots, \alpha_a$. En otras palabras el término α_j es reemplazado por $\alpha_2 x_2 + \alpha_3 x_3 + \ldots + \alpha_a x_a$, donde x_j es el valor de X_j para un individuo en el que A está en el nivel j, $j = 1,2, \ldots, a$. Entonces hay α-*1* parámetros asociados con el efecto del *factor A.*

2.8.2 Estimación del modelo de riesgos proporcionales

Ajustar el modelo de riesgos proporcionales a un conjunto observado de tiempos de supervivencia implica estimar los coeficientes desconocidos de las covariables explicativas X_1, X_2, . . . , X_p en el componente lineal del modelo b_1, b_2, . . . , b_p. La función de riesgo inicial $h_0(t)$ también necesita ser estimada, aunque estos dos componentes del modelo pueden ser estimados por separado. Primero son estimadas las *b's* y estos estimados son usados para construir un estimador de la función de riesgo inicial.

Los coeficientes b en el modelo de riesgos proporcionales pueden ser estimados a través del método de *máxima verosimilitud*. Para usar este método primero es necesario obtener la verosimilitud de de los datos muestrales. Esto es la verosimilitud conjunta de los datos observados como una función de los parámetros desconocidos en el modelo; para el caso del modelo de riesgos proporcionales esta es una función de los tiempos esperados de supervivencia y de los parámetros b desconocidos en la componente lineal del modelo. Las estimaciones de *b's* son entonces aquellos valores que son más verosímiles con base en los datos observados.

Suponemos que existe información disponible de n individuos, de los cuales hay r distintos tiempos de falla, entonces existen $n - r$. Asumiremos que sólo un individuo observa la falla en cada tiempo, así que no hay empates en los tiempos de falla. Los tiempos de falla ordenados serán denotados por $t_1 < t_2 <. . . < t_r$, así que t_j es el j-ésimo tiempo de falla ordenado. El conjunto de individuos que están en riesgo al tiempo t_j será denotado por $R(t_j)$, el cual representa el grupo de individuos que aún no experimentan la falla y no están censurados justo antes de t_j; esta cantidad es llamada el *conjunto en riesgo*.

Cox (1972) demostró que la función de verosimilitud para el modelo de riesgos proporcionales está dada por:

$$L(b) = \prod_{j=1}^{r} \frac{exp(b^T x_j)}{\sum_{l \in R(t_j)} exp(b^T x_l)}$$

donde x_j es el vector de covariables para el individuo que falla en el tiempo t_j. La suma en el denominador de esta función de verosimilitud es la suma de los valores de $exp(b^T x)$ sobre todos los individuos que están en riesgo en el tiempo t_j.

Ahora supongamos que existen n tiempos de falla observados denotados por t_1, t_2, ... , t_n, y que δ_i es un indicador del evento, el cual toma valor de cero si el tiempo t_j, es censurado por la derecha y 1 en otro caso. La función de verosimilitud anterior puede ser expresada como:

$$\left\{ \prod_{j=1}^{r} \frac{exp(b^T x_j)}{\sum_{l \in R(t_j)} exp(b^T x_l)} \right\}^{\delta_i}$$

La correspondiente función de log-verosimilitud está dada por:

$$logL(b) = \sum_{j=1}^{n} \delta_i \left\{ b^T x_j - log \sum_{l \in R(t_j)} exp(b^T x_l) \right\}$$

Los estimadores máximo verosímiles de los parámetros b pueden ser encontrados al maximizar la función de log-verosimilitud anterior. El proceso de maximización generalmente es logrado usando el método iterativo de *Newton-Raphson*.

Cuando es usado un paquete estadístico para ajustar un modelo de riesgos proporcionales, los parámetros estimados son acompañados por sus errores estándar, los cuales pueden ser usados para obtener intervalos de confianza de los parámetros b. Si el intervalo de confianza construido para b no incluye al cero, es una evidencia de que el valor del parámetro estimado es distinto de cero. Más específicamente para evaluar la hipótesis nula $b=0$ se utiliza el estadístico $\hat{b}/se(\hat{b})$; el resultado de este estadístico es evaluado en una distribución normal estándar para obtener su *p-valor* correspondiente.

Al interpretar el *p-valor* de un parámetro b_j , es necesario reconocer que la hipótesis nula que está siendo evaluada es $b_j=0$ en presencia de los otros términos en el modelo, en general, los estimadores individuales de los parámetros b en un modelo de riesgos proporcionales no son independientes uno del otro; por lo tanto, los resultados de evaluar diferentes hipótesis sobre los parámetros b por separado pueden ser difícil de interpretar.

Existen métodos implementados en algunos paquetes estadísticos, usados para seleccionar de forma automática las variables explicativas que deberían ser incluidas en un modelo de riesgos proporcionales. Estos métodos son conocidos como *paso a paso* ya que en cada paso una variable es agregada o eliminada del conjunto de variables explicativas basándose en un criterio de selección previamente especificado.

2.8.3 Validación del supuesto de riesgos proporcionales

Entre los métodos más usados para verificar el supuesto del modelo se encuentran los gráficos y las pruebas de residuales.

Los métodos gráficos consisten en observar la estimación de la función de supervivencia para distintos grupos; si se cruzan las curvas, entonces el supuesto no se cumple.

Los residuales son valores que pueden ser calculados para cada individuo dentro del estudio, y tienen la característica de que su comportamiento es conocido, al menos aproximadamente cuando el modelo ajustado es satisfactorio. La justificación teórica rigurosa está fuera de los objetivos de esta tesis. A continuación se enlistan los más comunes:

- Residuales de Cox-Snell
- Residuales de Devianza
- Residuales de Schonfeld

Todos estos métodos se encuentran implementados en la mayoría de los paquetes estadísticos.

2.8.4 Comparación de modelos alternativos

En la selección de un modelo ajustado a ciertos tiempos de supervivencia, se busca encontrar la dependencia de la función de riesgos a una o más variables explicativas, durante este proceso son ajustados y comparados diferentes modelos de riesgos proporcionales con componentes lineales que contienen diferentes conjuntos de términos.

Suponemos que son contemplados dos modelos para un conjunto particular de datos, el Modelo(1) y el Modelo(2), donde el Modelo(1) contiene un subconjunto de los términos en el Modelo(2). Se dice entonces que el Modelo(1) está anidado paramétricamente en el Modelo(2). Específicamente supongamos que son incluidas p en el Modelo(1), entonces la función de riesgo bajo este modelo puede ser expresada como

$$exp\left\{b_1x_1 + b_2x_2 + ... + b_px_p\right\} h_0(t).$$

También suponemos que $p + q$ variables explicativas son incluidas en el Modelo(2), entonces la función de riesgo bajo este modelo es

$$exp\left\{b_1x_1 + ... + b_px_p + b_{p+1}x_{p+1} + ... + b_{p+q}x_{p+q}\right\} h_0(t).$$

El Modelo(2) contiene q variables explicativas adicionales X_{p+1}, X_{p+2}, . . . , X_{p+q}. El problema estadístico se encuentra en determinar si las q variables adicionales mejoran significativamente el modelo, de lo contrario deberían ser omitidas y el Modelo(1) sería más adecuado.

Previamente observamos que cuando hay un número de variables explicativas de posible relevancia, el efecto de cada una no puede ser estudiado independiente de las otras; esto es, el efecto de cada término incluido en el modelo depende de los otros términos que ya estén en el modelo. Por ejemplo, en el Modelo(1) el efecto de cualquiera de las p variables explicativas depende de las $p-1$ variables que ya han sido ajustadas.

Con el objeto de comparar modelos alternativos a un conjunto de tiempos de falla, es necesario un estadístico que mida el grado de ajuste de cada modelo. Dado que la función de verosimilitud resume la información que contienen los datos sobre los parámetros desconocidos en un modelo, un estadístico de resumen adecuado es el valor de la función de verosimilitud cuando los parámetros son reemplazados por sus estimadores máximo-verosímiles, la cual denotaremos como $\hat{L}$. Para un conjunto de datos dado, mientras más grande es el valor de la función maximizada, mejor es el ajuste entre el modelo y los datos observados.

La devianza es un estadístico utilizado para conocer que tan cercanos son los valores predichos por el modelo ajustado a los datos originales. Consideramos el caso extremo de ajustar un modelo *saturado* donde el número de parámetros iguala al número de observaciones, en este caso el modelo *saturado* no provee simplificación alguna, sin embargo, es posible comparar cualquier modelo propuesto con el modelo *saturado* para determinar que tan bien se ajusta a la información el modelo propuesto. La devianza D está definida como:

$$D_p = 2\ [\ log\ \hat{L}(s)\ -\ log\ \hat{L}(p)]$$

Donde $log\ \hat{L}(s)$ representa la función de log-verosimilitud del modelo saturado y $log\ \hat{L}(p)$ la del modelo propuesto.

Si consideramos nuevamente el Modelo(1) y el Modelo(2) y sea el valor de la función de verosimilitud maximizada para cada modelo $\hat{L}(1)$ y $\hat{L}(2)$ respectivamente. Los dos modelos pueden ser comparados con base en las diferencias de sus devianzas; una gran diferencia entre sus devianzas llevaría a la concluir que las q variables adicionales en el Modelo(2) mejoran la adecuación del modelo. Para tamaños de muestras suficientemente grandes la diferencia entre las devianzas de los modelos tiene una distribución ji-cuadrada con grados de libertad igual a la diferencia entre el número de parámetros b ajustados en cada modelo. En el caso del Modelo(1) y el Modelo(2) tenemos que el estadístico $X^2 = D_1 - D_2 = 2\ [\ log\ \hat{L}(2) - log\ \hat{L}(1)]$ tiene una distribución ji-cuadrada con q grados de libertad bajo la hipótesis nula que los coeficientes b_{p+1}, $b_{p+2}, \ldots, b_{p+q}$ son cero.

2.9 Modelo de vida acelerada

A pesar de que el modelo de riesgos proporcionales ha sido ampliamente utilizado en el análisis de tiempos de supervivencia, existen relativamente pocas distribuciones para los tiempos de supervivencia que pueden ser usadas con este modelo. Un modelo que incluye un rango de distribuciones más amplio es el *modelo de vida acelerada*. En circunstancias donde el supuesto de riesgos proporcionales no se cumple, modelos basados en esta familia pueden resultar buenos ajustes.

El modelo de vida acelerada es un modelo general para análisis de supervivencia, en el que las variables explicativas medidas en un individuo se asume que actúan de manera multiplicativa en escala de tiempo.

Para ilustrar el modelo de vida acelerada suponemos que existen dos tratamientos para pacientes denotados por *S*, el tratamiento estándar y *N,* el nuevo tratamiento. Bajo un *modelo de vida acelerada,* el tiempo de supervivencia de un individuo bajo el nuevo tratamiento es un múltiplo de del tiempo de supervivencia para un individuo bajo el tratamiento estándar. Por lo tanto, el efecto del nuevo tratamiento se dice que *acelera* o *desacelera* el paso del tiempo. Bajo este supuesto, la probabilidad de que un individuo bajo el nuevo tratamiento sobreviva más allá del tiempo t es la probabilidad de que un individuo bajo el tratamiento estándar sobreviva más allá del tiempo t/ϕ.

Sean $S_S(t)$ y $S_N(t)$ las funciones de supervivencia para individuos en los dos grupos de tratamiento. Entonces, el modelo de vida acelerada especifica que

$$S_N(t) = S_S(t/\phi)$$

para cualquier valor del tiempo de supervivencia t. Una interpretación de este modelo es que el tiempo de vida de un individuo en el nuevo tratamiento es ϕ veces el tiempo el tiempo de vida que hubiera experimentado bajo el tratamiento estándar. El parámetro ϕ entonces, refleja el impacto del nuevo tratamiento en la escala de tiempo de referencia. Cuando el objeto de estudio es el tiempo de muerte de un paciente valores de ϕ menores a 1 corresponden a una aceleración del tiempo de muerte para un individuo asignado a un nuevo tratamiento, relativo a un individuo que se encuentra en el tratamiento estándar, por lo cual, el tratamiento estándar sería más efectivo. Por otra parte, si el objeto de estudio es el tiempo de recuperación de un paciente, de ϕ menores a 1 se encontrarían cuando el efecto del nuevo tratamiento. Por lo tanto el valor ϕ^{-1} es definido como el *factor de aceleración.*

De la relación entre la función de supervivencia, la función de densidad y la función de riesgo previamente establecidas, las relaciones entre la función de densidad y la función de riesgo para los individuos entre los dos grupos de tratamiento son:

$$f_N(t) = \phi^{-1} f_S(t/\phi),$$

y

$$h_N(t) = \phi^{-1} h_S(t/\phi).$$

Sea X una variable indicadora que toma valor cero cuando un individuo recibe el tratamiento estándar y la unidad cuando recibe un nuevo tratamiento. La función de riesgo puede entonces ser expresada como:

$$h_i(t) = \phi^{-x_i} h_0(t/\phi^{x_i}),$$

Cuando $x_i = 0$ en esta expresión se muestra que $h_0(t)$ es la función de riesgo para un individuo bajo el tratamiento estándar.

El parámetro ϕ debe ser no-negativo, por lo tanto es conveniente definir $\phi = e^{\alpha}$. El modelo de vida acelerada entonces es

$$h_i(t) = e^{-\alpha x_i} h_0(t/e^{\alpha x_i})$$

El modelo de vida acelerada de la ecuación anterior puede ser generalizado cuando el valor de p variables explicativas han sido reconocidas para los individuos en el estudio. Entonces la función de riesgo para el i-ésimo individuo al tiempo t se puede escribir como:

$$h_i(t) = e^{-\eta_i} h_0(t/e^{\eta i}),$$

donde

$$\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_p x_{pi}$$

es la componente lineal del modelo, en el cual x_{ji} es el valor de la j-ésima variable explicativa, X_j, $j=1,2, \ldots ,p,$ para el i-ésimo individuo, $i=1,2, \ldots , n.$ De igual forma que en el modelo de riesgos proporcionales, la función de riesgo inicial $h_0(t)$ representa el riesgo al tiempo t para un individuo para el que los valores de las p variables explicativas son iguales a cero. La función de supervivencia correspondiente para el i-ésimo individuo es

$$S_i(t) = S_0 \{t/exp(\eta_i)\}$$

Donde $S_0(t)$ es la función inicial de supervivencia.

2.9.1 Representación log-lineal del modelo de vida acelerada

Los modelos de vida acelerada paramétricos para tiempos de supervivencia son unificados adoptando una representación log-lineal del modelo de general.

Si consideramos un modelo log-lineal para la variable aleatoria T asociada al tiempo de vida del individuo i en un estudio de supervivencia, entonces

$$logT_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_p x_{pi} + \sigma\epsilon_i$$

En este modelo $\alpha_1, \alpha_2, ..., \alpha_p$ representan los coeficientes de p variables explicativas X_1, X_2, ..., X_p, mientras que μ y σ son conocidos como parámetros de intercepción y escala, respectivamente. La cantidad ϵ_i es una variable aleatoria usada para modelar la desviación de los valores de $logT_i$ de la parte lineal del modelo, además se asume que sigue una distribución de probabilidad particular. En esta representación del modelo los parámetros α reflejan el efecto que cada variable explicativa tiene en el tiempo de supervivencia; valores positivos sugieren que los tiempos de supervivencia incrementan conforme incrementan los valores de las variables explicativas.

Para mostrar la relación que existe entre la representación log-lineal del modelo y el modelo general consideremos la función de supervivencia de la variable aleatoria T_i , la variable aleatoria asociada al tiempo de supervivencia el i-ésimo individuo

$$S_i(t) = P(T_i \geq t) = P\left\{exp(\mu + \alpha' x_i + \sigma\epsilon_i) \geq t\right\}$$

Ahora, $S_i(t)$ puede ser escrita de la forma

$$S_i(t) = P\left\{exp(\mu + \sigma\epsilon_i) \geq t/exp(\alpha' x_i)\right\}$$

Y la función base de supervivencia en donde $x = 0$, es

$$S_0(t) = P\left\{exp(\mu + \sigma\epsilon_i) \geq t\right\}$$

Se sigue entonces que,

$$S_i(t) = S_0\left\{t/exp(\alpha' x_i)\right\}$$

Que es la forma general del modelo de vida acelerada. De las relaciones entre las funciones características de supervivencia se puede deducir la función de riesgo es equivalente también a la del modelo original.

El modelo log-lineal también puede ser usado para dar una forma general de la función de supervivencia para el i-ésimo individuo que es

$$S_i(t) = P\left(T_i \geq t\right) = P\left(logT_i \geq logt\right)$$

Por lo tanto

$$S_i(t) = P\left(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_p x_{pi} + \sigma\epsilon_i \geq logt\right),$$

$$= P\left(\epsilon_i \geq \frac{logt - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - ... - \alpha_p x_{pi}}{\sigma}\right)$$

Entonces la función de supervivencia para el i-ésimo individuo puede ser expresada como:

$$S_i(t) = S_{\epsilon_i}\left(\frac{logt - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - ... - \alpha_p x_{pi}}{\sigma}\right)$$

Este resultado muestra que la función de supervivencia de T_i puede ser encontrada a partir de la función de supervivencia de la distribución de ϵ_i.También demuestra que un modelo de vida acelerada puede ser derivado de muchas distribuciones de probabilidad de ϵ_i.

2.9.2 Modelos paramétricos de vida acelerada

Las distribuciones de ϵ_i que son más comúnmente utilizadas en los modelos de vida acelerada son aquellas en las cuales sus percentiles tienen formas más simples. Decisiones particulares para las distribuciones de ϵ_i en la formulación log-lineal del modelo de vida acelerada nos llevan a distribuciones asociadas al tiempo de supervivencia del individuo i:

- Si se asume una distribución normal de ϵ_i el modelo resultante es el log-normal.
- Si se asume una distribución de valores extremos de ϵ_i el modelo resultante es el Weibull.
- Si se asume una distribución logística de ϵ_i el modelo resultante es el log-logístico.

La función de riesgo acumulado de ϵ_i es encontrada a partir de $H_{\epsilon_i}(\epsilon) = -logS_{\epsilon_i}(\epsilon)$ y si es deseado conocerla, la función de densidad está dada por $f_{\epsilon_i}(\epsilon) = h_{\epsilon_i}(\epsilon)S_{\epsilon_i}(\epsilon)$.La función de riesgo y supervivencia de las distribuciones de ϵ_i que llevan a modelos comúnmente utilizados en modelos de vida acelerada para tiempos de supervivencia se encuentra resumida en la Tabla 2.3:

Tabla 2.3: Resumen de modelos paramétricos de vida acelerada

Distribucion de T_i	$S_{\epsilon_i}(\epsilon)$	$h_{\epsilon_i}(\epsilon)$
Weibull	$\exp(-e^{\epsilon})$	e^{ϵ}
Log-logistica	$(1+e^{\epsilon})^{-1}$	$(1+e^{\epsilon})^{-1}$
Log-normal	$1-\Phi(\epsilon)$	$\exp(-\epsilon^2/2)/\{1-\Phi(\epsilon)\}\sqrt{(2\pi)}$

A partir de la función de supervivencia y de riesgo de ϵ_i la función de supervivencia y riesgo de T_i pueden ser encontradas dada las siguientes relaciones

$$S_i(t) = S_{\epsilon_i}(\frac{logt - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - ... - \alpha_p x_{pi}}{\sigma}),$$

$$h_i(t) = \frac{1}{\sigma t} h_{\epsilon_i}(\frac{logt - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - ... - \alpha_p x_{pi}}{\sigma}),$$

La representación log-lineal del modelo de Weibull y el modelo log-logístico llevan a parametrizaciones distintas de las distribuciones de los tiempos de supervivencia previamente vistas.

2.9.3 Ajuste y comparación de modelos de vida acelerada

Los modelos de vida acelerada son ajustados usando el método de máxima verosimilitud. La función de verosimilitud es obtenida de la representación log-lineal del modelo, con la que después de utilizar métodos iterativos se obtienen los mejores estimadores. La función de verosimilitud de los n tiempos de supervivencia observados, $t_1, t_2, \ldots, t_n$, está dada por:

$$L(\alpha, \mu, \sigma) = \prod_{i=1}^{n} \{f_i(t_i)\}^{\delta_i} \{S_i(t_i)\}^{1-\delta_i}$$

donde $f_i(t_i)$ y $S_i(t_i)$ son la función de densidad y de supervivencia para el i-ésimo individuo al tiempo t_i y $\boldsymbol{\delta}_i$ es el evento indicador para la i-ésima observación, así que $\boldsymbol{\delta}_i$ es la unidad si la i-ésima observación es un evento y cero si es censurado. Ahora de la ecuación:

$$S_i(t_i) = S_{\epsilon_i}(z_i)$$

donde

$$z_i = (logt_i - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - ... - \alpha_p x_{pi})/\sigma,$$

y diferenciando con respecto a t obtenemos

$$f_i t_i = \frac{1}{\sigma t_i} f_{\epsilon_i}(z_i).$$

La función de verosimilitud puede ser expresada en términos de la función de supervivencia y densidad de ϵ_i, dando

$$L(\alpha, \mu, \sigma) = \prod_{i=1}^{n} (\sigma t_i)^{-\delta_i} \left\{ f_{\epsilon_i}(z_i) \right\}^{\delta_i} \left\{ S_{\epsilon_i}(z_i) \right\}^{1-\delta_i}.$$

La función de log-verosimilitud es entonces

$$logL(\alpha, \mu, \sigma) = \sum_{i=1}^{n} \left\{ -\delta_i log(\sigma t_i) + \delta_i log f_{\epsilon_i}(z_i) + (1 - \delta_i) log S_{\epsilon_i}(z_i) \right\}$$

y los estimadores máximo-verosímiles de los *p+2* parámetros desconocidos son encontrados al maximizar esta función utilizando el procedimiento de Newton Raphson. Después de ajustar un modelo, el valor del estadístico – *2 log* $\hat{L}$ puede ser calculado y usado al hacer comparaciones entre modelos anidados, de la misma forma que en el modelo de riesgos proporcionales, para comparar dos modelos, la diferencia en los valores del estadístico – *2 log* $\hat{L}$ para los dos modelos es comparada con puntos porcentuales de una distribución ji-cuadrada con grados de libertad iguales a la diferencia en el número de parámetros α incluidos en la componente lineal del modelo.

Una vez que un buen modelo ha sido identificado, estimaciones de la función de riesgo y de supervivencia pueden ser obtenidos y graficados. El modelo ajustado puede ser interpretado en términos de los valores estimados de los factores de aceleración para individuos particulares, o, en términos de la mediana y otros percentiles de la distribución de los tiempos de supervivencia.

En particular el p-ésimo percentil de la distribución de los tiempos de supervivencia, para un individuo cuyo vector de variables explicativas es x_i, está dado por:

$$\hat{t}_i(p) = exp\left\{ \hat{\sigma} \epsilon_i(p) + \hat{\mu} + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i} + ... + \hat{\alpha}_p x_{pi} \right\}.$$

Capítulo III

Ajuste y análisis de los modelos de supervivencia

3.1 Introducción

En esta sección se describirá el procedimiento seguido para analizar la supervivencia de los establecimientos de la ZMVM. Primero se realizará un análisis exploratorio de la base de datos generada a partir de los registros del DENUE y las probabilidades de cierre asignadas por estrato de personal ocupado y sector económico.

Después se aplicarán métodos no-paramétricos (Kaplan Meier) para describir la supervivencia entre distintos grupos; podremos concluir si la supervivencia de los establecimientos es más alta en alguno de las dos entidades o si existen diferencias significativas entre los distintos grupos de personal ocupado y sector económico.

Posteriormente se ajustará un modelo paramétrico de vida acelerada seleccionando una distribución conocida que se ajuste a nuestros datos, de las covariables disponibles en la base de datos previamente descrita se seleccionarán las que tengan un efecto significativo y se incluirán como factores de riesgo en el modelo.

Con los resultados obtenidos se compararán los distintos métodos utilizados para evaluar la supervivencia de los establecimientos.

Por último se realizará un seguimiento a la supervivencia de los establecimientos a través del tiempo, esto con el fin de observar si existen zonas geográficas con un crecimiento más acelerado del cierre de establecimientos.

3.2 Análisis Exploratorio

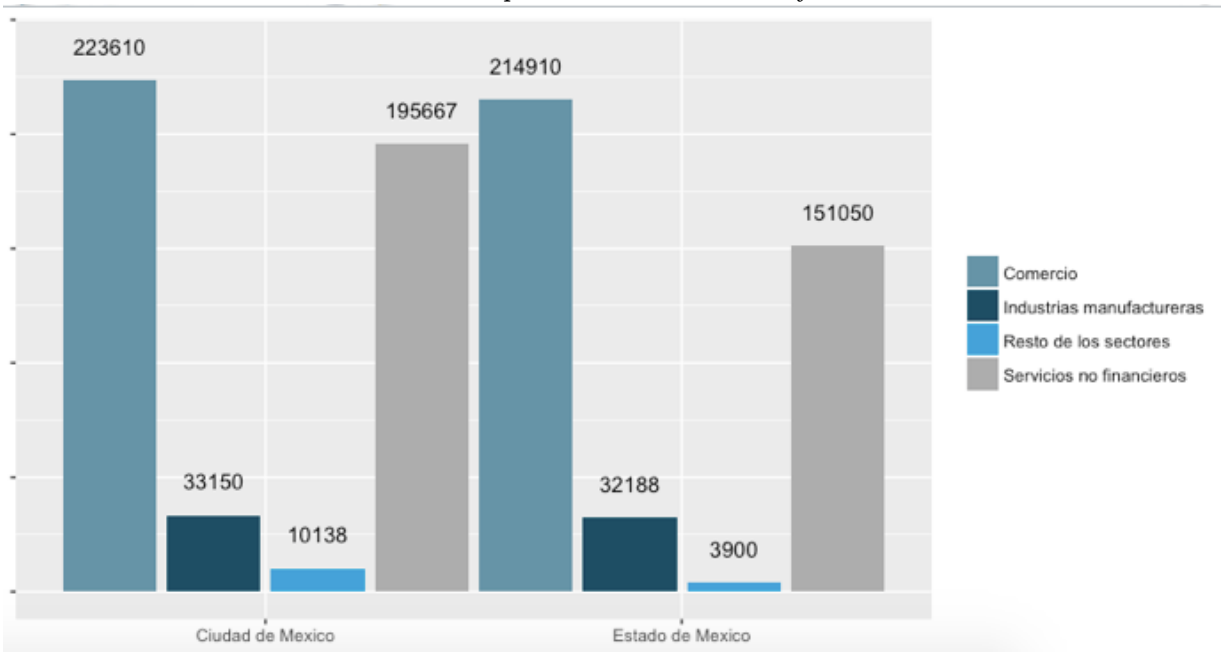
La base de datos generada contiene 864,613 observaciones las cuales corresponden a los establecimientos de la ZMVM registrados en la base del DENUE, la descripción de las variables disponibles se muestra en la Tabla 3.1:

Tabla 3.1: Variables utilizadas para la caracterización de los negocios y su supervivencia

Variables	Descripción	Valores
edad	Tiempo de supervivencia	años
estr_poc	Estrato de personal ocupado	6 niveles: 0 a 5, 6 a 10, 11 a 30, 30 a 50, 51 a 100, 101 a 250
sec_eco	Sector económico	Comercio, Manufacturas, Servicios no financieros, Resto
cve_ent	Clave de Entidad Federativa	9 = Ciudad de México 15 = Estado de México
cve_mun	Clave de Municipio	
tipo_est	Tipo de Establecimiento	0=fijo, 1=semifijo
remun_prom	Remuneración anual promedio municipal	Miles de pesos
pib_per_cap	PIB per cápita municipal	Miles de pesos
im_t	Índice de marginación municipal	[0,2.1610]
gm	Grado de marginación municipal	3 niveles: muy bajo, bajo, medio
indicador	Indicador de falla	0 = censura, 1 = cierre

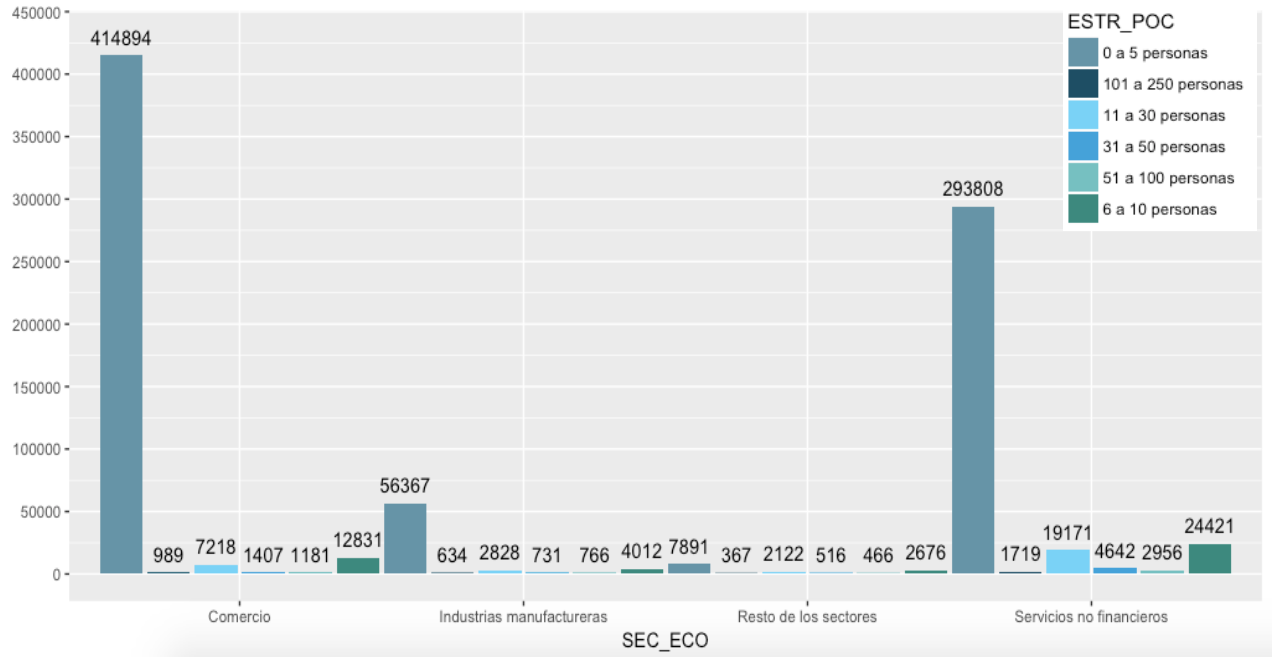
Podemos observar que la distribución de los negocios por sector económico en la ZMVM es similar a la distribución nacional siendo el sector comercio el predominante, seguido de cerca por los servicios no financieros. (ver gráfica 3.1).

Gráfica 3.1: Número de establecimientos por Entidad Federativa y Sector Económico



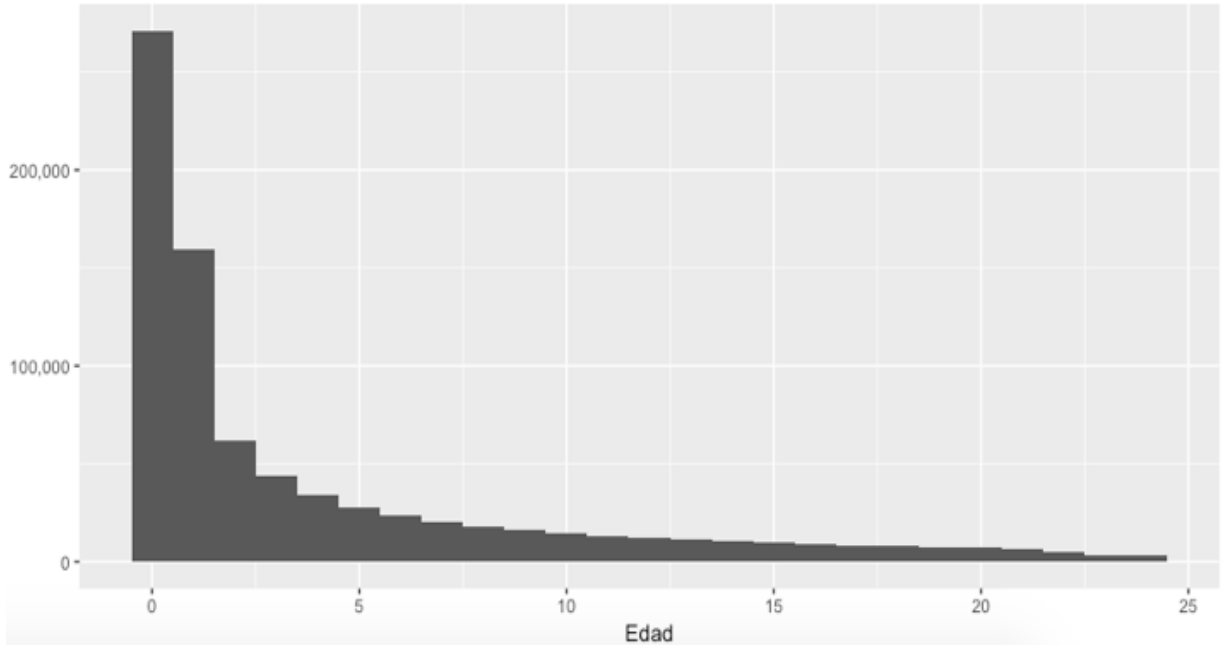
Los negocios de mayor tamaño por personal ocupado se encuentran en el sector de los servicios no financieros como se observa en la Gráfica 3.2:

Gráfica 3.2: Número de establecimientos por Sector Económico y personal ocupado



Mediante el histograma de la Gráfica 3.3 podemos tener una idea más cercana de la forma de la distribución de los tiempos de falla; podemos observar que existe una gran acumulación de eventos en los dos primeros años.

Gráfica 3.3: Histograma de frecuencias de los tiempos de falla



Para conocer la relación que existe entre el evento de cierre y las variables numéricas podemos apoyarnos en la siguiente matriz de correlaciones de la gráfica 3.4:

Gráfica 3.4: Nivel de correlación entre variables municipales y cierre de establecimientos

Podemos notar que existe una fuerte correlación entre las variables a nivel municipal, siendo claro que a mayor ingreso (Remuneración promedio y PIB per cápita), menor es el nivel de marginación. Éstas a su vez tienen un efecto similar, aunque de una forma mucho menos notoria, en el cierre de los establecimientos; es decir, a mayor nivel de ingreso es menos frecuente el cierre de un establecimiento y a mayor nivel de marginación más ocurre el cierre de un negocio.

Dado que las categorías de clasificación de los negocios en nuestra base de datos se pueden definir como *exhaustivas* y *mutuamente excluyentes*, además de ser una *muestra* de una población total. El siguiente paso del análisis exploratorio consiste en realizar pruebas de independencia mediante el uso de tablas de contingencia.

En la Tabla 3.2 se evalúa la independencia entre el nivel de marginación municipal y el cierre de un negocio, con el objetivo de evitar valores observados distintos de cero se juntaron en una sola categoría los negocios clasificados en el nivel de marginación medio y bajo.

Tabla 3.2: Tabla de contingencia de marginación municipal vs evento de cierre

Evento/Marginación	Medio/Bajo	Muy bajo	Total
Censura	735	57,597	**58,332**
Cierre	73,696	732,585	**806,281**
Total	**74,431**	**790,182**	**864,613**

Se obtiene un estadístico de prueba $\chi^2 = 4{,}293.48$ con g.l.= 1 por lo tanto es rechazada la hipótesis de independencia. ($p<0.001$), es decir el cierre de empresas depende del grado de marginación municipal.

En la Tabla 3.3 se observa la relación que existe entre las variables sector económico y estrato de personal ocupado:

Tabla 3.3: Tabla de contingencia de personal ocupado vs sector económico

Personal/Sector	Comercio	Industrias manufactureras	Resto de los sectores	Servicios no financieros	Total
0 a 5 personas	414,894	56,367	7,891	293,808	**772,960**
101 a 250 personas	989	634	367	1,719	**3,709**
11 a 30 personas	7,218	2,828	2,122	19,171	**31,339**
31 a 50 personas	1,407	731	516	4,642	**7,296**
51 a 100 personas	1,181	766	466	2,956	**5,369**
6 a10 personas	12,831	4,012	2,676	24,421	**43,940**
Total	**438,520**	**65,338**	**14,038**	**346,717**	**864,613**

Al evaluar la hipótesis de independencia se obtiene un estadístico de prueba $\chi^2 = 38{,}993.5$ con g.l.= 15 por lo tanto, también es rechazada la hipótesis de independencia. ($p<0.001$), es decir el tamaño de las empresas (personal contratado) depende del sector al que pertenece la empresa.

Tabla 3.4: Tabla de contingencia de personal ocupado vs Entidad

Estrato Personal/Entidad	Ciudad de México	Mexico	Total
0 a 5 personas	394,674	378,286	**772,960**
101 a 250 personas	2,893	816	**3,709**
11 a 30 personas	23,818	7,521	**31,339**
31 a 50 personas	5,979	1,317	**7,296**
51 a 100 personas	4,332	1,037	**5,369**
6 a10 personas	30,869	13,071	**43,940**
Total	**462,565**	**402,048**	**864,613**

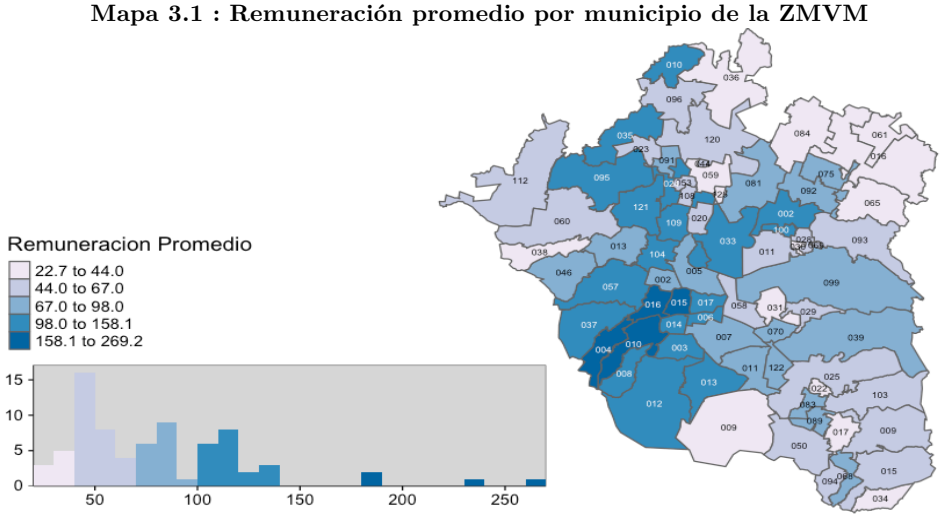
De igual manera cuando se compara la independencia entre las variables estrato de personal ocupado y entidad federativa (ver Tabla 3.4) rechazamos la hipótesis de independencia ($\chi^2 = 18{,}048.22$ con g.l.= 5, $p<0.0001$), es decir el tamaño de las empresas depende de la entidad en donde se ubica la empresa.

Estos resultados son importantes ya que cuando evaluemos el efecto en la supervivencia de cada una de esas variables podremos obtener mejores conclusiones.

Sería redundante realizar una prueba de independencia del evento de cierre con variables como el estrato de personal ocupado o el sector económico, ya que, de antemano sabemos que no existe la independencia debido a la forma en que se construyeron las probabilidades de cierre.

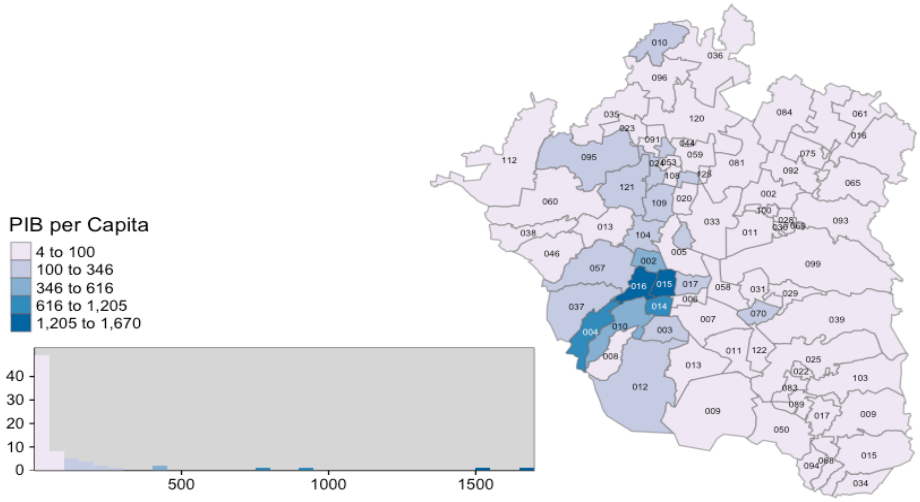
Por último se presentan mapas geográficos donde se muestra el nivel de las variables, Remuneración Promedio (ver Mapa 1.3), Índice de Marginación (ver Mapa 1.4) y PIB per cápita en los municipios de la ZMVM (ver Mapa 1.4).

Mapa 3.1 : Remuneración promedio por municipio de la ZMVM



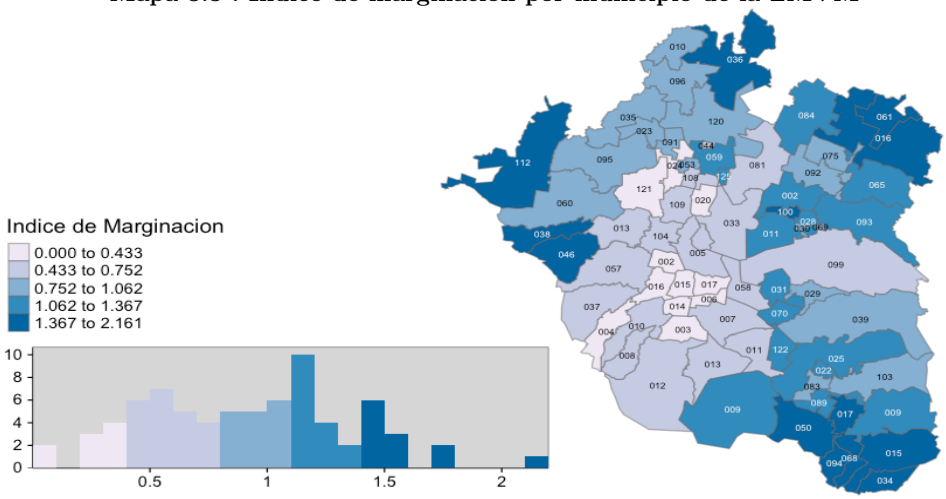
Podemos notar que los municipios más alejados del centro económico de la Ciudad de México son los que menor remuneración promedio obtienen anualmente, sobresaliendo por encima del resto las delegaciones Miguel Hidalgo, Cuauhtémoc y Cuajimalpa con remuneraciones anuales promedio de 269.24 mil, 230.93 mil y 188.44 mil pesos respectivamente. Siendo el promedio de la ZMVM de 112.42 mil pesos.

Mapa 3.2 : PIB per cápita por municipio de la ZMVM



En el caso del PIB per cápita tenemos las mismas delegaciones sobresalientes del resto, agregando además a la Delegación Benito Juárez. El promedio de esta variable es de 299.04 mil pesos.

Mapa 3.3 : Índice de marginación por municipio de la ZMVM

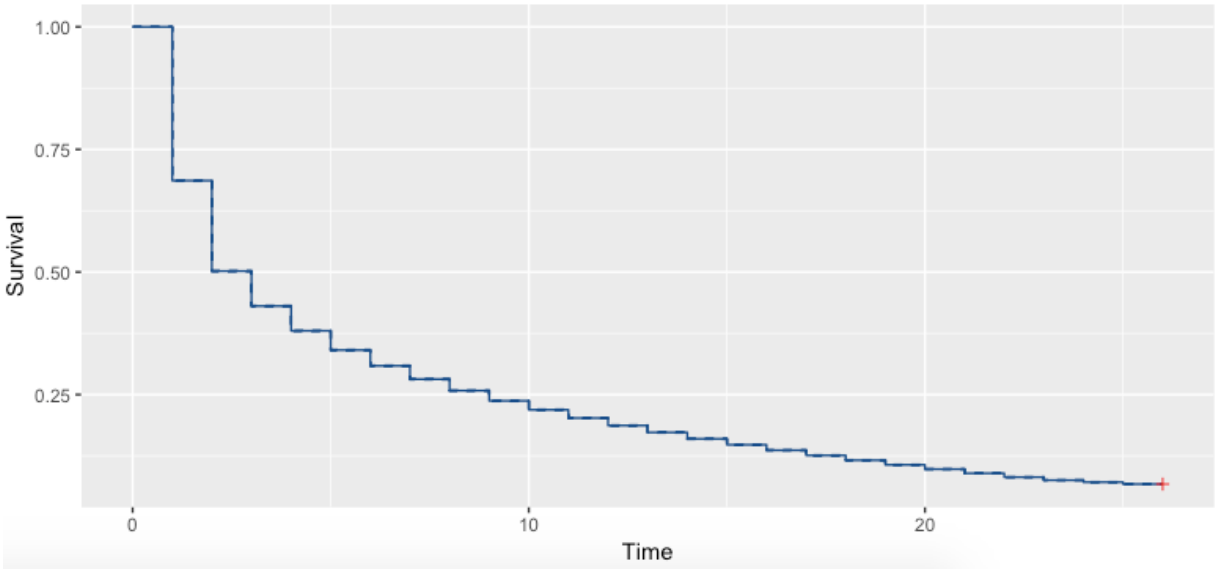


En el caso del índice de marginación se realizó una traslación dejando en nivel cero la delegación con menor nivel y a partir de allí todas tienen un valor positivo, podemos observar cómo también sobresalen los municipios más alejados al centro de la Ciudad de México como los que tienen mayor nivel de marginación.

3.3 Estimación no paramétrica (Kaplan Meier)

Al obtener una estimación de la función de supervivencia sobre el total de negocios mediante el método de Kaplan-Meier obtenemos la función escalonada mostrada en la Gráfica 3.5.

Gráfica 3.5: Función de supervivencia para establecimientos de la ZMVM por el método de KM



En la Tabla 3.5 se resumen los resultados de la estimación, podemos observar que los intervalos de confianza son muy cercanos al valor estimado, esto es debido al gran tamaño de muestra disponible.

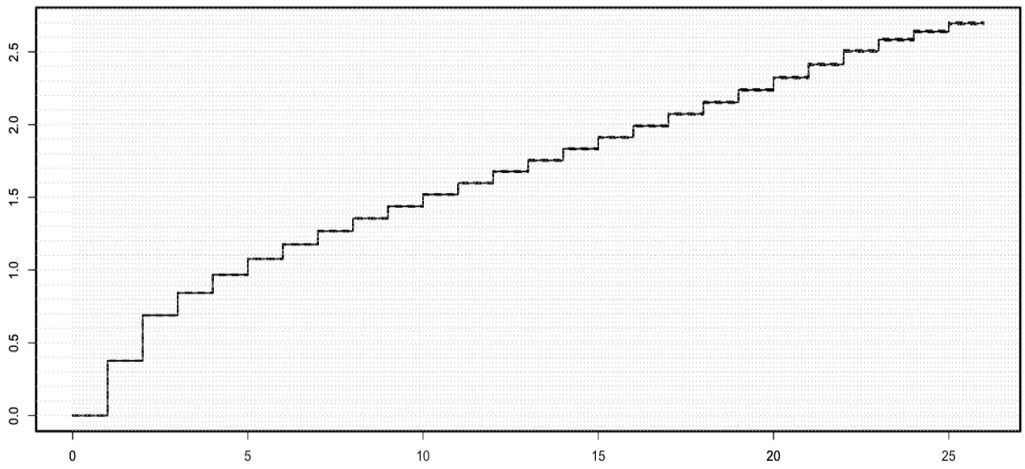
Tabla 3.5: Estimación de la supervivencia de los establecimientos mediante el método de KM

Tiempo(años)	Población en riesgo	No. de eventos	Supervivencia	Intervalo de confianza inferior 95%	Intervalo de confianza superior 95%
1	864,613	271,185	0.6864	0.6854	0.6873
2	593,428	159,585	0.5018	0.5007	0.5028
3	433,843	61,658	0.4305	0.4294	0.4315
4	372,185	43,624	0.3800	0.3790	0.3810
5	328,561	34,032	0.3406	0.3396	0.3416
6	294,529	27,701	0.3086	0.3076	0.3096
7	266,828	23,407	0.2815	0.2806	0.2825
8	243,421	20,313	0.2580	0.2571	0.2590
9	223,108	17,911	0.2373	0.2364	0.2382
10	205,197	15,945	0.2189	0.2180	0.2198
11	189,252	14,472	0.2021	0.2013	0.2030
12	174,780	13,219	0.1869	0.1860	0.1877
13	161,561	12,097	0.1729	0.1721	0.1737
14	149,464	11,262	0.1598	0.1591	0.1606
15	138,202	10,464	0.1477	0.1470	0.1485
16	127,738	9,698	0.1365	0.1358	0.1372
17	118,040	9,169	0.1259	0.1252	0.1266
18	108,871	8,532	0.1161	0.1154	0.1167
19	100,339	8,131	0.1066	0.1060	0.1073
20	92,208	7,511	0.0980	0.0973	0.0986
21	84,697	7,279	0.0895	0.0889	0.0901
22	77,418	6,872	0.0816	0.0810	0.0822
23	70,546	5,319	0.0754	0.0749	0.0760
24	65,227	3,535	0.0714	0.0708	0.0719
25	61,692	3,360	0.0675	0.0669	0.0680

Por definición la supervivencia en el tiempo 0 es igual a 1. Se observa que durante los primeros dos años casi la mitad de los nuevos establecimientos habrán cerrado y que para el año 20 quedará solamente cerca de un 10% del total inicial.

La función estimada de riesgo acumulado nos ayuda a observar cómo el comportamiento de éste es más elevado durante los primeros cinco años como se observa en la Gráfica 3.6, por este motivo, podemos rechazar la hipótesis de que la función de riesgo sea constante.

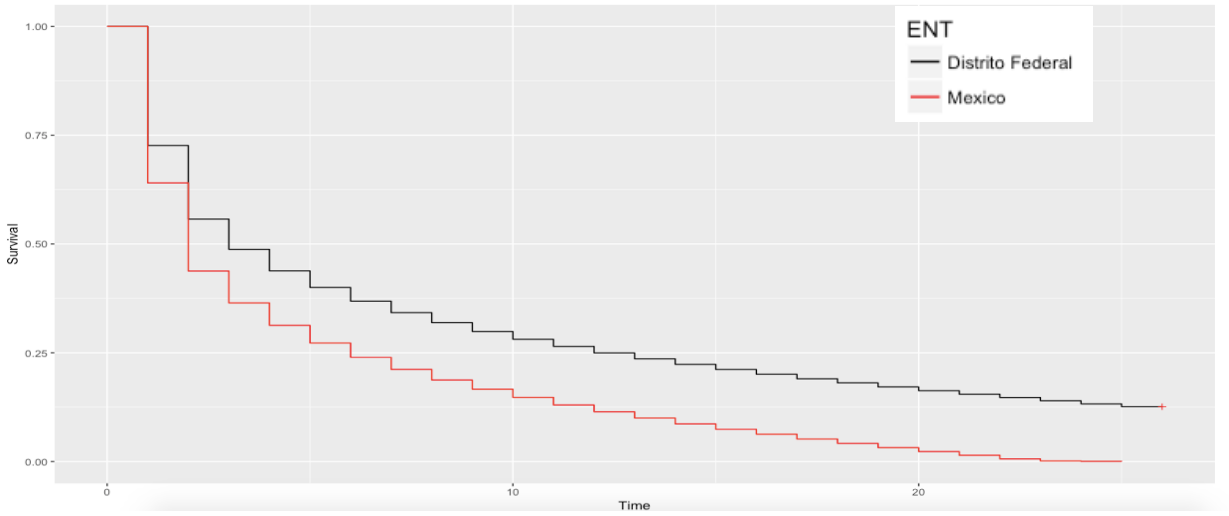
Gráfica 3.6: Función de riesgo acumulado para establecimientos de la ZMVM.



3.3.1 Comparación de la supervivencia entre grupos

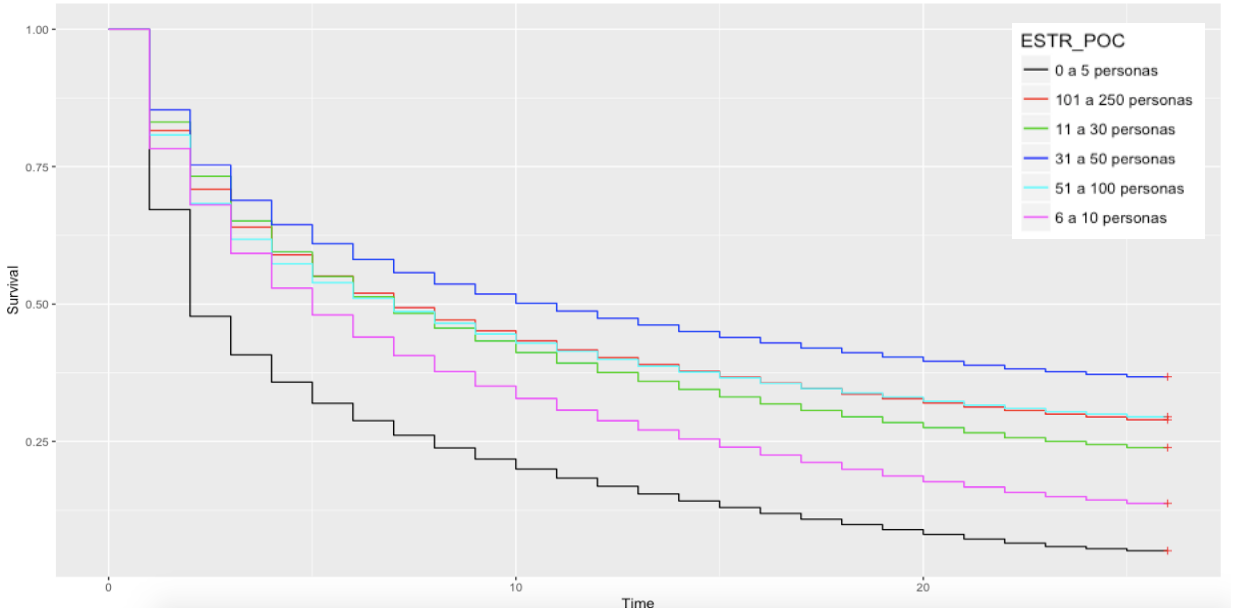
A continuación se presentan las estimaciones de la función de supervivencia para los distintos estratos con los que se pueden caracterizar los negocios de la ZMVM. En la Gráfica 3.5 se muestra que la supervivencia de negocios es más alta en la Ciudad de México comparada con la del Estado de México.

Gráfica 3.5: Funciones de supervivencia estimadas para cada entidad



Según los resultados obtenidos, la supervivencia de los establecimientos de *0 a 5 personas* es considerablemente menor que en el resto de los tamaños. Cabe destacar que el grupo de *31 a 50 personas* es el que menor cierres de establecimientos ha experimentado y no el de *101 a 250 personas* como se observa en la Gráfica 3.6:

Gráfica 3.6: Funciones de supervivencia estimadas para cada estrato de personal ocupado



Además la supervivencia de los negocios pertenecientes al *resto de los sectores* es mayor que en el de los sectores *manufacturero*, *comercio* y *servicios no financieros* , mientras que el sector más propenso al cierre de negocios es del *comercio*; esto es debido a que el tamaño de empresas que forman parte de este sector son en un 96.2 % de *0 a 5 personas*.

Gráfica 3.7: Funciones de supervivencia estimadas para cada sector económico

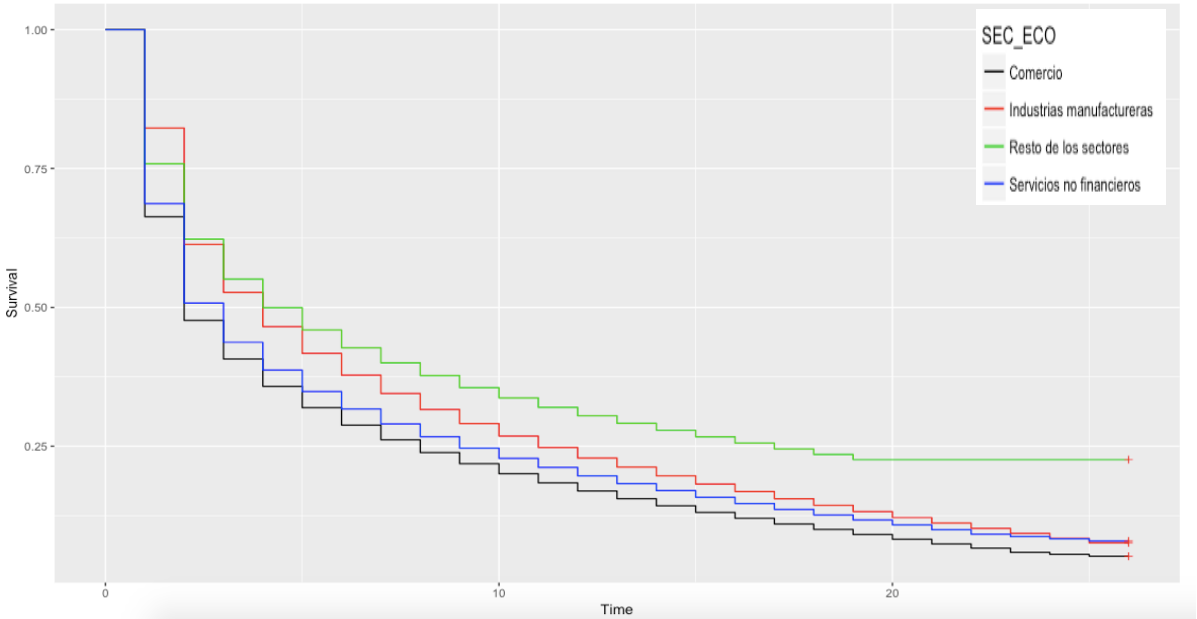


Tabla 3.6: Estimación de la supervivencia de los establecimientos para los distintos grupos de Entidad, Sector Económico y Personal ocupado

Tiempo(años)	CDMX	EdoMex	Resto sectores	Industrias manufactureras	Servicios no financieros	Comercio	0-5	6-10	11-30	31-50	51-100	101-250
1	0.726	0.640	0.759	0.823	0.687	0.663	0.672	0.783	0.831	0.853	0.808	0.816
2	0.557	0.438	0.623	0.614	0.508	0.477	0.478	0.680	0.733	0.753	0.683	0.709
3	0.488	0.365	0.551	0.527	0.437	0.407	0.408	0.592	0.651	0.689	0.618	0.640
4	0.438	0.313	0.500	0.465	0.387	0.358	0.358	0.529	0.595	0.644	0.573	0.589
5	0.400	0.272	0.459	0.417	0.348	0.319	0.319	0.480	0.550	0.610	0.539	0.551
6	0.369	0.240	0.427	0.378	0.317	0.288	0.288	0.440	0.513	0.581	0.510	0.520
7	0.342	0.212	0.400	0.345	0.290	0.262	0.261	0.406	0.483	0.557	0.486	0.494
8	0.319	0.188	0.377	0.316	0.267	0.239	0.238	0.377	0.456	0.536	0.465	0.471
9	0.299	0.166	0.355	0.291	0.247	0.218	0.218	0.351	0.433	0.518	0.446	0.451
10	0.281	0.147	0.337	0.268	0.228	0.200	0.200	0.328	0.412	0.502	0.429	0.433
11	0.265	0.130	0.320	0.248	0.212	0.184	0.183	0.307	0.393	0.487	0.414	0.417
12	0.250	0.115	0.305	0.229	0.197	0.169	0.168	0.288	0.375	0.474	0.399	0.403
13	0.236	0.100	0.291	0.212	0.183	0.155	0.155	0.271	0.359	0.462	0.387	0.390
14	0.223	0.087	0.279	0.197	0.170	0.143	0.142	0.254	0.345	0.450	0.376	0.377
15	0.212	0.074	0.267	0.182	0.158	0.131	0.130	0.240	0.331	0.440	0.366	0.367
16	0.201	0.063	0.256	0.168	0.147	0.120	0.119	0.225	0.318	0.429	0.356	0.356
17	0.190	0.052	0.245	0.155	0.136	0.110	0.108	0.212	0.306	0.420	0.346	0.346
18	0.181	0.042	0.235	0.144	0.126	0.100	0.099	0.199	0.295	0.411	0.338	0.336
19	0.171	0.032	0.226	0.132	0.117	0.091	0.089	0.187	0.285	0.404	0.331	0.328
20	0.163	0.023	0.226	0.121	0.108	0.082	0.081	0.177	0.275	0.396	0.323	0.320
21	0.155	0.014	0.226	0.111	0.100	0.074	0.073	0.167	0.266	0.389	0.316	0.312
22	0.147	0.006	0.226	0.102	0.092	0.066	0.065	0.157	0.257	0.382	0.310	0.306
23	0.140	0.001	0.226	0.093	0.087	0.059	0.059	0.150	0.250	0.377	0.304	0.300
24	0.133	0.001	0.226	0.084	0.083	0.055	0.055	0.143	0.244	0.372	0.300	0.295
25	0.126	0.001	0.226	0.076	0.079	0.052	0.051	0.137	0.239	0.368	0.295	0.290
26	0.126	0.001	0.226	0.076	0.079	0.052	0.051	0.137	0.239	0.368	0.295	0.290

En la Tabla 3.6 se muestran las distintas estimaciones de la supervivencia obtenidas para los distintos grupos; para validar las conclusiones referentes a las diferencias en la supervivencia de los distintos grupos de establecimientos se realizaron pruebas de *Log-Rangos* a las diferentes estratificaciones de los negocios, los resultados se muestran en la Tabla 3.7, es claro que la hipótesis de igualdad en la supervivencia de los distintos grupos se rechaza en los tres casos.

Tabla 3.7: Prueba de igualdad para distintas funciones de supervivencia

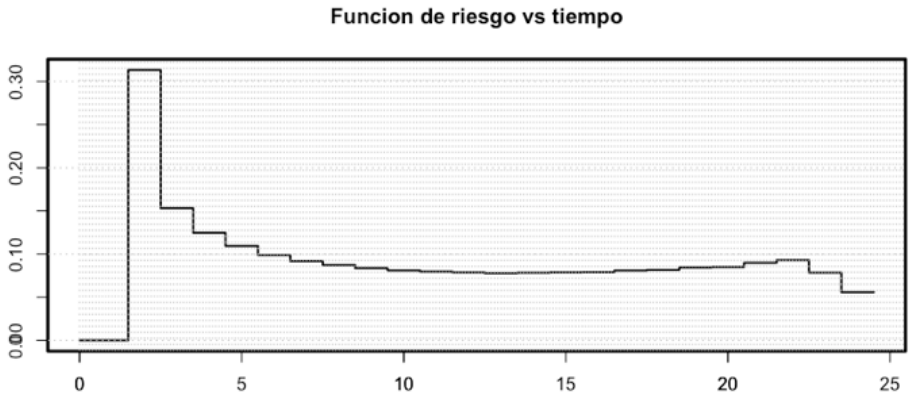
Estratificación	χ^2	g.l.	sig.
Entidad	52,092	1	p<(0.0001)
Personal Ocupado	26,073	5	p<(0.0001)
Sector Económico	6,552	3	p<(0.0001)

Antes de iniciar el ajuste de un modelo paramétrico donde sea posible la inclusión de variables explicativas, es importante mencionar que el ajuste de un modelo semi-paramétrico como el de riesgos proporcionales sería inadecuado ya que la hipótesis de riesgos proporcionales no se cumple, esto lo podemos observar en las Gráficas 3.6 y 3.7 donde es claro que la función estimada de supervivencia para distintos grupos se intersectan.

3.4 Ajuste de un modelo paramétrico

Para la selección de un modelo paramétrico primero se observó la función de riesgo, esto con el objetivo de descartar distribuciones que no pudieran modelar nuestros tiempos de falla; la estimación de ésta función se observa en la Gráfica 3.8:

Gráfica 3.8: Función de riesgo estimada.



Como previamente se había descartado un comportamiento constante de la función de riesgo la distribución exponencial claramente no sería útil.

La función de riesgo observa un comportamiento creciente durante el primer año y decreciente casi de inmediato; se evaluará el ajuste con las distribuciones Weibull, Log-normal y Log-logística debido a que nos permiten modelar funciones de riesgo con comportamientos decrecientes.

En la Gráfica 3.9 se muestra la comparación del ajuste entre los modelos Weibull, Log-logístico y Log-normal comparadas con la estimación empírica de la supervivencia mediante el método de Kaplan-Meier.

Para evaluar cuál es el mejor ajuste entre las dos distribuciones se utilizará el criterio de información de Akaike (AIC), el cual está definido por $AIC = 2k - 2 \log \hat{L}$, donde k es el número de parámetros incluidos en cada modelo. Para un conjunto de modelos candidatos es preferible el que tiene el valor mínimo en el AIC.

Gráfica 3.9: Ajuste de distribuciones Weibull, Log-logística y Log-normal

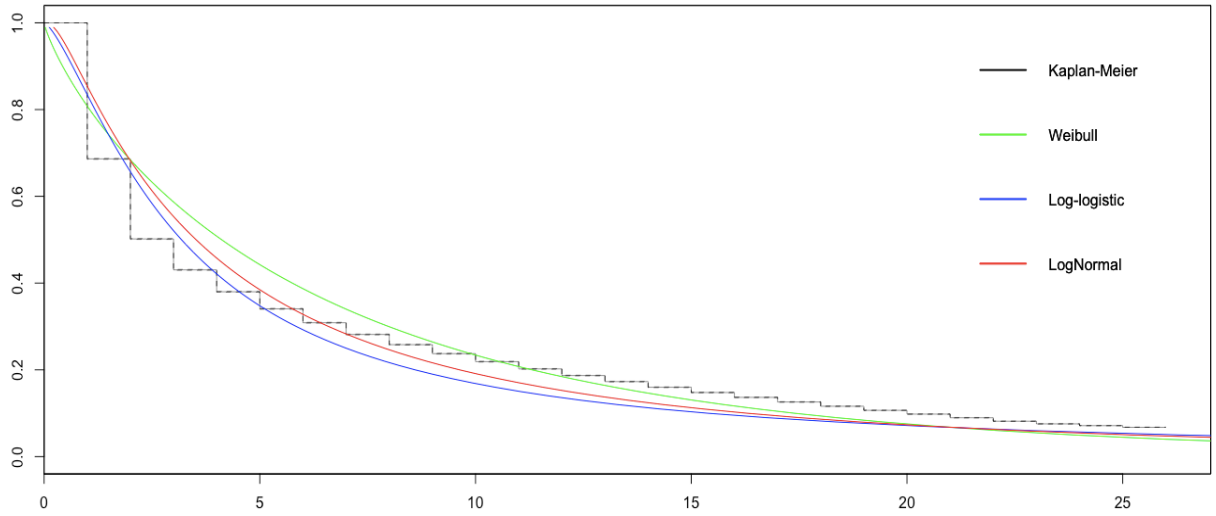


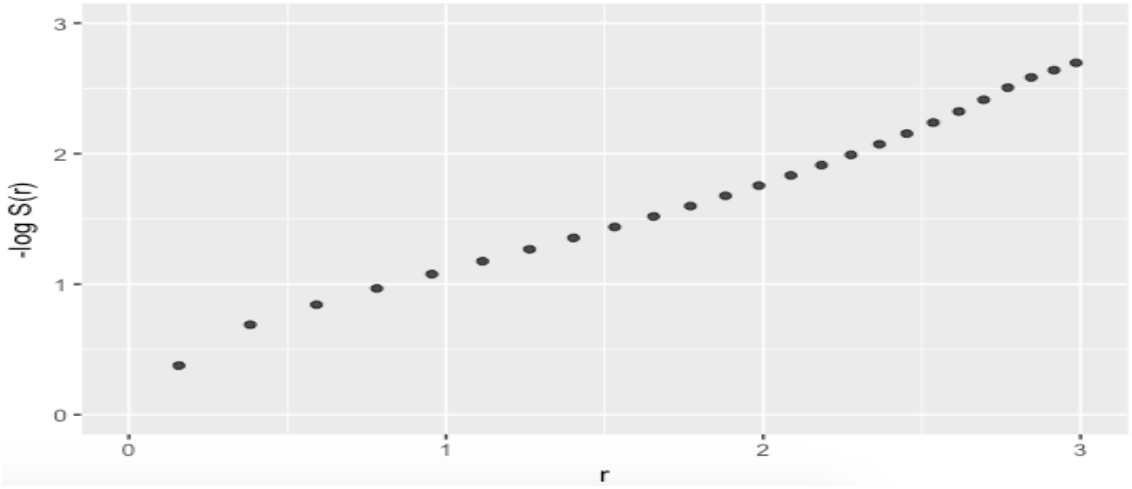
Tabla 3.8: AIC para los modelos ajustados

Modelo	-2 *log* $\hat{L}$	AIC
Log-normal	4,467,386	4,467,390
Log-logístico	4,505,936	4,505,940
Weibull	4,688,674	4,688,678

La distribución elegida sería la Log-normal, como se indica en la Tabla 3.8; para validar esa elección se realizó una prueba gráfica de residuales de Cox-Snell.

Como se muestra en la Gráfica 3.10 los residuales r_i graficados contra *-log* $S_R(r_i)$ donde $S_R(r_i)$ es la estimación de la supervivencia de Kaplan Meier de los residuales, se observa que la gráfica es relativamente cercana a una línea recta con pendiente 1 e intersección en cero, por lo tanto, la selección de una distribución Log-normal sería adecuada (Lee, 2003; p.215).

Gráfica 3.10: Prueba de Residuales de Cox-Snell



En la siguiente sección se evaluará la inclusión de covariables en el modelo paramétrico; el análisis consiste en evaluar si existe influencia de las variables explicativas en los tiempos de supervivencia.

3.4.1 Inclusión de variables explicativas

Para identificar las covariables que deberían ser incluidas en el modelo Log-normal de vida acelerada, los valores del estadístico - *2 log* $\hat{L}$, serán evaluados en todos los modelos ajustados con todas las combinaciones de variables explicativas originales en la base de datos, posteriormente se agregarán una a una las variables a nivel municipal para validar si tienen un efecto significativo en el modelo.

Las variables disponibles a incluir en los modelos son: entidad *(ent),* sector económico *(sec_eco),* personal ocupado *(estr_poc).* A nivel municipal se encuentran las variables: Remuneración promedio *(remun_prom),* PIB per cápita*(pib_per_cap),* Índice de marginación *(im_t).*

Tabla 3.9: Estadístico -2 *log* $\hat{L}$ para los modelos ajustados con las variables originales

Variables en el modelo	**-2 *log* $\hat{L}$**
ninguna	4,467,386
entidad	4,435,658
sec_eco	4,460,102
estr_poc	4,438,600
entidad+sec_eco	4,429,132
entidad+estr_poc	4,413,714
sec_eco+estr_poc	4,434,670
entidad+sec_eco+estr_poc	4,409,616

Como se muestra en la Tabla 3.9 según el estadístico -2 *log* $\hat{L}$, el modelo que mejor ajuste tiene incluye las variables: *entidad*, *sector económico* y *estrato de personal ocupado*, comparado con el modelo que contiene las variables *entidad* y *estrato de personal ocupado* la diferencia en el estadístico -2 *log* $\hat{L}$, es de 4,098 la cual se distribuye χ^2 con 3 grados de libertad ($p<0.0001$) por lo tanto, existe evidencia para concluir que la variable *sector económico* tiene un efecto significativo en la supervivencia.

Al incluir en el modelo seleccionado las variables a nivel municipal obtenemos los siguientes resultados del estadístico -2 *log* $\hat{L}$*:*

Tabla 3.10: Cambios en -2 $log\ \hat{L}$ al incluir variables a nivel municipal

Variables en el modelo	-2 $log\ \hat{L}$
entidad+sec_eco+estr_poc+remun_prom	4,409,616
entidad+sec_eco+estr_poc+pib_per_capita	4,409,614
entidad+sec_eco+estr_poc+im_t	4,409,616
entidad+sec_eco+estr_poc+remun_prom+pib_per_capita+im_t	4,409,608

Como se puede observar en la Tabla 3.10 los cambios en el modelo al agregar cada una de las variables explicativas a nivel municipal no son significativos, ni de forma individual, ni actuando en conjunto; por lo tanto, se rechaza su inclusión en el modelo.

Este era un resultado esperado considerando que la correlación de estas variables con el indicador de falla es muy baja como se mostró en el diagrama de correlaciones de la Gráfica 3.4.

Tabla 3.11: Estimación de los coeficientes en el modelo seleccionado

Modelo Ajustado	est	se	exp(est)	z	p
meanlog	2.5290	0.0171			
sdlog	1.1598	0.0009			
ent **mexico**	-0.4039	0.0025	0.6677	-159.0130	0.0000
sec_eco **industrias manufactureras**	0.0733	0.0111	1.0760	6.6156	0.0000
sec_eco **servicios no financieros**	-0.2010	0.0103	0.8179	-19.5704	0.0000
sec_eco **comercio**	-0.2266	0.0103	0.7973	-22.0241	0.0000
estr_poc **51 a 100 personas**	-0.2904	0.0216	0.7479	-13.4318	0.0000
estr_poc **101 a 250 personas**	-0.2578	0.0242	0.7728	-10.6407	0.0000
estr_poc **11 a 30 personas**	-0.2718	0.0157	0.7620	-17.3123	0.0000
estr_poc **6 a 10 personas**	-0.5418	0.0153	0.5817	-35.5094	0.0000
estr_poc **0 a 5 personas**	-0.9503	0.0143	0.3866	-66.4317	0.0000

Las variables explicativas incluidas en el modelo son estadísticamente significativas como se observa en la Tabla 3.11; estas se integraron con la característica de ser factores con distintos niveles, siendo el nivel de referencia para cada variable el grupo en el que la supervivencia fuera mayor; por ejemplo para la variable entidad existen dos niveles (0=Ciudad de México y 1=Estado de México) por lo que al interpretar la estimación del coeficiente se concluye que el pertenecer al Estado de México reduce la supervivencia de una empresa en un 33 %.

La variable sector económico tiene 4 niveles (0=Resto de los sectores, 1=Industrias manufactureras, 2=Servicios no financieros, 3=Comercio). La interpretación de los coeficientes mostrados en la Tabla 3.11 nos muestra que pertenecer al sector de las *industrias manufactureras* mejora la supervivencia en un 7.6 % esto sucede debido a que en los primeros años observados, que es donde se observan más cierres, las *industrias manufactureras* observan una mayor

supervivencia. Por otra parte pertenecer a los sectores *servicios no financieros* y *comercio* afectan la supervivencia en un 18% y 20% respectivamente.

Para la variable que indica el estrato de personal ocupado existen 6 niveles; el grupo base corresponde a los negocios que se encuentran en el estrato de 31 a 50 personas ocupadas, para el estrato de 51 a 100 , 101 a 250 y 11 a 30 personas la supervivencia se reduce en promedio 24%, mientras que para los negocios de 6 a 10 y de 0 a 5 personas la supervivencia se ve afectada en un 42% y 61% respectivamente.

3.5 Análisis comparativo

Bajo los resultados obtenidos podemos darnos cuenta que los modelos no-paramétricos resultan una herramienta útil inclusive cuando los tiempos de falla cuentan con una distribución conocida, ya que se puede comparar las observaciones contra la distribución ajustada y validar gráficamente si el modelo puede resultar o no apto. Además siempre son útiles como un primer acercamiento al analizar los tiempos de falla.

Como se describió en el Capitulo II no siempre es factible ajustar a un conjunto de datos una distribución paramétrica; por eso las aproximaciones no-paramétricas juegan un papel fundamental en el análisis de supervivencia. El método de Kaplan Meier nos permitió observar que existían diferencias significativas entre los distintos grupos en los que se puede dividir la población de negocios. Además existen pruebas que nos permiten validar estadísticamente las diferencias entre los distintos grupos.

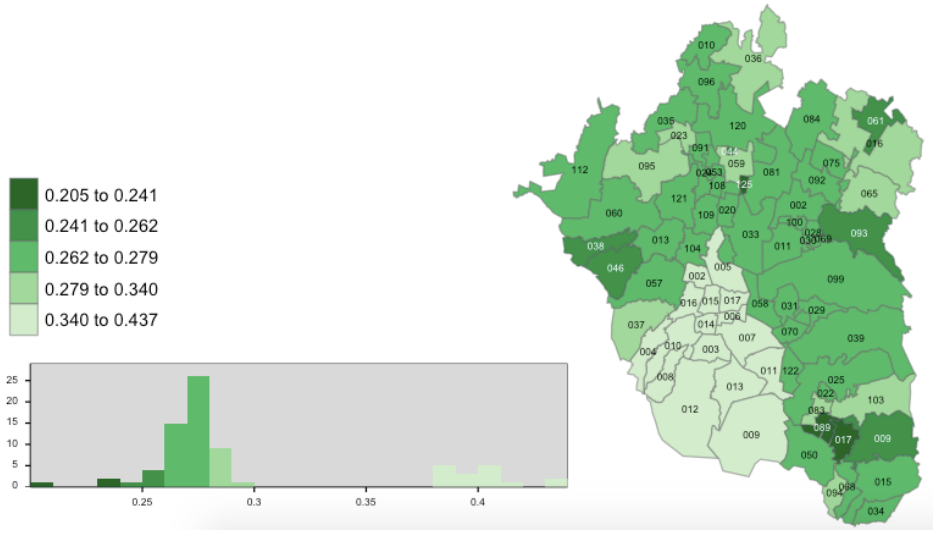
Una desventaja que existe en los modelos no paramétricos es la inclusión de variables explicativas o factores de riesgo; si bien en un modelo no-paramétrico un factor con distintos niveles puede servir para conocer las diferencias en la supervivencia entre distintos grupos de la población en un modelo paramétrico o semi-paramétrico se puede medir la influencia de ese factor en los tiempos de falla aunado a otras variables dentro del mismo modelo.

Una alternativa para la inclusión de covariables es el modelo de riesgos proporcionales, ya que no es necesario asumir una distribución específica para los datos observados y se puede medir el efecto de las covariables en los tiempos de falla. Éste método ha sido muy usado en estudios clínicos ya que puede medir el efecto de dichas variables en la supervivencia de pacientes, en el caso de los establecimientos de la ZMVM no se cumplió tal supuesto, por lo tanto se procedió con el ajuste de un modelo completamente paramétrico.

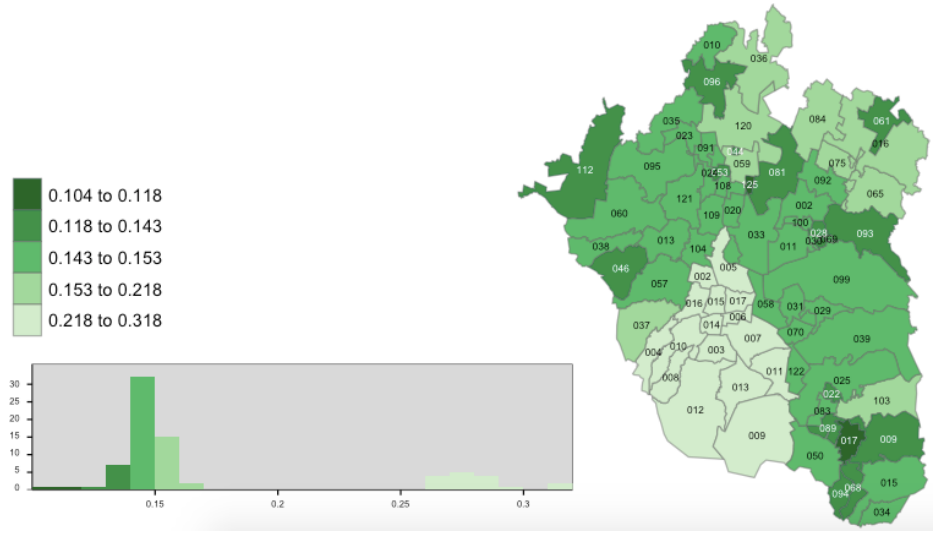
3.6 Evolución geográfica de la supervivencia

En los siguientes mapas se presenta un seguimiento quinquenal a la intensidad de la proporción de cierres de establecimientos en la ZMVM. Esta proporción se obtuvo con base en la simulación realizada de los tiempos de falla, en cada quinquenio se dividió el número de cierres entre el número total de establecimientos activos en ese periodo dentro de cada municipio.

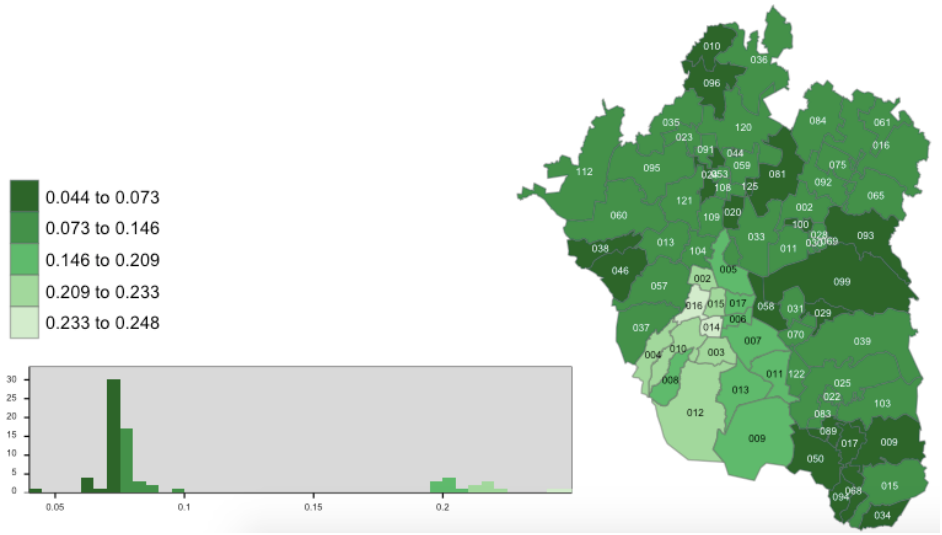
Mapa 3.4 : Proporción de cierres de 1 a 5 años en la ZMVM



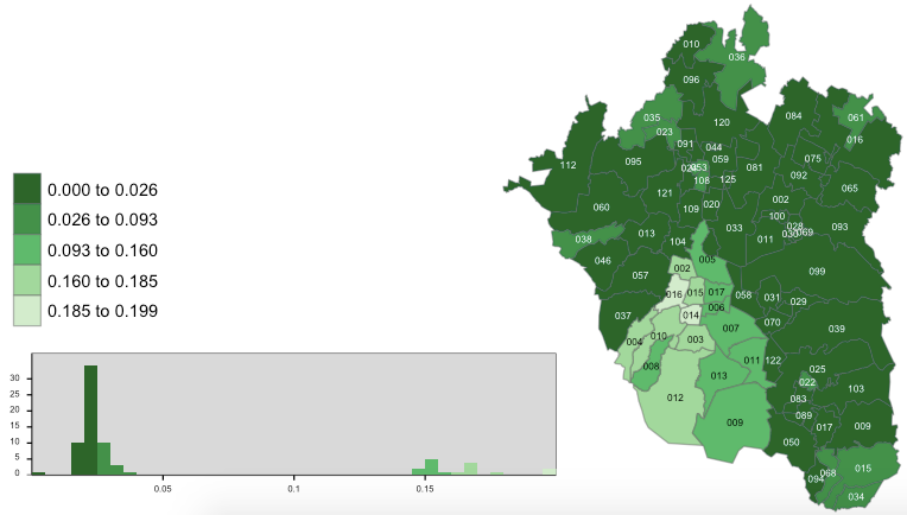
Mapa 3.5 : Proporción de cierres de 6 a 10 años en la ZMVM



Mapa 3.6 : Proporción de cierres de 11 a 15 años en la ZMVM



Mapa 3.5 : Proporción de cierres de 16 a 20 años en la ZMVM



Podemos observar que la proporción de cierres en el Estado de México es considerablemente mayor comparada con la Ciudad de México, esto también se debe a que el tamaño de la muestra disponible en el DENUE para los establecimientos del Estado de México era considerablemente menor y en la simulación de cierres realizada estos eran más probables de experimentar el evento.

También podemos afirmar que en los municipios y delegaciones donde menos proporción de cierres se observa son aquellos cuyos niveles de marginación son menores, esto es debido a que concentran gran parte de los negocios de mayores tamaños.

Conclusiones

En la presentación de resultados de este trabajo se observó cuales son los factores que más impacto tienen en el cierre de establecimientos, entre estos destacan principalmente, el tamaño de de una empresa medido por el número de personal ocupado, además del sector económico al que pertenezcan, también se identificó que son más propensas a cerrar las empresas cuya localización se encuentra en lugares con un menor desarrollo sociodemográfico.

Un punto a destacar es la fuerte dependencia que existe entre las variables utilizadas para estratificar la población, por ejemplo podemos concluir que el bajo nivel de supervivencia de las industrias del sector comercio se explica en gran parte porque cerca del 95% de estas se encuentran en el estrato de personal ocupado de 0 a 5 personas.

Otro de los principales hallazgos fue que la experiencia cierre en los negocios de 31 a 50 personas es significativamente menor que entre el resto de los tamaños por personal ocupado.

El presente trabajo además se presentó una aplicación del análisis de supervivencia en una base de datos generada con base en la experiencia de la esperanza de vida de los negocios. Dentro de los diferentes métodos que existen para analizar la supervivencia las estimaciones no paramétricas nos permitieron validar las diferencias que existen entre los diferentes estratos de los negocios.

Se logró ajustar un modelo paramétrico de vida acelerada donde se pudo conocer el efecto de cada una de las covariables explicativas incluidas en el modelo, no obstante, las variables que se agregaron a nivel municipal como el nivel de ingresos promedio o el índice de marginación no resultaron significativas para explicar su efecto directo en la supervivencia.

Como conclusión es importante destacar el hecho de que existe la posibilidad de aplicación de los modelos de supervivencia en diversos campos no tan conocidos como lo es la demografía económica además de que partiendo de resultados obtenidos en estudios previos es posible complementar este tipo de análisis.

Anexos

Dentro de este capítulo se presentan los códigos en el lenguaje de programación R utilizados para la generación de la base de datos con el indicador de supervivencia así como el procedimiento seguido para el ajuste de los distintos modelos de supervivencia.

A1. Generación de la base de datos con indicador de supervivencia

```
require(sampling) #Paquete para seleccionar los eventos=cierre establecimiento.
require(plyr) #Paquete para agregar datos.
require(dplyr) #Paquete para agregar datos.

##Se carga la base del DENUE de Ciudad de México

DENUE_INEGI_09_<read.csv("/DENUE_INEGI_09_.csv",encoding="latin")
DENUE_INEGI_09_<-subset.data.frame(DENUE_INEGI_09_,ESTR_POC=="0 a 5
personas"|ESTR_POC=="6 a 10 personas"|ESTR_POC=="11 a 30 personas"|ESTR_POC=="31
a 50 personas"|ESTR_POC=="51 a 100 personas"|ESTR_POC=="51 a 100
personas"|ESTR_POC=="101 a 250 personas")

##Creación de una variable de EDAD

DENUE_INEGI_09_$EDAD<-0

##Se carga la variable con información de la probabilidad (qx) del evento

tabla.cmdx<-read.csv("qxcdmx.csv",encoding="latin")

##Se preparan dos bases de trabajo

DENUE_INEGI_09_->base_origen
DENUE_INEGI_09_->base

##Se prepara la base para acumular los establecimientos que experimentan el evento.

evento_cum<-NULL
for (i in 1:25)

{
  ## Se asocian las probabilidades del evento dependiendo del ESTR_POC, SEC_ECO, EDAD
```

```
  merge(base,tabla.cmdx,by=c("ESTR_POC","SEC_ECO","EDAD"))->base

  ## Se excluyen los estratos cuya suma  de probabilidades es menor a 1

  base<-ddply(base,.(ESTR_POC,SEC_ECO,EDAD),transform,prob=sum(qx))
  base<-base[base$prob>=1,-ncol(base)]

  ## De acuerdo de las probabilidades del evento se calculan el numero de eventos para cada
estrato definido por "SEC_ECO","EDAD"y "ESTR_POC"

  tamagno<-aggregate(x=base$qx,by=list(base$EDAD,base$SEC_ECO,base$ESTR_POC), FUN
= sum)
  tamagno1<-trunc(tamagno$x)

 ## Ordenamiento de la base por "SEC_ECO","EDAD"

  base<-base[order(base$ESTR_POC,base$SEC_ECO,base$EDAD),]

 ## Selección de las unidades con el evento de acuerdo a la probabilidad qx

  s=strata(base,stratanames=c("ESTR_POC","SEC_ECO","EDAD"),
  size=as.numeric(tamagno1),method="srswor",description=TRUE)

  ## Se crea una base las unidades con el evento de acuerdo

  base[s$ID_unit,-ncol(base) ] ->evento

  ## Se crea un indicador del evento y se almacena para acumular todas las unidades con el
evento.

  evento$indicador<-1
  evento_cum<-rbind.fill(evento_cum,evento)

  ## De la base original se excluyen las unidades que experimentaron el evento.

  base<-base_origen[-match(evento_cum$D_LLAVE,base_origen$D_LLAVE),]

  ## Las unidades que no experimentaron el evento, envejecen un año.

  base$EDAD<-base$EDAD+i
}

## Se aplica la censura a las unidades que no experimentaron el evento después de 35 anos.

base$indicador<-0

## Ordenamiento de la base de unidades con censura, por "SEC_ECO" y "EDAD"

base<-base[order(base$ESTR_POC,base$SEC_ECO,base$EDAD),]
```

```
#Se unen las bases con los indicadores de evento=1, censura=0

rbind.fill(base,evento_cum)->DENUE_INEGI_09_surv

##Se repite el proceso para los establecimientos del Estado de México

DENUE_INEGI_15_ <- read.csv("/DENUE_INEGI_15_.csv",encoding="latin")
DENUE_INEGI_15_<-subset.data.frame(DENUE_INEGI_15_,ESTR_POC=="0 a 5
personas"|ESTR_POC=="6 a 10 personas"|ESTR_POC=="11 a 30 personas"|ESTR_POC=="31
a 50 personas"|ESTR_POC=="51 a 100 personas"|ESTR_POC=="51 a 100
personas"|ESTR_POC=="101 a 250 personas")

DENUE_INEGI_15_$EDAD<-0

tabla.edomex<-read.csv("/qxmex.csv",encoding="latin")

DENUE_INEGI_15_->base_origen
DENUE_INEGI_15_->base

evento_cum<-NULL
for (i in 1:25)
{
  merge(base,tabla.edomex,by=c("ESTR_POC","SEC_ECO","EDAD"))->base
  base<-ddply(base,.(ESTR_POC,SEC_ECO,EDAD),transform,prob=sum(qx))
  base<-base[base$prob>=1,-ncol(base)]
  tamagno<-aggregate(x=base$qx,by=list(base$EDAD,base$SEC_ECO,base$ESTR_POC), FUN
= sum)
  tamagno1<-trunc(tamagno$x)
  base<-base[order(base$ESTR_POC,base$SEC_ECO,base$EDAD),]
  s=strata(base,stratanames=c("ESTR_POC","SEC_ECO","EDAD"),
  size=as.numeric(tamagno1),method="srswor",description=TRUE)
  base[s$ID_unit,-ncol(base) ] ->evento
  evento$indicador<-1
  evento_cum<-rbind.fill(evento_cum,evento)
  base<-base_origen[-match(evento_cum$D_LLAVE,base_origen$D_LLAVE),]
  base$EDAD<-base$EDAD+i
}

base$indicador<-0
base<-base[order(base$ESTR_POC,base$SEC_ECO,base$EDAD),]
rbind.fill(base,evento_cum)->DENUE_INEGI_15_surv

## Se unen las dos base de la Ciudad de Mexico y el Estado de Mexico
```

```
rbind.fill(DENUE_INEGI_15_surv,DENUE_INEGI_09_surv)->DENUE_INEGI_09_15_surv
table(DENUE_INEGI_09_15_surv[DENUE_INEGI_09_15_surv$indicador==1, ]$EDAD)
write.csv(DENUE_INEGI_09_15_surv,"DENUE_INEGI_09_15_survival.csv")
```

A2. Ajuste de modelos de supervivencia

```
require(MASS)
require(survival)
require(fitdistrplus)
require(actuar)
require(muhaz)
require(survMisc)
require(ggplot2)
require(flexsurv)
require(plyr)
require(survminer)
require(dplyr)
require(rms)

## Se carga base de datos con los eventos de supervivencia

base<-read.csv("basezm.csv",encoding="latin")
base$EDAD<-base$EDAD+1
base$ESTR_POC<-factor(base$ESTR_POC,levels=c("31 a 50 personas","51 a 100 personas","101
a 250 personas","11 a 30 personas","6 a 10 personas","0 a 5 personas"))
base$SEC_ECO<-factor(base$SEC_ECO,levels=c("Resto de los sectores","Industrias
manufactureras","Servicios no financieros","Comercio"))

## creamos un objeto tipo supervivencia, indicamos el tiempo y el evento

data<-Surv(base$EDAD,base$indicador)

## ajustamos un modelo de KM

KM.loglog<-survfit(formula=data~1,conf.type="log-log")

## grafica de la función de supervivencia

ggsurvplot(surv.entidad,ggtheme = theme_gray())

## grafica de la función de riesgo acumulado
```

```
plot(surv.entidad,fun="cumhaz",main = "Función de riesgo acumulado", xlab= "Tiempo
",ylab="",lwd=2, col ="black")
box(lwd=3, col = "black")
axis(1, seq(0,35,10))
axis(2, seq(0,31,1))
abline(h = seq(0,3,.1), v =seq(0,35,.1),lty=3,col ="gray")

## supervivencia para distintos grupos

surv.entidad<-survfit(Surv(EDAD,indicador)~ENT,data=base)
ggsurv(surv.entidad,surv.col=c(1:2))

surv.sector<-survfit(Surv(EDAD,indicador)~SEC_ECO,data=base)
ggsurv(surv.sector,surv.col=c(1:4))

surv.personal<-survfit(Surv(EDAD,indicador)~ESTR_POC,data=base)
ggsurv(surv.personal,surv.col=c(1:6))

## se valida que las diferencias entre distintos grupos sean significativas

prueba.entidad<-survdiff(Surv(base$EDAD,base$indicador)~base$ENT)
prueba.sector<-survdiff(Surv(base$EDAD,base$indicador)~base$SEC_ECO)
prueba.personal<-survdiff(Surv(base$EDAD,base$indicador)~base$ESTR_POC)

## estimación función de riesgo##

riesgo<-kphaz.fit(base$EDAD,base$indicador,q=1,method="nelson")
kphaz.plot(riesgo,main = "Funcion de riesgo",lwd=2, col ="black")
box(lwd=3, col = "black")
axis(1, seq(0,35,10))
axis(2, seq(0,3,1))
abline(h = seq(0,3,.1), v =seq(0,35,.1),lty=3,col ="gray")


## Creamos un objeto del tipo supervivencia

base$SurvObj <- with(base, Surv(EDAD, indicador))


## Kaplan-Meier sin agrupamiento

km.null <- survfit(data = base, SurvObj ~ 1)
```

```
plot(km.null,main = "Funcion de supervivencia",lwd=0.5)
box(lwd=1, col = "black")

## Parametricos weibull , log logístico y log normal


weibull.null <- survreg(data = base, SurvObj ~ 1, dist = "weibull")
lines(x = predict(weibull.null, type = "quantile", p = seq(0.01, 0.99, by=.01))[1,],
      y = rev(seq(0.01, 0.99, by = 0.01)),
      col = "green")

loglogistic.null <- survreg(data = base, SurvObj ~ 1, dist = "loglogistic")
lines(x = predict(loglogistic.null, type = "quantile", p = seq(0.01, 0.99, by=.01))[1,],
      y = rev(seq(0.01, 0.99, by = 0.01)),
      col = "blue")

lognormal.null <- survreg(data = base, SurvObj ~ 1, dist = "lognormal")
lines(x = predict(lognormal.null, type = "quantile", p = seq(0.01, 0.99, by=.01))[1,],
      y = rev(seq(0.01, 0.99, by = 0.01)),
      col = "red")

legend(x = "topright",
       legend = c("Kaplan-Meier", "Weibull", "Log-logistic","LogNormal"),
       lwd = 1.5, bty = "n",
       col = c("black", "green","blue","red"))

## se valida el mejor ajuste mediante el criterio de información Akaike

extractAIC(loglogistic.null)
extractAIC(weibull.null)
extractAIC(lognormal.null)

## inclusión de variables explicativas en el modelo log-normal

lognormal.ninguna<-flexsurvreg(Surv(EDAD,indicador)~1,data=base,dist = "lognormal")
lognormal.entidad<-flexsurvreg(Surv(EDAD,indicador)~ENT,data=base,dist = "lognormal")
lognormal.sector<-flexsurvreg(Surv(EDAD,indicador)~SEC_ECO,data=base,dist = "lognormal")
lognormal.personal<-flexsurvreg(Surv(EDAD,indicador)~ESTR_POC,data=base,dist =
"lognormal")
lognormal.entidad.sector<-flexsurvreg(Surv(EDAD,indicador)~ENT + SEC_ECO ,data=base,dist
= "lognormal")
lognormal.entidad.personal<-flexsurvreg(Surv(EDAD,indicador)~ENT +
ESTR_POC,data=base,dist = "lognormal")
```

```
lognormal.sector.personal<-flexsurvreg(Surv(EDAD,indicador)~SEC_ECO +
ESTR_POC,data=base,dist = "lognormal")
lognormal.entidad.sector.personal<-flexsurvreg(Surv(EDAD,indicador)~ENT + SEC_ECO +
ESTR_POC,data=base,dist = "lognormal")

## se elige el modelo cuyo estadistico -2logl sea menor

loglogistic.entidad.sector.personal

## se evalúa la significancia dentro del modelo elegido de las variables a nivel municipal de
manera individual

lognormal.modelo.remuneracion<-flexsurvreg(Surv(EDAD,indicador)~ENT + SEC_ECO +
ESTR_POC+REMUN_PROM ,data=base,dist = "lognormal")
lognormal.modelo.pibpercapita<-flexsurvreg(Surv(EDAD,indicador)~ENT + SEC_ECO +
ESTR_POC+PIB_PER_CAP,data=base,dist = "lognormal")
lognormal.modelo.im_t<-flexsurvreg(Surv(EDAD,indicador)~ENT + SEC_ECO +
ESTR_POC+IM_T,data=base,dist = "lognormal")

lognormal.municipales <- flexsurvreg(Surv(EDAD,indicador)~ENT + SEC_ECO + ESTR_POC
+ REMUN_PROM +PIB_PER_CAP +IM_T+GM,data=base,dist = "lognormal") #de manera
conjunta

## se verifica mediante el estadístico -2logl y los estimadores de los coeficientes de las variables
municipales que no son estadísticamente significativos para el modelo

## se realiza una prueba de residuales Cox.Snell

lognormal<-survreg(Surv(EDAD,indicador)~1,data=base,dist = "lognormal")

linFit <- predict(lognormal, type="lp")
sderr <- (log(base$EDAD)-linFit)/lognormal$scale

CoxSnellResidual <- function (standRes, weight=1, dist)
{
  standRes <- standRes[rep(seq_len(length(standRes)), weight)]
  if (dist=="lognormal") {csr <- -log(1-pnorm(standRes))}
  else if (dist=="weibull") {csr <- -log(exp(-exp(standRes)))}
}
cxsn <- CoxSnellResidual(standRes=sderr, dist="lognormal")
survfit(cxsn, SurvObj ~ 1)

kmcs = survfit(Surv(cxsn,base$indicador)~1)$surv
```

```
uniquecxsn<-unique(sortcxsn)
sortcxsn<-sort(cxsn)

cxsn<-as.data.frame(uniquecxsn)
kmcs<-as.data.frame(kmcs)

ggplot(cxsn,aes(x=uniquecxsn, y=-log(kmcs$kmcs)))+
  geom_point(shape=19,alpha=3/4)+
  xlim(0,3)+
  ylim(0,3)+
  labs(x="r",y="-log S(r)")

## Riesgos Proporcionales

ph<-coxph(Surv(EDAD,indicador)~ENT+SEC_ECO+ESTR_POC,data=base)
testph<-cox.zph(ph)
```

Bibliografía

[1] Collet, D.,(2003). *Modelling Survival Data in Medical Research*. Chapman and Hall.

[2] Lee, Elisa T., y Wang, John W., (2003). *Statistical methods for survival data analysis.* John Wiley and Sons.

[3] Lee, Elisa T., y Go, Oscar T., (1997). Survival Analysis in Public Health Research. *Annual Review of Public Health*, 18, 105-134.

[4] Wienke, Andreas (2010). *Frailty Models in Survival Analysis.* Chapman and Hall/CRC.

[5] Instituto Nacional de Estadística y Geografía (México). Análisis de la demografía de los establecimientos 2012 : metodología / Instituto Nacional de Estadística y Geografía.-- México : INEGI, c2013. 28 p.

[6] Instituto Nacional de Estadística y Geografía (México). Censos económicos 2014 : Ciudad de México / Instituto Nacional de Estadística y Geografía.-- México : INEGI, c2016. 90 p.

[7] Instituto Nacional de Estadística y Geografía (México). Documento metodológico de la demografía de los negocios en México Estadísticas Experimentales, Demografía Económica / Instituto Nacional de Estadística y Geografía.-- México : INEGI, c2016. 24 p.

[8] Instituto Nacional de Estadística y Geografía (México). Censos económicos 2014 : México / Instituto Nacional de Estadística y Geografía.-- México : INEGI, c2016. 94 p.

[9] Instituto Nacional de Estadística y Geografía (México). Censos económicos 2014 : Micro, pequeña, mediana y gran empresa: estratificación de los establecimientos / Instituto Nacional de Estadística y Geografía.-- México : INEGI, c2015. 221 p.

[10] Instituto Nacional de Estadística y Geografía (México). Censos económicos 2014 : Resumen de los resultados definitivos / Instituto Nacional de Estadística y Geografía.-- México : INEGI, c2016. 20 p.