



BENEMÉRITA UNIVERSIDAD
AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS FÍSICO
MATEMÁTICAS

INFLUENCIA DE FACTORES SOCIO-AMBIENTALES EN LA
INCIDENCIA DE DENGUE:
UN MODELO PARA BAJA CALIFORNIA SUR Y UN MODELO
GENERAL PARA LA REPÚBLICA MEXICANA

TESIS

QUE PARA OBTENER EL TÍTULO DE
LICENCIADA EN MATEMÁTICAS APLICADAS

PRESENTA

IRENE MARCELINO SALVADOR

DIRECTOR DE TESIS

Dr. JORGE VELÁZQUEZ CASTRO

PUEBLA, PUE.

NOVIEMBRE 2016

A mi familia, motivo de superación...

*”Las matemáticas son un lugar donde puedes hacer cosas
que no puedes hacer en el mundo real”.*

Marcus du Sautoy

Agradecimientos

A ese ser divino al que llamamos Dios por darme una maravillosa familia, la dicha de culminar la licenciatura, por iluminar cada momento de mi vida.

Mi profundo y sincero agradecimiento a mi asesor Dr. Jorge Velázquez Castro, por confiar en mí dándome la oportunidad de llevar a cabo este proyecto de tesis, por compartirme en gran cantidad de su conocimiento, apoyo, tiempo, paciencia, amistad y motivación para lograr culminar la tesis.

Agradezco a cada uno de los maestros que contribuyeron con su conocimiento y experiencias a lo largo de mi formación académica, por hacer que descubriera lo increíble y maravilloso del saber, por hacerme descubrir nuevos horizontes, en particular a mis sinodales, los doctores: Hortensia Josefina Reyes Cervantes, Francisco Solano Tajonar Sanabria y Víctor Hugo Vázquez Guevara, quiénes dedicaron de su tiempo para revisar mi tesis, hacerme observaciones y proporcionarme sugerencias para enriquecerla.

Agradezco especialmente a mis padres Marcela Salvador Morales y Jaime Marcelino Contreras, por darme la vida, amor, apoyo, comprensión, consejos, confianza, por su gran y verdadera amistad, por hacer de mí una persona de bien, por creer en mí, por reconfortarme cuando lo requiero, por los grandes sacrificios que hacen, los amo con todo mi ser, gracias por ser mis padres son los mejores.

A mis hermanos por ser los mejores amigos, por compartir su conocimiento, experiencias, por brindarme aliento, amor y alegría a mi vida, los amo. A mis sobrinas nueva alegría para la familia, motivo de ser y dar lo mejor de mí.

A mis amigos y compañeros, por brindarme apoyo, confianza, por compartir momentos académicos y de vida.

Introducción

Normalmente, muchas de las enfermedades son causadas por virus o parásitos minúsculos. Como caso particular si un mosquito pica a un humano o animal infectado puede contraer un virus junto con la sangre que succiona. El mosquito y el virus no se dañan el uno al otro, pero el virus se aloja y desarrolla en el interior del mosquito. Permitiendo así, transmitir el virus a otros seres humanos al picarlos. Esta es la dinámica básica de las enfermedades transmitidas por vectores y se le llama vector al agente que hace la función de vehículo de transmisión del virus.

El fenómeno que se analizó en este trabajo es la enfermedad del dengue cuyo vector es el mosquito *Aedes Aegypti*. El dengue es un problema importante de salud pública en México. Actualmente es considerada una de las enfermedades más importantes con respecto a su magnitud epidemiológica [18].

Su transmisión presenta un comportamiento complejo debido a la variedad de serotipos y a los desplazamientos de los individuos infectados. Además, los factores socio-ambientales juegan un papel importante para la reproducción y traslado del mosquito portador del virus. Según Getis [15], por medio de un estudio en Perú, encontró que hay una gran cantidad de vectores dentro y alrededor de las casas. En este trabajo se estudiará la influencia de cinco variables sociales y cuatro variables ambientales que propician el incremento del mosquito *Aedes Aegypti* y por tanto, la incidencia de dengue en la población. La identificación y clasificación de los factores sociales y ambientales que potencian la transmisión tiene un gran impacto en políticas de salud pública e identificación de zonas de riesgo. Se ha puesto énfasis en la elaboración de modelos epidemiológicos con factores que tomen en cuenta la propagación del mosquito. La tesis se divide en cuatro capítulos.

En el capítulo uno se revisa la teoría para elaborar un modelo de regresión lineal simple (MRLS), la técnica de mínimos cuadrados para estimar los coeficientes del MRLS y los supuestos con sus propiedades de ésta técnica. También se analiza la teoría de los modelos de regresión lineal múltiple (MRLM), la estimación de los coeficientes de regresión desde un enfoque matricial, los supuestos y propiedades de los estimadores de mínimos cuadrados para el MRLM, la inferencia estadística para contrastar un modelo completo con un modelo reducido. Además se introduce el concepto de correlación, sus propiedades y su empleo como un estadístico de bondad de ajuste en la regresión lineal.

En el capítulo dos se estudia el análisis de componentes principales (ACP). Se plantea el problema de ACP y se describe dicho análisis en tres enfoques diferen-

II

tes. Para mayor claridad de lo que se quiere obtener con el ACP se describe de forma explícita el cálculo de la primera, segunda y j -ésima componente principal. Se describen las propiedades y se da una interpretación de las componentes principales.

En el capítulo tres se describe la relación del mosquito *Aedes Aegypti* con el dengue, chikungunya y zika, se hace una descripción del ciclo biológico del mosquito *Aedes Aegypti*, además se exponen las características de cada virus que propician las tres enfermedades virales que pueden ser transmitidas por el *Aedes Aegypti*, finalmente se detalla la importancia del control y factores que influyen en el crecimiento de población del mosquito.

En el capítulo cuatro se aplican técnicas y conceptos expuestos en capítulos anteriores con el objetivo de obtener modelos predictivos de incidencia de dengue en función de las variables sociales y ambientales. Se describen las variables sociales y ambientales a utilizar para la búsqueda de los modelos de regresión lineal. Los datos fueron reportados por el Sistema Nacional de Vigilancia Epidemiológica (SINAVE), la Dirección General de Epidemiología (DGE) [3], el Instituto Nacional de Estadística y Geografía (INEGI) [1], el Servicio Meteorológico Nacional (SMN), la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) [10], la Secretaría de Desarrollo Social (SEDESOL) [5] y la Comisión Nacional del agua (CONAGUA).

Se encuentran dos MRLS para el estado de Baja California Sur con factores ambientales e incidencia de dengue como variable dependiente del año 2014. Posteriormente se obtienen dos modelos de regresión lineal simple para México considerando datos agregados de los 32 estados. En el primer modelo se hace uso de la variable temperatura mínima y el otro usando precipitaciones.

Se propone también un MRLM para México que relacione variables sociales y ambientales, salvo que ahora los datos son los reportados en el año 2011 y considerando 27 estados. Para determinar el MRLM se procede primero con un ACP para reducir la dimensión y tomar componentes principales que son combinación lineal de las variables originales. Se realiza el cálculo de los coeficientes de mínimos cuadrados del MRLM para México, además de que se estiman dos MRLM uno denominado completo y el otro reducido. Se realiza una prueba de hipótesis con una distribución F , para contrastar estos dos últimos modelos. Se exponen de manera resumida resultados y conclusiones así como el trabajo futuro.

En el apéndice A se exponen algunos conceptos y demostraciones matemáticas que se emplean en el desarrollo del trabajo. En el apéndice B se incluye gran parte del código en python usado para generar las distintas gráficas y resultados del análisis de estudio.

Índice general

Introducción	I
1. Regresión Lineal	1
1.1. Regresión Lineal Simple	1
1.1.1. Estimación de los coeficientes del MRLS	2
1.1.1.1. Método de mínimos cuadrados para el MRLS	2
1.1.1.2. Supuestos y propiedades de los estimadores de mínimos cuadrados para el MRLS	4
1.2. Regresión Lineal Múltiple	7
1.2.1. Estimación de los coeficientes del MRLM	8
1.2.1.1. Método de mínimos cuadrados para el MRLM	8
1.2.2. MRLM enfoque matricial	9
1.2.2.1. Estimación matricial de los coeficientes para el MRLM	9
1.2.2.2. Supuestos y propiedades de los estimadores de mínimos cuadrados para el MRLM	10
1.2.3. Inferencia para la significancia de la regresión	14
1.3. Correlación	16
1.3.1. Estimación de ρ	17
1.3.2. Características del coeficiente de correlación muestral (R)	19
2. Análisis de Componentes Principales (ACP)	20
2.1. Planteamiento del problema	21
2.1.1. Enfoque descriptivo	21
2.1.2. Enfoque estadístico	22
2.1.3. Enfoque geométrico	22
2.2. Cálculo de las componentes principales	23
2.2.1. Cálculo de la primera componente principal	24
2.2.2. Cálculo de la segunda componente principal	26
2.2.3. Cálculo de la j -ésima componente principal	27
2.2.4. Generalización de las componentes principales	29

2.3.	Propiedades de las componentes principales	30
2.4.	Interpretación de las componentes principales	31
3.	Aedes Aegypti y su Relación con el Dengue, Chikungunya y Zika	32
3.1.	Ciclo biológico del Aedes Aegypti	32
3.2.	Aedes Aegypti como vector de virus	35
3.3.	Zika	35
3.4.	Chikungunya	36
3.5.	Dengue	37
3.6.	Importancia del control y factores que influyen en la población del Aedes Aegypti	39
4.	Análisis de Efectos Ambientales y Sociales en la Incidencia de Dengue en México	40
4.1.	Datos de estudio	40
4.2.	Efectos ambientales en la incidencia de dengue en Baja California Sur (BCS)	41
4.2.1.	Efecto de temperatura en la incidencia de dengue en BCS	42
4.2.2.	Retardo en la aparición de brotes epidemiológicos usando el factor de temperatura para BCS	45
4.2.3.	Efecto de la precipitación en la incidencia de dengue en BCS	46
4.2.4.	Retardo en la aparición de brotes epidemiológicos usando el factor de precipitación para BCS	49
4.3.	Efectos ambientales en la incidencia de dengue en México	50
4.3.1.	Efecto de temperatura en la incidencia de dengue en México	50
4.3.2.	Efecto de precipitación en la incidencia de dengue en Méxi- co	52
4.4.	Inclusión de factores sociales en un modelo de regresión lineal múltiple de incidencia de dengue	53
4.4.1.	Análisis de correlación entre los factores Socio-ambientales e incidencia de dengue en México	55
4.4.2.	Análisis de componentes principales	60
4.4.3.	Cálculo de los estimadores de mínimos cuadrados para el MRLM para México	62
4.4.4.	Prueba de mejor ajuste entre modelo lineal completo y reducido	66
	Conclusiones	68
	Apéndice	69

<i>ÍNDICE GENERAL</i>	v
Índice de figuras	89
Índice de cuadros	91
Bibliografía	92

Capítulo 1

Regresión Lineal

Gran parte de los pronósticos estadísticos del tiempo está basado en la técnica conocida como regresión lineal, la cual puede ser simple o múltiple, dependiendo del número de variables independientes que se tengan para obtener una respuesta. Se detalla cada una de esta clasificación en las siguientes secciones.

1.1. Regresión Lineal Simple

Es una técnica estadística para el modelado e investigación que permite estudiar la relación entre una variable respuesta Y que depende de una variable independiente o predictora X y medir la fuerza de asociación. La relación resultante puede ser empleada para hacer predicciones, optimizaciones o control de la variable dependiente mediante la variable independiente. Si se encuentra una asociación, se infiere que la variación en X puede causar variación en Y , o viceversa.

Muchas funciones matemáticas diferentes se pueden usar para modelar una respuesta que sea de una o más variables independientes. Esta sección sólo se basará en el estudio de una sola variable independiente.

Es decir, si se desea una ecuación que describa la relación entre dos variables y pueda ser utilizada para predecir valores desconocidos, es necesario ajustar las observaciones a un modelo matemático. Para esto, el investigador dispone de una muestra de tamaño n de observaciones de la población. El conjunto de observaciones se pueden representar de la forma (X_i, Y_i) . Es posible elaborar un diagrama de dispersión del conjunto de los n pares de observaciones. Esto ayudará a observar el comportamiento que deberá seguir el modelo a obtener y que nos describa la relación entre la variable dependiente e independiente. Un modelo en el cual se proponga una relación lineal entre las variables de un conjunto de n pares de observaciones (X_i, Y_i) se dice que es un modelo de regresión lineal simple el cual se define como:

Definición 1.1

Un modelo estadístico o modelo de regresión lineal simple (MRLS) relaciona una respuesta aleatoria Y dependiente con una variable X independiente de la forma:

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (1.1)$$

en donde β_0, β_1 son parámetros desconocidos, β_0 es la ordenada al origen, β_1 es la pendiente y ϵ es un error aleatorio con media cero y varianza σ^2 .

Como la media del error aleatorio ϵ es cero, una ecuación equivalente a (1.1) es

$$E[Y] = \beta_0 + \beta_1 X.$$

Se dice que ϵ es el término que determina el error que se tiene para dar un modelo exacto.

1.1.1. Estimación de los coeficientes del MRLS

Una vez definido el modelo de regresión lineal simple se procede a estimar los coeficientes de regresión β_0 y β_1 , para esto se utiliza el método de “regresión de mínimos cuadrados ordinarios”, así β_0, β_1 serán los parámetros que mejor ajusten a la recta de regresión, que minimiza la suma de las distancias verticales al cuadrado entre los puntos de los datos Y_i y la recta estimada $\hat{Y}$.

1.1.1.1. Método de mínimos cuadrados para el MRLS

¹ El objetivo del método de mínimos cuadrados es encontrar los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ de β_0 y β_1 que minimicen la suma de los errores cuadráticos (SSE):

$$SSE = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

En donde, SSE representa la variabilidad observada que no es explicada o determinada por el MRLS.

Por tanto los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ deben satisfacer las siguientes ecuaciones:

$$\begin{aligned} \left. \frac{\partial SSE}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1} &= -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0, \\ \left. \frac{\partial SSE}{\partial \beta_1} \right|_{\hat{\beta}_0, \hat{\beta}_1} &= -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0. \end{aligned}$$

¹Esta sección está basada principalmente en [16], [22].

Después de un poco de álgebra, de las ecuaciones anteriores se tiene

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n X_i &= \sum_{i=1}^n Y_i, \\ \hat{\beta}_0 \sum_{i=1}^n X_i + \hat{\beta}_1 \sum_{i=1}^n X_i^2 &= \sum_{i=1}^n Y_i X_i. \end{aligned} \quad (1.2)$$

Las ecuaciones (1.2) son conocidas como ecuaciones normales de mínimos cuadrados y la solución de éstas dan como resultado $\hat{\beta}_0$ y $\hat{\beta}_1$, estimadores de mínimos cuadrados de β_0 y β_1 .

Definición 1.2

Las estimaciones de mínimos cuadrados de la ordenada al origen y la pendiente del modelo de regresión lineal simple son

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned} \quad (1.3)$$

en donde $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ y $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

Así, la recta de regresión ajustada es

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

Una forma de escribir el numerador y denominador de $\hat{\beta}_1$ es la siguiente

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{(\sum_{i=1}^n Y_i)(\sum_{i=1}^n X_i)}{n}}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} = \frac{S_{xy}}{S_{xx}},$$

en donde

$$S_{xx} = \sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n} = \sum_{i=1}^n (X_i - \bar{X})^2$$

y

$$S_{xy} = \sum_{i=1}^n X_i Y_i - \frac{(\sum_{i=1}^n Y_i)(\sum_{i=1}^n X_i)}{n} = \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}),$$

además,

$$S_{yy} = \sum_{i=1}^n Y_i^2 - \frac{(\sum_{i=1}^n Y_i)^2}{n} = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

El coeficiente de regresión estimado $\hat{\beta}_1$ es de gran importancia ya que indica la forma en que están relacionadas X y Y .

Ahora bien, una forma conveniente de escribir SSE para su cálculo es:

$$\begin{aligned}
 SSE &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\
 &= \sum_{i=1}^n [(Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})]^2 \\
 &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) + \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2 \\
 &= S_{yy} - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1^2 S_{xx} = S_{yy} - \hat{\beta}_1 S_{xy}.
 \end{aligned} \tag{1.4}$$

1.1.1.2. Supuestos y propiedades de los estimadores de mínimos cuadrados para el MRLS

² Para determinar el modelo de regresión lineal simple, se establecen algunas suposiciones estadísticas necesarias.

- $X_1, X_2, \dots, X_n$ son observaciones de la variable independiente.
- β_0 y β_1 son parámetros desconocidos que determinan el modelo de regresión $E[Y] = \beta_0 + \beta_1 X$, este modelo es lineal respecto a los parámetros β_0 y β_1 . Notemos que si en lugar de una recta se utilizara una curva con grado mayor a uno en X pero grado uno en los coeficientes β_0 y β_1 , el modelo de regresión sigue siendo lineal.
- $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ son variables aleatorias independientes distribuidas normalmente, con media cero y varianza σ^2 , esto es

$$\epsilon_i \sim N(0, \sigma^2); \quad i = 1, 2, \dots, n.$$

Lo anterior se generaliza en el siguiente resultado.

Teorema 1.3

Las variables aleatorias Y_i son independientes, y cada una de ellas tiene una distribución normal con media $\mu = \beta_0 + \beta_1 X_i$ y varianza σ^2 , es decir,

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2); \quad i = 1, 2, \dots, n.$$

²Basado principalmente en [14], [16].

Del Teorema 1.3 se sigue que $\hat{\beta}_0$ y $\hat{\beta}_1$ dependen de los valores de Y por lo que los estimadores de mínimos cuadrados de los coeficientes de regresión son vistos como variables aleatorias. Se estudiarán algunas propiedades de $\hat{\beta}_0$ y $\hat{\beta}_1$. Consideremos primero la propiedad de sesgo para el valor esperado de $\hat{\beta}_1$,

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}} E\left[\sum_{i=1}^n Y_i(X_i - \bar{X})\right] \\ &= \frac{1}{S_{xx}} E\left[\sum_{i=1}^n (X_i - \bar{X})(\beta_0 + \beta_1 X_i + \epsilon_i)\right] \\ &= \frac{1}{S_{xx}} \left\{ E\left[\beta_0 \sum_{i=1}^n (X_i - \bar{X})\right] + E\left[\beta_1 \sum_{i=1}^n X_i(X_i - \bar{X})\right] \right\} \\ &\quad + \frac{1}{S_{xx}} \left\{ E\left[\sum_{i=1}^n \epsilon_i(X_i - \bar{X})\right] \right\} \\ &= \frac{1}{S_{xx}} \beta_1 S_{xx} = \beta_1. \end{aligned}$$

Puesto que $\sum_{i=1}^n (X_i - \bar{X}) = 0$ y $\sum_{i=1}^n X_i(X_i - \bar{X}) = S_{xx}$, además, por hipótesis se tiene que $E(\epsilon_i) = 0$, por lo que $\hat{\beta}_1$ es un estimador insesgado de β_1 . Ahora bien, notemos que la varianza de $\hat{\beta}_1$, dado que $V(\epsilon_i) = \sigma^2$ es

$$V(\hat{\beta}_1) = V\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2} V\left[\sum_{i=1}^n Y_i(X_i - \bar{X})\right]. \quad (1.5)$$

Como las variables $\{Y_i\}$ no están correlacionadas ya que las variables $\{\epsilon_i\}$ no lo están, la suma de la ecuación (1.5) es la suma de las varianzas, por lo que

$$V(\hat{\beta}_1) = \frac{1}{S_{xx}^2} \sigma^2 \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{\sigma^2}{S_{xx}}.$$

Para calcular la esperanza y varianza de $\hat{\beta}_0$ se procede de manera similar que en $\hat{\beta}_1$ y se obtiene que

$$E(\hat{\beta}_0) = \beta_0$$

y

$$V(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} \right] = \frac{\sigma^2 \sum X_i^2}{(nS_{xx})}.$$

Por tanto, $\hat{\beta}_0$ es un estimador insesgado de β_0 . La covarianza de las variables $\hat{\beta}_0$ y $\hat{\beta}_1$ no es cero y está determinada por

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{X}\sigma^2}{S_{xx}}.$$

Notemos que $\hat{\beta}_0, \hat{\beta}_1$ están correlacionadas y por tanto, son dependientes a menos que $\bar{X} = 0$. Estos resultados se resumen en el siguiente Teorema.

Teorema 1.4

Distribuciones de los estimadores de mínimos cuadrados.

Los estimadores de la ordenada al origen β_0 , la pendiente β_1 y la recta de regresión $E[Y] = \beta_0 + \beta_1 X$ tienen las siguientes distribuciones:

$$\begin{aligned}\hat{\beta}_0 &\sim N\left[\beta_0; \frac{\sigma^2 \sum X_i^2}{nS_{xx}}\right] \\ \hat{\beta}_1 &\sim N\left[\beta_1; \frac{\sigma^2}{S_{xx}}\right] \\ \hat{Y}_{X_0} &\sim N(\beta_0 + \beta_1 X_0; \sigma_{\hat{Y}_{X_0}}^2)\end{aligned}$$

en donde X_0 es un valor fijo de X y $\sigma_{\hat{Y}_{X_0}}^2 = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}} \right]$.

Con el fin de hacer inferencias usando los estimadores, es necesario estimar σ^2 , que nos indicará qué tan buena será la predicción del ajuste obtenido.

Al explicar la idea de mínimos cuadrados se dijo que los estimadores minimizan:

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{\epsilon}_i^2.$$

Así, $\hat{\epsilon}_i = Y_i - \hat{Y}_i$, puede considerarse un estimador de ϵ_i , ahora bien, se calcula la varianza por medio de $\hat{\epsilon}_i$ ya que ϵ_i es una variable aleatoria con media cero y varianza σ^2 , el estimador insesgado para σ^2 es S^2 ya que $E(S^2) = \sigma^2$, esto se establece en el siguiente Teorema.

Teorema 1.5

Una estimación insesgada de σ^2 es:

$$S^2 = \frac{SSE}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2},$$

en donde $S_{yy} = \sum_{i=1}^n (Y_i - \bar{Y})^2$ y $S_{xy} = \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})$.

La demostración se muestra en el Apéndice A.1.1.

1.2. Regresión Lineal Múltiple

En la sección 1.1 se estudiaron conceptos y la técnica de regresión lineal simple en donde hay, sólo una variable de regresión independiente, no obstante hay muchos problemas de investigación en los que una sola variable independiente no es suficiente. Cuando en el problema de estudio hay diversas variables independientes que se relacionan linealmente con una variable dependiente se puede usar la técnica de regresión lineal múltiple.

El modelo de regresión lineal múltiple (MRLM) es una herramienta estadística poderosa y flexible a disposición de investigadores. Está técnica es utilizada en distintas áreas. Por ejemplo, la respuesta de un animal de laboratorio a algún fármaco puede depender de la dosis, edad, y el peso del animal. O bien, que el administrador de un hospital le interese estudiar las admisiones de diversas comunidades a las que atiende el hospital, y desee determinar qué factores parecen ser los que influyen en las diferencias observadas en las tasas de admisión.

En el MRLM se supone que existe una relación lineal entre alguna variable Y dependiente cuyos valores queremos predecir y k variables independientes. En donde $X_1, X_2, \dots, X_k$ son variables o bien, son variables aleatorias observables, es decir, cualquier inferencia que se haga de los datos de la muestra sólo se aplica al conjunto de valores de X observados.

En el análisis de regresión lineal múltiple, la ecuación de regresión ya no define una recta en el plano, sino un hiperplano en un espacio multidimensional. La definición de este tipo de modelos se da a continuación [26].

Definición 1.6

Un modelo estadístico lineal múltiple o modelo de regresión lineal múltiple (MRLM) relaciona una respuesta aleatoria Y con un conjunto de variables independientes $X_1, X_2, \dots, X_k$ de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon, \quad (1.6)$$

en donde $\beta_0, \beta_1, \dots, \beta_k$ son parámetros desconocidos, las variables $X_1, X_2, \dots, X_k$ toman valores conocidos y ϵ es un error aleatorio con media cero y varianza σ^2 , por tanto

$$E(Y) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k.$$

Al igual que en la sección uno, ϵ es el error aleatorio el cual representa la incapacidad para dar un modelo exacto.

1.2.1. Estimación de los coeficientes del MRLM

³ Al igual que en la sección 1.1.1, en donde sólo hay una variable independiente X , se estiman los coeficientes de regresión lineal múltiple por medio del método de mínimos cuadrados tomando en cuenta los n datos

$$(X_{i1}, X_{i2}, \dots, X_{ik}, Y_i); \quad i = 1, 2, \dots, n.$$

Notemos que X_{i1} es el i -ésimo valor de la variable X_1 , X_{i2} es el i -ésimo valor de la variable X_2 y así sucesivamente.

Para los parámetros desconocidos $\beta_j; j = 0, 1, 2, \dots, k$ se estimarán los valores de los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$.

1.2.1.1. Método de mínimos cuadrados para el MRLM

Cada observación satisface el modelo:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \epsilon_i; \quad i = 1, 2, \dots, n. \quad (1.7)$$

Así, la función de $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ que tiene que ser minimizada es:

$$\begin{aligned} SSE_M &= \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n [Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_k X_{ik}]^2 \\ &= \sum_{i=1}^n \left(Y_i - \beta_0 - \sum_{j=1}^k \beta_j X_{ij} \right)^2. \end{aligned}$$

Diferenciando parcialmente SSE_M con respecto a los $\beta_0, \beta_1, \dots, \beta_k$ se encuentra que los estimadores de mínimos cuadrados $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ deben satisfacer:

$$\begin{aligned} \left. \frac{\partial SSE_M}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} &= -2 \sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j X_{ij} \right) = 0, \\ \left. \frac{\partial SSE_M}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} &= -2 \sum_{i=1}^n \left(Y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j X_{ij} \right) X_{ij} = 0, \quad j = 1, \dots, k. \end{aligned}$$

³Esta sección se basa principalmente en [22].

Después de simplificar las ecuaciones anteriores, se obtienen las ecuaciones de mínimos cuadrados

$$\begin{aligned}
 n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n X_{i1} + \hat{\beta}_2 \sum_{i=1}^n X_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n X_{ik} &= \sum_{i=1}^n Y_i, \\
 \hat{\beta}_0 \sum_{i=1}^n X_{i1} + \hat{\beta}_1 \sum_{i=1}^n X_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n X_{i1}X_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n X_{i1}X_{ik} &= \sum_{i=1}^n X_{i1}Y_i, \\
 &\vdots \\
 \hat{\beta}_0 \sum_{i=1}^n X_{ik} + \hat{\beta}_1 \sum_{i=1}^n X_{ik}X_{i1} + \hat{\beta}_2 \sum_{i=1}^n X_{ik}X_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n X_{ik}^2 &= \sum_{i=1}^n X_{ik}Y_i.
 \end{aligned} \tag{1.8}$$

Nótese que hay $k + 1$ ecuaciones, que corresponde al número de coeficientes de regresión desconocidos y las soluciones de las ecuaciones son los estimadores de los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$.

1.2.2. MRLM enfoque matricial

Al ajustar un modelo de regresión lineal múltiple es más conveniente expresar las operaciones por medio de matrices. Algunos resultados y deducciones son análogos a la sección 1.2.1.1.

El MRLM está dado por la ecuación (1.7), en donde hay n observaciones independientes, $Y_1, Y_2, \dots, Y_n$, en $\mathbf{Y}$. Se definen las siguientes matrices con $X_{i0} = 1, i = 1, \dots, n$.

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{pmatrix} X_{10} & X_{11} & X_{12} & \dots & X_{1k} \\ X_{20} & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{n0} & X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Así, las n ecuaciones que representan a Y_i como función de $\mathbf{X}$, $\boldsymbol{\beta}$ y $\boldsymbol{\epsilon}$ matricialmente se escriben como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

1.2.2.1. Estimación matricial de los coeficientes para el MRLM

⁴ Una vez descrito y definido el modelo de regresión lineal múltiple de forma matricial, se procede a estimar los coeficientes de regresión del vector $\boldsymbol{\beta}$, para esto

⁴Esta sección se basa principalmente en [22], [27].

se utiliza el método de “regresión de mínimos cuadrados ordinarios”, como en la sección 1.1.1.1, así, el vector $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)^t$, contiene a las estimaciones de los parámetros que mejor ajusten al hiperplano de regresión. Para calcular un estimador $\hat{\beta}$ de β es necesario minimizar:

$$SSE_M = \sum_{i=1}^n \epsilon^2 = (\mathbf{Y}_i - \hat{\mathbf{Y}})^2 = \mathbf{Y}^t \mathbf{Y} - \hat{\beta}^t \mathbf{X}^t \mathbf{Y} = (\mathbf{Y} - \mathbf{X}\beta)^t (\mathbf{Y} - \mathbf{X}\beta).$$

De modo que en el valor estimado se tiene que cumplir:

$$\frac{\partial SSE_M}{\partial \beta_0} = 0.$$

Después de algunos pasos de álgebra se encuentra que

$$(\mathbf{X}^t \mathbf{X}) \hat{\beta} = \mathbf{X}^t \mathbf{Y}.$$

De lo anterior se establece el siguiente resultado.

Teorema 1.7

Las ecuaciones de mínimos cuadrados para el MRLM son

$$(\mathbf{X}^t \mathbf{X}) \hat{\beta} = \mathbf{X}^t \mathbf{Y}.$$

Cuyas soluciones están dadas por

$$\hat{\beta} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y},$$

en donde $\mathbf{X}^t$ es la transpuesta de $\mathbf{X}$ y $(\mathbf{X}^t \mathbf{X})^{-1}$ es la inversa de $\mathbf{X}^t \mathbf{X}$ además de que este último producto de matrices es no singular.

1.2.2.2. Supuestos y propiedades de los estimadores de mínimos cuadrados para el MRLM

Las propiedades estadísticas de los estimadores de mínimos cuadrados $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ pueden determinarse bajo ciertas suposiciones de los errores $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ que son idénticas al caso de la regresión lineal simple. Se consideran como variables aleatorias independientes distribuidas normalmente, con media cero y varianza σ^2 , esto es:

$$\epsilon_i \sim N(0, \sigma^2); \quad i = 1, 2, \dots, n.$$

Así, se tiene el siguiente resultado:

Teorema 1.8

Las variables aleatorias Y_i son independientes, y cada una de ellas tiene una distribución normal con media $\mu_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik}$ y varianza $\sigma_i^2 = \sigma^2$, es decir,

$$Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik}, \sigma^2).$$

Los estimadores $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ son insesgados

Esta propiedad se puede demostrar de la siguiente forma:

$$\begin{aligned} E(\hat{\beta}) &= E[(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}] = E[(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t (\mathbf{X} \beta + \epsilon)] \\ &= E[(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{X} \beta + ((\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \epsilon)] = \beta. \end{aligned}$$

Ya que $E[\epsilon] = 0$ y $(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{X} = \mathbf{I}$, la matriz identidad. Por tanto, $\hat{\beta}$ es un estimador insesgado de β .

Las varianzas de los estimadores $\hat{\beta}$ se obtienen de los elementos de la inversa de la matriz $(\mathbf{X}^t \mathbf{X})$, los elementos fuera de la diagonal de $(\mathbf{X}^t \mathbf{X})$ son las sumas de productos cruzados de los elementos de las columnas de $\mathbf{X}$, mientras que los elementos de la diagonal de $(\mathbf{X}^t \mathbf{X})$ representan las sumas de los cuadrados de los elementos en las columnas de $\mathbf{X}$.

La inversa de $(\mathbf{X}^t \mathbf{X})$ multiplicada por la constante σ^2 representa la matriz de varianza-covarianza de los coeficientes de regresión $\hat{\beta}$.

Los elementos de la diagonal de $\sigma^2 (\mathbf{X}^t \mathbf{X})^{-1}$ son las varianzas de $\hat{\beta}$, mientras que los elementos fuera de la diagonal son las covarianzas.

Matricialmente se escribe de la siguiente forma

$$\mathbf{C} = (\mathbf{X}^t \mathbf{X})^{-1} = \begin{pmatrix} C_{00} & C_{01} & C_{02} & \dots & C_{0k} \\ C_{10} & C_{11} & C_{12} & \dots & C_{1k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{k0} & C_{k1} & C_{k2} & \dots & C_{kk} \end{pmatrix},$$

en donde se observa que $\mathbf{C}$ es una matriz simétrica dado que $(\mathbf{X}^t \mathbf{X})$ lo es. Por tanto, se tiene que

$$V(\hat{\beta}_j) = \sigma^2 C_{jj}, \quad j = 0, 1, \dots, k$$

y

$$\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}; \quad j \neq i.$$

Así, la matriz de covarianza de $\hat{\beta}$ es una matriz simétrica de dimensión $(k \times k)$ y el jj -ésimo elemento es la varianza de $\hat{\beta}_j$, mientras que el ij -ésimo elemento representa a la covarianza entre $\hat{\beta}_i$ y $\hat{\beta}_j$, esto es

$$\text{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}^t \mathbf{X})^{-1} = \sigma^2 \mathbf{C}.$$

Para obtener las estimaciones de las varianzas, es necesario sustituir a σ^2 por un estimador apropiado S_M^2 . De manera similar al caso de regresión lineal simple, la estimación de σ^2 está definida en términos de la suma de cuadrados de los errores:

$$SSE_M = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (\mathbf{Y}_i - \hat{\mathbf{Y}}_i)^2 = \mathbf{Y}^t \mathbf{Y} - \hat{\beta}^t \mathbf{X}^t \mathbf{Y}.$$

Entonces, se tiene que

$$S_M^2 = \frac{SSE_M}{[n - (k + 1)]} = \frac{\mathbf{Y}^t \mathbf{Y} - \hat{\beta}^t \mathbf{X}^t \mathbf{Y}}{[n - (k + 1)]}.$$

Las propiedades mencionadas para el caso de regresión lineal se resumen en el siguiente Teorema.

Teorema 1.9

Propiedades de los estimadores de mínimos cuadrados (MRLM):

1. $E[\hat{\beta}_i] = \beta_i; \quad i = 0, 1, \dots, k.$
2. $V(\hat{\beta}_i) = \sigma^2 C_{ii}$ donde C_{ii} es el elemento de la fila i y la columna i de $(\mathbf{X}^t \mathbf{X})^{-1}$.
3. $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$ donde C_{ij} es el elemento de la fila i y la columna j de $(\mathbf{X}^t \mathbf{X})^{-1}$.
4. Un estimador insesgado de σ^2 es

$$S_M^2 = \frac{SSE_M}{[n - (k + 1)]} = \frac{\mathbf{Y}^t \mathbf{Y} - \hat{\beta}^t \mathbf{X}^t \mathbf{Y}}{[n - (k + 1)]}.$$

Nótese que hay $k + 1$ valores β_i desconocidas en el modelo.

Si, además, las ϵ_i variables aleatorias, para $i = 1, 2, \dots, n$ están distribuidas normalmente,

5. Cada $\hat{\beta}_i$ está distribuida normalmente, entonces
6. Los estadísticos S_M^2 y $\hat{\beta}_i$ son independientes para cada $i = 0, 1, 2, \dots, k$.
7. La variable aleatoria

$$\frac{[n - (k + 1)]S_M^2}{\sigma^2}$$
 tiene una distribución χ^2 con $n - (k + 1)$ grados de libertad.

Se puede notar que el Teorema 1.9, es una generalización de las propiedades para el caso de regresión lineal simple, así como también S_M^2 es una medida de la variación de los errores de predicción [26].

La suma de cuadrados del error y la regresión son de la misma forma y juegan el mismo papel que en el caso de regresión lineal simple. Lo anterior nos conduce a la siguiente definición.

Definición 1.10

En el análisis de regresión lineal múltiple se tienen las siguientes sumas de cuadrados

- Suma de los cuadrados de la regresión: $SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$.
- Suma de los cuadrados del error: $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$.
- Suma total de cuadrados de las desviaciones: $SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$.

Se cumple la identidad de suma de cuadrados siguiente

$$SST = SSR + SSE,$$

que es equivalente a escribir

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2,$$

y se tienen k grados de libertad asociados con SSR , para SST se tienen $n - 1$ grados de libertad. Al realizar la resta SSE tiene $[n - (k + 1)]$ grados de libertad.

Así, la estimación de σ^2 está determinada por la suma de cuadrados del error dividida entre sus grados de libertad.

Por consiguiente se anuncia el siguiente Teorema [27].

Teorema 1.11

Para el modelo de regresión lineal

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

La suma de cuadrados del error se puede escribir como

$$SSE = SST - SSR,$$

en donde $SST = (\mathbf{Y}^t \mathbf{Y}) - \frac{(\sum_{i=1}^n Y_i)^2}{n}$, $SSR = \sum_{j=0}^k (\hat{\beta}^t \mathbf{X}^t \mathbf{Y}) - \frac{(\sum_{i=1}^n Y_i)^2}{n}$.

La demostración se puede ver en el Apéndice A1.2.

1.2.3. Inferencia para la significancia de la regresión

⁵ En problemas de regresión lineal múltiple existen pruebas de hipótesis sobre los parámetros del modelo que son apropiadas para medir la adecuación del mismo. En esta sección se expondrá una técnica para una prueba de hipótesis.

Recordemos que $\epsilon_i \sim iidN[0; \sigma^2]$, es decir que las variables aleatorias ϵ_i son independientes e idénticamente distribuidas normalmente con media cero y varianza σ^2 .

Nótese que se está trabajando con un MRLM por lo que es común tener un modelo lineal con muchas variables independientes y se desea ajustar un modelo que contenga un subconjunto de las variables independientes $X_1, X_2, \dots, X_k$ sin reducir significativamente la precisión. Supongamos que se conoce el modelo

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \beta_{p+1} X_{p+1} + \dots + \beta_k X_k + \epsilon,$$

que denota el modelo lineal con todas las variables independientes, es decir, el **modelo completo**.

Consideremos el siguiente modelo lineal con un subconjunto de variables independientes, el cual es un **modelo reducido**.

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon.$$

Notemos que $p < k$, además, se puede calcular la suma de cuadrados de las desviaciones para cada modelo denotadas por SSE_C y SSE_R , respectivamente.

⁵Sección basada principalmente en [26].

Por tanto, el juego de hipótesis apropiadas es

$$H_0 : \beta_{p+1} = \beta_{p+2} = \dots = \beta_k = 0$$

vs

$$H_a : \beta_j \neq 0, \quad \text{al menos para un } j, j = p+1, \dots, k.$$

Se espera que el modelo completo prediga con un error menor que el modelo reducido. Sea $SSE_A = (SSE_R - SSE_C)$, donde SSE_A se le suele llamar **suma de cuadrados asociada** con las variables $X_{p+1}, X_{p+2}, \dots, X_k$ ajustada para las variables $X_1, X_2, \dots, X_p$, si el valor de SSE_A es grande habrá más evidencia de apoyar la hipótesis alternativa, es decir, que $X_{p+1}, X_{p+2}, \dots, X_k$ aportan información para la predicción de Y .

Si H_0 es verdadera, un estadístico de prueba está determinado en función de SSE_A , note que

$$SSE_R = SSE_C + (SSE_R - SSE_C) = SSE_C + SSE_A, \quad (1.9)$$

y como se supone H_0 verdadera, entonces

$$\frac{SSE_R}{\sigma^2}$$

tiene una distribución de probabilidad χ^2 con k grados de libertad. De la ecuación (1.9) la SSE_R se puede dividir en dos partes SSE_C y SSE_A , con un razonamiento semejante se tiene que:

$$\chi_1^2 = \frac{SSE_R}{\sigma^2}, \quad \text{tiene } (n - [p + 1]), \quad \text{grados de libertad.}$$

$$\chi_2^2 = \frac{SSE_C}{\sigma^2}, \quad \text{tiene } (n - [k + 1]), \quad \text{grados de libertad.}$$

$$\chi_3^2 = \frac{SSE_A}{\sigma^2}, \quad \text{tiene } (k - p), \quad \text{grados de libertad.}$$

En donde, n es el tamaño de la muestra, p el número de variables del modelo reducido y k es el número de variables del modelo completo.

Por otra parte, se tiene que χ_2^2 y χ_3^2 son estadísticamente independientes.

Recordemos la siguiente definición:

Definición 1.12

Sean W_1 y W_2 variables aleatorias independientes con distribución χ^2 con ν_1 y ν_2 grados de libertad, respectivamente. Entonces se dice que

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

tiene una distribución F con ν_1 grados de libertad en el numerador y ν_2 grados de libertad en el denominador.

De la definición 1.12 se tiene que

$$F = \frac{\chi^2_3/(k-p)}{\chi^2_2/(n-[k+1])} = \frac{(SSE_R - SSE_C)/(k-p)}{SSE_C/(n-[k+1])} = \frac{SSE_A/(k-p)}{SSE_C/(n-[k+1])}.$$

Ahora bien, si

$$H_0 : \beta_{p+1} = \beta_{p+2} = \dots = \beta_k = 0$$

es verdadera, entonces tiene una distribución F con $\nu_1 = k - p$ grados de libertad en el numerador y $\nu_2 = n - (k + 1)$ grados de libertad en el denominador. Así, para valores grandes de F favorecen el rechazo de H_0 . Si deseamos una prueba con nivel de significancia α se tiene que

$$F > F_\alpha,$$

es la región de rechazo.

1.3. Correlación

Cabe mencionar que el análisis de regresión estadística estudia la relación entre una variable independiente X y la media de una variable dependiente Y . La variable predictora X no es aleatoria, mientras que la respuesta Y sí lo es. Nuestro interés se enfoca en la relación entre X y la media de Y . En donde se busca una ecuación de línea recta de la forma $E[Y] = \beta_0 + \beta_1 X$, la cual nos sirve para estimar respuestas futuras y respuestas medias para valores de X dados.

En el análisis de correlación, X e Y son variables aleatorias y son de igual interés. El análisis de correlación es una herramienta útil cuyo objetivo es determinar si existe o no una relación lineal entre estas variables aleatorias.

Sea ρ el coeficiente de correlación de Pearson, la medida de asociación lineal entre dos variables aleatorias. El coeficiente ρ está determinado por la covarianza de X e Y , definida a continuación.

Definición 1.13

Sean X e Y variables aleatorias con medias μ_X y μ_Y respectivamente. La covarianza entre X e Y , está dada por:

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - E[X]E[Y].$$

Notemos lo siguiente:

- Si valores pequeños de X se asocian con valores pequeños de Y y valores grandes de X con valores grandes de Y , entonces $(X - \mu_X)$ y $(Y - \mu_Y)$ tienen el mismo signo algebraico y por tanto, una covarianza positiva.
- Si valores pequeños de X se asocian con valores grandes de Y y viceversa, entonces $(X - \mu_X)$ y $(Y - \mu_Y)$ tienen signos contrarios y por tanto, generan una covarianza negativa.

El problema que ahora se presenta es que el parámetro ρ no está acotado. Asume cualquier valor real y su magnitud carece de sentido. Para resolver el problema se divide la covarianza entre $\sqrt{(Var(X))(Var(Y))}$, obteniéndose así la siguiente definición:

Definición 1.14

Coeficiente de correlación de Pearson.

Sean X e Y variables aleatorias con medias μ_X , μ_Y y varianzas σ_X^2 , σ_Y^2 respectivamente. La correlación ρ entre X e Y es

$$\rho = \frac{Cov(X, Y)}{\sqrt{(Var(X))(Var(Y))}}.$$

1.3.1. Estimación de ρ

⁶ Notemos que $Cov(X, Y)$ y ρ no pueden calcularse si no se conoce la distribución de probabilidad de las variables X e Y , ya que sólo son parámetros teóricos. Se pretende estimar los valores de $Cov(X, Y)$ y ρ a partir de un conjunto de datos. La $Cov(X, Y)$ puede expresarse como la diferencia de sus medias teóricas, es decir:

$$Cov(X, Y) = E[XY] - E[X]E[Y]. \quad (1.10)$$

La $Cov(X, Y)$ puede calcularse sustituyendo cada media teórica por su media muestral correspondiente. Esto es

⁶Sección basada principalmente en [21].

$$\begin{aligned}
E[XY] & \text{ por } \frac{\sum_{i=1}^n x_i y_i}{n}, \\
E[X] & \text{ por } \frac{\sum_{i=1}^n x_i}{n}, \\
E[Y] & \text{ por } \frac{\sum_{i=1}^n y_i}{n}.
\end{aligned} \tag{1.11}$$

Por otra parte, se tiene que

$$Var(X) = E[X^2] - (E[X])^2,$$

y

$$Var(Y) = E[Y^2] - (E[Y])^2. \tag{1.12}$$

Sustituyendo las medias teóricas por las medias muestrales determinadas por la ecuación (1.11) en las ecuaciones (1.10) y (1.12) se obtiene la estimación del coeficiente de correlación ρ , dado en la siguiente definición.

Definición 1.15

La medida de asociación entre dos variables X e Y se estima con el coeficiente de correlación muestral designado también por R , donde:

$$\hat{\rho} = R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

Podemos expresar a R en cantidades conocidas

$$R = \hat{\beta}_1 \left(\frac{S_{xx}}{S_{yy}} \right)^{1/2},$$

de modo que si despejamos a $\hat{\beta}_1$, ésta es igual al coeficiente de correlación muestral R multiplicado por la raíz cuadrada del cociente de valores de la dispersión de Y entre X . Notemos que, $\hat{\beta}_1$ y R están muy relacionados, aunque la información que proporcionan es diferente ya que el coeficiente de correlación muestral R mide la asociación lineal entre Y y X , mientras que $\hat{\beta}_1$ mide el cambio esperado en la media de Y por un cambio unitario en X .

La relación es más clara si elevamos al cuadrado a R .

$$R^2 = \hat{\beta}_1^2 \left(\frac{S_{xx}}{S_{yy}} \right) = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}},$$

en la expresión anterior tenemos que el numerador es la suma de los cuadrados de la regresión (SSR) y el denominador la suma total de cuadrados (SST), esto se resume en la siguiente definición.

Definición 1.16

El estadístico de bondad del ajuste en la regresión lineal múltiple representa una cantidad que se denota por R^2 , denominado coeficiente de determinación, dado por

$$R^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{SST - SSE}{SST} = \frac{SSR}{SST}.$$

Ahora sea, $R^2 = 1 - [SSE/SST]$ que se puede interpretar como la proporción de la variación total en las Y_i , explicada por las covariables $X_1, X_2, \dots, X_k$ en un modelo de regresión lineal múltiple o proporción de la varianza en Y explicada por la regresión.

Nótese que $0 \leq R^2 \leq 1$ y por definición se tiene que el coeficiente de determinación se le usa como una medida de bondad del ajuste de los datos a la recta, ya que a mayor R^2 menor es SSE , que nos indicará que el modelo de regresión está bien ajustado.

1.3.2. Características del coeficiente de correlación muestral (R)

Ahora visualizando a R como un índice de asociación entre X e Y debe cumplir las siguientes características:

- El coeficiente de correlación muestral R varía en el intervalo $[-1, 1]$.
- A medida que R se aproxime a uno o menos uno, habrá mayor evidencia de que en el modelo de regresión lineal simple la variable X contribuye a explicar a Y . Esto implica la SSE es cercana a cero. Si $R = 1$ o $R = -1$ los puntos observados caen exactamente sobre una recta, es decir, hay una correlación ya sea positiva o negativa perfecta entre X e Y , el signo depende de $\hat{\beta}_1$.
- Si R tiene un valor cercano a cero, no existe una relación lineal entre X e Y . Si esto ocurre es porque SSR tiene un valor pequeño en relación con la SST .

Así, el coeficiente de correlación muestral proporciona una medida cuantitativa de la dependencia entre dos variables.

Cuanto más cerca es $|R|$ de 1, cuanto más estrechamente relacionadas son las variables; si $|R| = 1$, entonces una variable se puede predecir exactamente de la otra. Cabe destacar que tanto R como R^2 deben usarse únicamente como índices descriptivos bajo las suposiciones del modelo propuesto.

Capítulo 2

Análisis de Componentes Principales (ACP)

¹ Un problema en el análisis de datos multivariantes es la reducción de la dimensionalidad, es decir, describir los valores de las k , variables que definen una observación por un pequeño subconjunto $p < k$ de ellas que representen adecuadamente la información con las p variables que son combinación lineal de las variables originales.

La técnica de análisis de componentes principales es debida a Hotelling (1933), aunque sus orígenes están en los ajustes ortogonales por mínimos cuadrados que fueron introducidos por K. Pearson (1901). Sin embargo, el ACP se popularizó hasta la aparición de ordenadores. La utilidad del ACP radica en dos formas:

- Admite la representación óptima en un espacio de dimensión pequeño p de observaciones de un espacio k -dimensional. Así, con el ACP se identifican las posibles variables latentes o no observables que generan los datos.
- Para facilitar la interpretación de los datos al hacer una transformación ortogonal de las k variables originales en p nuevas variables no correlacionadas entre sí, denominadas componentes principales.

Las componentes principales se obtienen en forma decreciente de importancia. Cuando su varianza es mayor, existe mayor información.

Las k componentes principales son combinación lineal de las p variables medibles y se espera que la variabilidad de algunas de estas componentes, en donde las más importantes expliquen la mayor parte de la variabilidad presente en los datos originales. Así, la dimensionalidad de los datos se ha reducido, es decir, hay una alta correlación entre alguna de las p variables que expresan lo mismo por lo que existen restricciones lineales sobre ellas.

¹Sección basada principalmente en [20], [25].

2.1. Planteamiento del problema

Si se disponen de los valores de k —variables en n elementos de una población en una matriz $\mathbf{X}$ de dimensión $n \times k$. Supondremos que se ha restado a cada variable su media, de tal forma que las variables de la matriz $\mathbf{X}$ tienen media cero y su matriz de covarianzas determinada por $1/n\mathbf{X}^t\mathbf{X}$.

Así, el problema a resolver es encontrar un espacio de dimensión reducida que represente adecuadamente los datos.

El problema puede abordarse desde tres perspectivas equivalentes, que serán discutidas a continuación.

2.1.1. Enfoque descriptivo

La idea intuitiva de este enfoque es encontrar un subespacio de dimensión menor que k tal que al proyectar sobre los puntos conserven su estructura con menos alteración.

Matemáticamente se tiene lo siguiente:

- Primero considere un subespacio de dimensión uno (una recta).
Se desea que las proyecciones de los puntos sobre esta recta mantengan lo más posible sus posiciones relativas.
- Ahora considere el caso de dos dimensiones ($k = 2$).
La condición de que la recta pase cerca de la mayoría de los puntos se logra con mantener las distancias entre los puntos originales y sus proyecciones sobre la recta, es decir, que sean lo más pequeñas posibles. Si se considera un punto X_i y una dirección $\mathbf{d}_1 = (d_{11}, \dots, d_{1k})^t$ (vector de norma uno), la proyección del punto $\mathbf{X}_i$ sobre esta dirección es el escalar:

$$Y_i = d_{11}X_{i1} + \dots + d_{1k}X_{ik} = \mathbf{d}_1^t \mathbf{X}_i$$

y el vector que representa ésta proyección es $Y_i \mathbf{d}_1$. Llamémosle r_i a la distancia entre el punto $\mathbf{X}_i$ y su proyección sobre la dirección $\mathbf{d}_1$, este criterio implica:

$$\text{minimizar} \left(\sum_{i=1}^n r_i^2 = \sum_{i=1}^n |\mathbf{X}_i - Y_i \mathbf{d}_1|^2 \right).$$

Como el primer término es constante, minimizar la suma de las distancias a la recta de todos los puntos $\sum_{i=1}^n r_i^2$ es equivalente a maximizar $\sum_{i=1}^n Y_i^2$, la suma al cuadrado de los valores de las proyecciones. Como las proyecciones Y_i son variables de media cero, maximizar la suma de sus cuadrados

equivale a maximizar su varianza, así, obtenemos el criterio de encontrar la dirección de proyección que maximice la varianza de los datos proyectados.

Es equivalente a buscar la dirección que pasa cerca de los puntos a buscar la dirección tal que los puntos proyectados sobre ella conserven lo más posible sus distancias relativas.

2.1.2. Enfoque estadístico

En este enfoque se quiere representar puntos k dimensionales con la mínima pérdida de información, para lo cual se tiene:

- En un espacio de dimensión uno es equivalente a sustituir las k variables originales por una nueva variable Y_1 que resuma óptimamente la información. Para esto, la nueva variable debe tener globalmente máxima correlación con las variables originales. Esto no será posible si la nueva variable toma el mismo valor en todos los elementos, así la condición para que se pueda predecir con la mínima pérdida de información en los datos observados es utilizar la variable de máxima variabilidad.
- Dimensión dos, con esta dimensión se obtiene el mejor subespacio de los datos. Para esto, se calcula el plano que mejor aproxima a los puntos, por lo cual el problema se reduce a encontrar una nueva dirección definida por un vector unitario d_2 , que sin pérdida de generalidad puede tomarse ortogonal a d_1 , con la condición de que la proyección de un punto sobre este eje maximice las distancias entre los puntos proyectados, esto equivale a encontrar una segunda variable Y_2 no correlacionada con la anterior, y que tenga varianza máxima. Note que la componente Y_p ($p < k$) tendrá varianza máxima entre todas las combinaciones lineales de las k variables originales con la condición de estar no correlacionada con las $Y_1, \dots, Y_{p-1}$ previamente obtenidas.

2.1.3. Enfoque geométrico

Analizando una vez más el problema desde un enfoque geométrico se tiene que:

- Si se observa la nube de puntos que se genera con los datos y si forman una elipse de concentración se pueden describir por su proyección en la dirección del eje mayor de la elipse de concentración. Puede demostrarse que éste eje es la recta que minimiza las distancias ortogonales con lo que volvemos al problema que ya se ha resuelto.

- En varias dimensiones se tendrían elipsoides de concentración, y la mejor aproximación a los datos es la proporcionada por su proyección sobre el eje mayor del elipsoide de concentración.

Intuitivamente la mejor aproximación en dos dimensiones es la proyección sobre el plano de los dos ejes mayores del elipsoide de concentración, así se pueden tomar a los ejes del elipsoide de concentración como nuevas variables originales que pasan de variables correlacionadas a variables ortogonales o no correlacionadas.

Para visualizar gráficamente el análisis de componentes principales, en la Figura 2.1 se presenta un esquema de una nube de puntos representando datos originales, en donde la primer componente principal corresponde al eje mayor de la elipse de concentración y la segunda componente principal está representada por el eje menor de la elipse de concentración.

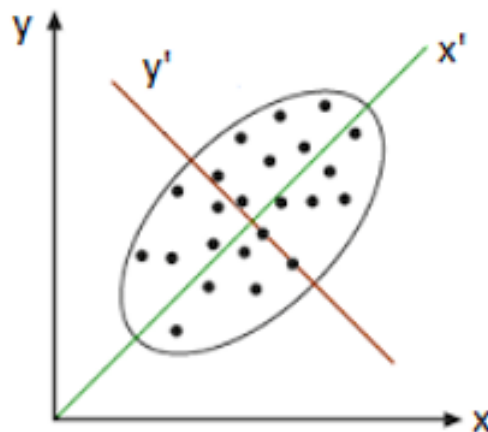


Figura 2.1: Elipse de concentración.

2.2. Cálculo de las componentes principales

En el ACP se pretende explicar la estructura de covarianza de un vector aleatorio buscando un nuevo sistema de ejes coordenados que indiquen las direcciones de mayor variabilidad.

Sean

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}$$

y

$$\Sigma = \mathbf{E}(\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^t = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{2k} & \sigma_{22}^2 & \dots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kk}^2 \end{pmatrix}.$$

En donde $\mathbf{m}$ es el vector promedio y $\mathbf{X}$ es el vector aleatorio con matriz de covarianza Σ . La matriz $\Sigma = \Sigma^t$ ya que es semidefinida positiva y por ser simétrica es autoadjunta por lo que sus eigenvalores son positivos y sus eigenvectores ortogonales.

En las siguientes secciones se explica la forma de calcular las componentes principales.

2.2.1. Cálculo de la primera componente principal

Se define a la primera componente principal como una combinación lineal de las variables originales que debe tener varianza máxima, es decir:

$$\mathbf{Y}_1 = d_{11}X_1 + d_{21}X_2 + \dots + d_{k1}X_k = \mathbf{d}_1^t \mathbf{X}.$$

Se obtiene un vector aleatorio $\mathbf{d}_1$ donde sus componentes son combinación lineal de los componentes del vector inicial, donde $\mathbf{d}_1 = (d_{11}, d_{12}, \dots, d_{1k})^t$ y se quiere elegir un $\mathbf{d}_1$ de modo que se maximice la varianza de $\mathbf{Y}_1$, es conveniente normalizar el vector $\mathbf{d}_1$, es decir, $\mathbf{d}_1^t \mathbf{d}_1 = 1$, así, la varianza para $\mathbf{Y}_1$ será:

$$Var(\mathbf{Y}_1) = Var(\mathbf{d}_1^t \mathbf{X}) = \mathbf{d}_1^t Var(\mathbf{X}) \mathbf{d}_1 = \mathbf{d}_1^t \Sigma \mathbf{d}_1. \quad (2.1)$$

Ahora bien, para maximizar (2.1) y que además tenga solución siendo este el problema a resolver, se hace uso del método de multiplicadores de Lagrange para maximizar $Var(\mathbf{Y}_1)$ sujeto a la restricción $\mathbf{d}_1^t \mathbf{d}_1 = 1$. Notemos que $\mathbf{d}_1$ es el vector desconocido que nos da la combinación lineal óptima. Así, la función a maximizar es

$$L_1(\mathbf{d}_1, \lambda_1) = \mathbf{d}_1^t \Sigma \mathbf{d}_1 - \lambda_1(\mathbf{d}_1^t \mathbf{d}_1 - 1), \quad (2.2)$$

derivando con respecto a $\mathbf{d}_1$ e igualando a cero se tiene:

$$\frac{\partial L_1(\mathbf{d}_1, \lambda_1)}{\partial \mathbf{d}_1} = 0 \quad , \text{ es decir,}$$

$$\frac{\partial L_1(\mathbf{d}_1, \lambda_1)}{\partial \mathbf{d}_1} = \frac{\partial}{\partial \mathbf{d}_1} \left[\mathbf{d}_1^t \sum \mathbf{d}_1 - \lambda_1 (\mathbf{d}_1^t \mathbf{d}_1 - 1) \right]$$

y si $\alpha = \mathbf{d}_1$, y empleando convención de suma sobre índices repetidos se tiene que:

$$\begin{aligned} \frac{\partial L_1(\alpha, \lambda_1)}{\partial \alpha_k} &= \frac{\partial}{\partial \alpha_k} \left[\alpha_i \left(\sum \right)_{ij} \alpha_j - \lambda_1 (\alpha_i \alpha_i - 1) \right] \\ &= \frac{\partial \alpha_i}{\partial \alpha_k} \left(\sum \right)_{ij} \alpha_j + \alpha_i \left(\sum \right)_{ij} \frac{\partial \alpha_j}{\partial \alpha_k} - \lambda_1 2 \alpha_i \frac{\partial \alpha_i}{\partial \alpha_k} \\ &= \delta_k^i \left(\sum \right)_{ij} \alpha_j + \alpha_i \left(\sum \right)_{ij} \delta_k^j - \lambda_1 2 \alpha_i \delta_k^i \quad (\text{usando deltas de Kronecker}) \\ &= \left(\sum \right)_{kj} \alpha_j + \alpha_i \left(\sum \right)_{ik} - 2 \lambda_1 \alpha_k \quad (\text{por ser simétrica}) \\ &= 2 \left(\sum \right)_{kj} \alpha_j - 2 \lambda_1 \alpha_k \end{aligned}$$

regresando a la notación de vector se tiene

$$\frac{\partial L_1(\mathbf{d}_1, \lambda_1)}{\partial \mathbf{d}_1} = 2 \sum \mathbf{d}_1 - 2 \lambda_1 \mathbf{Id}_1 = \mathbf{0}, \quad (2.3)$$

$$\frac{\partial L_1(\mathbf{d}_1, \lambda_1)}{\partial \lambda_1} = -(\mathbf{d}_1^t \mathbf{d}_1 - 1) = 0. \quad (2.4)$$

Por tanto, la solución de (2.3) no es más que un sistema lineal de ecuaciones.

$$2 \sum \mathbf{d}_1 - 2 \lambda_1 \mathbf{Id}_1 = \mathbf{0} \Rightarrow \left(\sum - \lambda_1 \mathbf{I} \right) \mathbf{d}_1 = \mathbf{0}.$$

Por el Teorema de Roché-Frobenius (Apéndice A1.3), para que el sistema tenga una solución distinta de cero, la matriz $(\sum - \lambda_1 \mathbf{I})$ tiene que ser singular. Esto es,

$$\left| \left(\sum - \lambda_1 \mathbf{I} \right) \right| = 0$$

en donde $\mathbf{d}_1$ es un eigenvector de la matriz $\sum$ y λ_1 su eigenvalor. Por otro lado, se tiene que

$$\left(\sum - \lambda_1 \mathbf{I} \right) \mathbf{d}_1 = \mathbf{0} \Leftrightarrow \sum \mathbf{d}_1 = \lambda_1 \mathbf{Id}_1 \quad (2.5)$$

y

$$Var(\mathbf{Y}_1) = Var(\mathbf{d}_1^t \mathbf{X}) = \mathbf{d}_1^t \sum \mathbf{d}_1 = \mathbf{d}_1^t \lambda_1 \mathbf{Id}_1 = \lambda_1 \mathbf{d}_1^t \mathbf{d}_1 = \lambda_1.$$

Por lo que, $Var(\mathbf{Y}_1) = \lambda_1$, como es la cantidad que se quiere maximizar, λ_1 será el mayor eigenvalor de la matriz $\sum$. El vector $\mathbf{d}_1$ define los coeficientes de cada variable en la primer componente principal.

2.2.2. Cálculo de la segunda componente principal

La segunda componente principal es una combinación lineal de las variables originales que tiene varianza máxima y no está correlacionada con la primera componente principal:

$$\mathbf{Y}_2 = d_{21}X_1 + \dots + d_{2k}X_k = \mathbf{d}_2^t \mathbf{X}.$$

Se obtiene $\mathbf{d}_2$ un vector aleatorio donde sus componentes son combinación lineal de los componentes del vector inicial, $\mathbf{d}_2 = (d_{21}, d_{22}, \dots, d_{2k})^t$, se elige $\mathbf{d}_2$ de modo de que se maximice la varianza de $\mathbf{Y}_2$, sujeto a $\mathbf{d}_2^t \mathbf{d}_2 = 1$ que está determinada por:

$$Var(\mathbf{Y}_2) = Var(\mathbf{d}_2^t \mathbf{X}) = \mathbf{d}_2^t Var(\mathbf{X}) \mathbf{d}_2 = \mathbf{d}_2^t \sum \mathbf{d}_2, \quad (2.6)$$

como se dijo antes que no deben estar correlacionada $\mathbf{Y}_2$ con $\mathbf{Y}_1$, es decir, $Cov(\mathbf{Y}_2, \mathbf{Y}_1) = 0$, se tiene que:

$$\begin{aligned} Cov(\mathbf{Y}_2, \mathbf{Y}_1) &= Cov(\mathbf{d}_2^t \mathbf{X}, \mathbf{d}_1^t \mathbf{X}) = \mathbf{d}_2^t Cov(\mathbf{X}, \mathbf{X}) \mathbf{d}_1 \\ &= \mathbf{d}_2^t Var(\mathbf{X}) \mathbf{d}_1 = \mathbf{d}_2^t \sum \mathbf{d}_1, \end{aligned}$$

se quiere que $\mathbf{d}_2^t \sum \mathbf{d}_1 = 0$, nótese que de la ecuación (2.5) se tiene que $\sum \mathbf{d}_1 = \lambda_1 \mathbf{d}_1$ que es la varianza de la primera componente principal que tiene la mayor variabilidad de los datos, como $\lambda_1 > 0$ tiene variabilidad positiva. Así

$$Cov(\mathbf{Y}_2, \mathbf{Y}_1) = \mathbf{d}_2^t \sum \mathbf{d}_1 = \mathbf{d}_2^t \lambda_1 \mathbf{d}_1 = \lambda_1 \mathbf{d}_2^t \mathbf{d}_1 = 0,$$

si y sólo si $\mathbf{d}_2^t \mathbf{d}_1 = 0$, es decir, que $\mathbf{d}_2$ y $\mathbf{d}_1$ son ortogonales. Luego las componentes principales $\mathbf{Y}_1$ y $\mathbf{Y}_2$ estarán no correlacionadas.

El problema es maximizar la ecuación (2.6), para esto se hará nuevamente uso de multiplicadores de Lagrange que incluye la restricción de que las direcciones debe tener modulo unitario, $\mathbf{d}_2^t \mathbf{d}_2 = 1$, así, la función objetivo a maximizar es

$$\begin{aligned} L_2(\mathbf{d}_1, \mathbf{d}_2, \lambda_1, \lambda_2) &= \mathbf{d}_1^t \sum \mathbf{d}_1 + \mathbf{d}_2^t \sum \mathbf{d}_2 - \lambda_1(\mathbf{d}_1^t \mathbf{d}_1 - 1) \\ &\quad - \lambda_2(\mathbf{d}_2^t \mathbf{d}_2 - 1) - \gamma(\mathbf{d}_2^t \mathbf{d}_1) \end{aligned} \quad (2.7)$$

se deriva con respecto a $\mathbf{d}_1$, $\mathbf{d}_2$, λ_1 y λ_2 e igualando a cero:

$$\frac{\partial L_2(\mathbf{d}_1, \mathbf{d}_2, \lambda_1, \lambda_2)}{\partial \mathbf{d}_1} = 2 \sum \mathbf{d}_1 - 2\lambda_1 \mathbf{Id}_1 = 0 \quad (2.8)$$

$$\frac{\partial L_2(\mathbf{d}_1, \mathbf{d}_2, \lambda_1, \lambda_2)}{\partial \mathbf{d}_2} = 2 \sum \mathbf{d}_2 - 2\lambda_2 \mathbf{d}_2 = 0 \quad (2.9)$$

$$\frac{\partial L_2(\mathbf{d}_1, \mathbf{d}_2, \lambda_1, \lambda_2)}{\partial \lambda_1} = -(\mathbf{d}_1^t \mathbf{d}_1 - 1) = 0 \quad (2.10)$$

$$\frac{\partial L_2(\mathbf{d}_1, \mathbf{d}_2, \lambda_1, \lambda_2)}{\partial \lambda_2} = -(\mathbf{d}_2^t \mathbf{d}_2 - 1) = 0. \quad (2.11)$$

Por tanto, la solución de las ecuaciones (2.8) y (2.9) no es más que un sistema lineal de ecuaciones. La primera dada en la ecuación (2.5) y la segunda se calcula de manera análoga, dada por la siguiente ecuación

$$2 \sum \mathbf{d}_2 - 2\lambda_2 \mathbf{Id}_2 = 0 \Rightarrow \left(\sum -\lambda_2 \mathbf{I} \right) \mathbf{d}_2 = 0 \Rightarrow \sum \mathbf{d}_2 = \lambda_2 \mathbf{Id}_2. \quad (2.12)$$

Así, se tiene que

$$Var(\mathbf{Y}_2) = Var(\mathbf{d}_2^t \mathbf{X}) = \mathbf{d}_2^t \sum \mathbf{d}_2 = \mathbf{d}_2^t \lambda_2 \mathbf{Id}_2 = \lambda_2 \mathbf{d}_2^t \mathbf{d}_2 = \lambda_2.$$

Por lo que, $Var(\mathbf{Y}_2) = \lambda_2$, como es la cantidad que se quiere maximizar, si λ_2 es el mayor eigenvalor de la matriz de covarianza ($\sum$). Si se toma a los eigenvectores de norma uno y sustituyendo en (2.7) se tiene que el máximo de esta función objetivo es $\lambda_1 + \lambda_2$ es claro que λ_1 y λ_2 son los dos eigenvalores mayores asociados a la matriz de covarianza ($\sum$), $\mathbf{d}_1$ y $\mathbf{d}_2$ sus correspondientes eigenvectores y cumplen que $\mathbf{d}_2^t \mathbf{d}_1 = 0$, (ver Apéndice A1.4). Así, la segunda componente principal está dada por $\mathbf{Y}_2 = \mathbf{X} \mathbf{d}_2$.

2.2.3. Cálculo de la j -ésima componente principal

La j -ésima componente principal es una combinación lineal de las variables originales que tiene varianza máxima y no está correlacionada con las $j - 1$ primeras componentes principales $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_{j-1}$, así se tiene que

$$\mathbf{Y}_j = d_{j1}X_1 + \dots + d_{jk}X_k = \mathbf{d}_j^t \mathbf{X}.$$

Se busca obtener el vector $\mathbf{d}_j$ definido por $\mathbf{d}_j = (d_{j1}, d_{j2}, \dots, d_{jk})^t$, y se debe elegir de modo que maximice la varianza de $\mathbf{Y}_j$, la cual se determina por:

$$Var(\mathbf{Y}_j) = Var(\mathbf{d}_j^t \mathbf{X}) = \mathbf{d}_j^t Var(\mathbf{X}) \mathbf{d}_j = \mathbf{d}_j^t \sum \mathbf{d}_j \quad (2.13)$$

como se dijo antes que no debe estar correlacionada $\mathbf{Y}_j$ con $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_{j-1}$ es decir, $Cov(\mathbf{Y}_j, \mathbf{Y}_i) = 0, i = 1, 2, \dots, j-1$. Por lo tanto

$$\begin{aligned} Cov(\mathbf{Y}_j, \mathbf{Y}_i) &= Cov(\mathbf{d}_j^t \mathbf{X}, \mathbf{d}_i^t \mathbf{X}) = \mathbf{d}_j^t Cov(\mathbf{X}, \mathbf{X}) \mathbf{d}_i \\ &= \mathbf{d}_j^t Var(\mathbf{X}) \mathbf{d}_i = \mathbf{d}_j^t \sum \mathbf{d}_i, \end{aligned}$$

se quiere que $\mathbf{d}_j^t \sum \mathbf{d}_i = 0$, note que se cumple $\sum \mathbf{d}_i = \lambda_i \mathbf{d}_i$ que es la varianza de la primera componente principal que tiene la mayor variabilidad de los datos, como $\lambda_i > 0$ tiene variabilidad positiva. Así,

$$Cov(\mathbf{Y}_j, \mathbf{Y}_i) = \mathbf{d}_j^t \sum \mathbf{d}_i = \mathbf{d}_j^t \lambda_i \mathbf{d}_i = \lambda_i \mathbf{d}_j^t \mathbf{d}_i = 0$$

si y sólo si $\mathbf{d}_j^t \mathbf{d}_i = 0$, es decir, que $\mathbf{d}_j$ y $\mathbf{d}_i$ son ortogonales para $i = 1, \dots, j-1$.

Luego las componentes principales $\mathbf{Y}_i$ y $\mathbf{Y}_j$ serán no correlacionadas.

El problema es maximizar la ecuación (2.13), para esto se hará nuevamente uso de la función Lagrangiana que incluye la restricción de que las direcciones debe tener módulo unitario, $\mathbf{d}_j^t \mathbf{d}_j = 1$, así, la función objetivo a maximizar es

$$\begin{aligned} L_j(\mathbf{d}_1, \dots, \mathbf{d}_j, \lambda_1, \dots, \lambda_j) &= \mathbf{d}_1^t \sum \mathbf{d}_1 + \mathbf{d}_2^t \sum \mathbf{d}_2 + \dots + \mathbf{d}_j^t \sum \mathbf{d}_j \\ &\quad - \lambda_1(\mathbf{d}_1^t \mathbf{d}_1 - 1) - \lambda_2(\mathbf{d}_2^t \mathbf{d}_2 - 1) - \dots - \lambda_j(\mathbf{d}_j^t \mathbf{d}_j - 1) - \gamma(\mathbf{d}_j^t \mathbf{d}_i), \end{aligned} \quad (2.14)$$

se deriva con respecto a $\mathbf{d}_1, \dots, \mathbf{d}_j, \lambda_1, \dots, \lambda_j$ e igualando a cero:

$$\frac{\partial L_j(\mathbf{d}_1, \dots, \mathbf{d}_j, \lambda_1, \dots, \lambda_j)}{\partial \mathbf{d}_j} = 2 \sum \mathbf{d}_j - 2\lambda_j \mathbf{I} \mathbf{d}_j = \bar{\mathbf{0}}. \quad (2.15)$$

$$\frac{\partial L_j(\mathbf{d}_1, \dots, \mathbf{d}_j, \lambda_1, \lambda_j)}{\partial \lambda_j} = -(\mathbf{d}_j^t \mathbf{d}_j - 1) = \bar{\mathbf{0}}. \quad (2.16)$$

Por tanto, la solución de las ecuaciones (2.15) no es más que un sistema lineal de ecuaciones, dado por:

$$2 \sum \mathbf{d}_j - 2\lambda_j \mathbf{I} \mathbf{d}_j = \bar{\mathbf{0}} \Rightarrow \left(\sum -\lambda_j \mathbf{I} \right) \mathbf{d}_j = \bar{\mathbf{0}} \Rightarrow \sum \mathbf{d}_j = \lambda_j \mathbf{I}. \quad (2.17)$$

Se tiene que

$$Var(\mathbf{Y}_j) = Var(\mathbf{d}_j^t \mathbf{X}) = \mathbf{d}_j^t \sum \mathbf{d}_j = \mathbf{d}_j^t \lambda_j \mathbf{I} \mathbf{d}_j = \lambda_j \mathbf{d}_j^t \mathbf{d}_j = \lambda_j.$$

Por lo que, $Var(\mathbf{Y}_j) = \lambda_j$, es la cantidad que se quiere maximizar, si λ_j es el mayor eigenvalor propio de la matriz $\sum$. Si se toma a los eigenvectores de norma uno y sustituyendo en (2.14) se tiene que el máximo de esta función objetivo es $\lambda_1 + \lambda_2 + \dots + \lambda_j$, es claro que $\lambda_1, \dots, \lambda_j$ son j eigenvalores mayores de la matriz $\sum$, $\mathbf{d}_1, \dots, \mathbf{d}_j$ son sus correspondientes eigenvectores, y cumplen que $\mathbf{d}_j^t \mathbf{d}_i = 0$, (ver Apéndice A1.4). Así, la j -ésima componente principal está dada por $\mathbf{Y}_j = \mathbf{X} \mathbf{d}_j$.

2.2.4. Generalización de las componentes principales

Generalizaremos ahora el cálculo de componentes principales a un espacio de dimensión p que mejor represente a los puntos. A este espacio, están asociados los p eigenvectores con mayores eigenvalores de Σ . Las direcciones que definen estos vectores se denominan direcciones principales de los datos.

Sean

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{bmatrix}, \mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1k} \\ d_{21} & d_{22} & \dots & d_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k2} & \dots & d_{kk} \end{pmatrix}, \mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}.$$

Como $\mathbf{X}$ tiene rango k , entonces hay tantas componentes principales como variables que se obtienen calculando los eigenvalores, $\lambda_1, \lambda_2, \dots, \lambda_k$ de la matriz de varianza y covarianza de las variables Σ , mediante $|\Sigma - \lambda I| = 0$, siendo sus vectores asociados $(\Sigma - \lambda_i I)\mathbf{d}_i = \mathbf{0}$. Esto es equivalente a calcular las varianzas:

$$Var(\mathbf{Y}_1) = \lambda_1,$$

$$Var(\mathbf{Y}_2) = \lambda_2,$$

$$\vdots$$

$$Var(\mathbf{Y}_k) = \lambda_k,$$

y la matriz de covarianza de $\mathbf{Y}$ es

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_k \end{pmatrix}.$$

Se tiene también que

$$\mathbf{\Lambda} = Var(\mathbf{Y}) = \mathbf{D}^t Var(\mathbf{X}) \mathbf{D} = \mathbf{D}^t \Sigma \mathbf{D}, \quad (2.18)$$

entonces $\Sigma = \mathbf{D} \mathbf{\Lambda} \mathbf{D}^t$ ya que $\mathbf{D}$ es una matriz ortogonal.

Por otra parte, las $\lambda_i \in \mathbb{R}$, puesto que Σ es simétrica y positiva, Σ es definida positiva, ahora bien:

- Si Σ es simétrica: Sean $r, s \in \{1, \dots, k\}$, $r \neq s$ entonces λ_r y λ_s son raíces distintas por lo tanto, sus vectores asociados son ortogonales.

- Si Σ es semidefinida positiva de rango p ; $p < k$, si $k - p$ variables fuesen combinación lineal de las demás, habría sólo p raíces positivas y las demás serían cero.

El vector de variables no correlacionadas $\mathbf{Y}$ que están relacionadas con las variables originales mediante:

$$\mathbf{Y} = \mathbf{X}\mathbf{D} \quad \text{donde} \quad \mathbf{D}^t\mathbf{D} = \mathbf{I}.$$

Se dice que una equivalencia al cálculo de las componentes principales es aplicar una transformación ortogonal $\mathbf{D}$ a las variables $\mathbf{X}$ para obtener $\mathbf{Y}$, variables nuevas no correlacionadas entre sí.

2.3. Propiedades de las componentes principales

A continuación se resumen las propiedades de las componentes principales:

- I) $E[\mathbf{Y}] = \bar{0}$.
- II) $Var(\mathbf{Y}) = \Lambda$.
- III) $Cov(\mathbf{Y}_i, \mathbf{Y}_j) = 0, i \neq j$, es decir, las Y_i son no correlacionadas.
- IV) $Var(\mathbf{Y}_1) \geq Var(\mathbf{Y}_2) \geq \dots \geq Var(\mathbf{Y}_k) \geq 0$.
- V) $\sum_{i=1}^k Var(\mathbf{Y}_i) = Tr(\Sigma)$.
- VI) $\prod_{i=1}^k Var(\mathbf{Y}_i) = |\Sigma|$.

Se omitirán las demostraciones de I), II), III) y IV), ya que se han hecho de manera explícita en secciones anteriores. Procederemos con las demostraciones de las últimas dos propiedades.

Sabemos por la ecuación (2.18) que $\Lambda = \mathbf{D}\Sigma\mathbf{D}^t$ entonces $\Sigma = \mathbf{D}\Lambda\mathbf{D}^t$, así se tiene que:

$$\text{v) } Tr(\Sigma) = Tr(\mathbf{D}\Lambda\mathbf{D}^t) = Tr(\Lambda) = \sum_{i=1}^k \lambda_i = \sum_{i=1}^k Var(\mathbf{Y}_i).$$

$$\text{vi) } |\Sigma| = |\mathbf{D}\Lambda\mathbf{D}^t| = |\mathbf{D}||\Lambda\mathbf{D}^t| = |\Lambda\mathbf{D}^t||\mathbf{D}| = |\Lambda| = \prod_{i=1}^k \lambda_i = \prod_{i=1}^k Var(\mathbf{Y}_i). \square$$

2.4. Interpretación de las componentes principales

El análisis de componentes principales tiene dos principales objetivos

1. Identificación y significado, de las nuevas variables denominadas **componentes principales**.
2. Reducción de la dimensionalidad de los datos originales.

Lo que se analiza al calcular los eigenvalores de la matriz de varianzas (o correlaciones) es si las primeras componentes aportan mayor variabilidad en los datos. Es decidir qué eigenvalores son grandes y cuáles son pequeños, de manera que las componentes principales correspondientes a los eigenvalores pequeños no se toman en cuenta por la escasa información que aportan.

El ACP se puede interpretar cómo elegir nuevos ejes coordenados, que coincidan con los ejes principales de los datos.

El análisis de componentes principales evalúa la dimensión efectiva de un conjunto de datos.

Capítulo 3

Aedes Aegypti y su Relación con el Dengue, Chikungunya y Zika

El mosquito de la especie *Aedes Aegypti* es el responsable de transmitir tres enfermedades que amenazan actualmente, zika, chikungunya y dengue.

Este mosquito es de origen africano. En 1827 fue reportado por primera vez en México y a principios del siglo XXI tuvo un mayor auge [12]. Sus hábitos son antropófilos y domésticos, con radicación de criaderos en la vivienda humana y en sus alrededores.

3.1. Ciclo biológico del *Aedes Aegypti*

Son insectos de metamorfosis completa. Durante su desarrollo ontogénico pasan por cuatro etapas.

- **Huevo:** La hembra del mosquito *Aedes Aegypti* ovipone de 100 a 200 huevos por lote. Estas pueden producir hasta cinco lotes de huevos durante toda su vida. El número de huevos depende del tamaño de la porción de sangre succionada, ya que la sangre es necesaria como proteína para el desarrollo de los huevos. Si el mosquito es interrumpido antes de que complete su alimentación (2 o 3 mg de sangre) puede alimentarse con sangre más de una vez entre cada postura [19]. Los huevos son puestos en superficies húmedas y lugares donde el agua se acumula o se almacena ya sea de lluvia o no. Los huevos eclosionan en el momento de contacto con el agua. Son puestos por separado y en distinto momento, es decir, pueden ser distribuidos en horas o días dependiendo de la disponibilidad de sustratos adecuados. Además, son colocados a diferentes distancias por encima de la línea de agua. Según la Organización Mundial de la Salud (OMS) los huevos miden aproxima-

damente un milímetro de longitud. En el momento de postura son blancos, pero adquieren un color negro brillante.

Son fecundados durante la postura y el desarrollo embrionario hasta su eclosión se completa en un lapso de dos a tres días si el ambiente es húmedo y cálido. El desarrollo puede prolongarse hasta cinco días con temperaturas bajas. Con posterioridad a ese periodo, los huevos son capaces de resistir desecación y temperaturas extremas sobreviviendo de siete meses a un año. Esto facilita su traslado a otros lugares por medio de recipientes secos [11].

- **Larva:** Las larvas que emergen inician un ciclo de cuatro estadios larvales alcanzando seis o siete milímetros de longitud final, son exclusivamente acuáticas. Esta fase es el período de mayor crecimiento. El cuerpo de la larva está comprendido por la cabeza, el tórax ovoides y el abdomen de nueve segmentos [11]. Se alimentan de material orgánico acumulado en las paredes y el fondo del recipiente. La posición de reposo en el agua es casi vertical, su desplazamiento acuático es serpenteante y son sensibles a la luz. La duración del desarrollo larval depende de la temperatura, la disponibilidad de alimentos y la densidad de larvas en el recipiente. En condiciones óptimas con temperaturas de 25°C a 29°C el período desde la eclosión hasta la pupación puede ser de cinco a siete días [11].

Los tres primeros estados larvales se desarrollan rápidamente, mientras que el cuarto demora más tiempo con mayor aumento de tamaño y peso, (a baja temperatura y escasez de alimento, el cuarto periodo larval puede prolongarse hasta siete meses previo a su transformación en pupa). Son incapaces de resistir temperaturas inferiores a 10°C, superiores a 45°C, ya que a menos de 13°C se impide su paso a pupa.

- **Pupa:** La pupa no requiere alimentación ya que se encuentra en un estado de reposo donde se producen modificaciones anatómico-fisiológicas, además de que reaccionan inmediatamente a estímulos externos y se desplazan activamente por todo el recipiente. Se mantienen en la superficie del agua debido a su flotabilidad, esto último facilita la emersión del insecto adulto. El período pupal dura de uno a tres días en condiciones favorables y con temperaturas entre 28°C y 32°C . La pupa tiene en la base del tórax un par de tubos respiratorios que atraviesan la superficie del agua permitiéndoles la respiración, en la base del abdomen poseen un par de aletas.
- **Adulto:** Una vez que emerge de la pupa, el insecto adulto permanece en reposo para permitir el endurecimiento del exoesqueleto y alas, el insecto es un pequeño volador de color negro, con manchitas blancas en el dorso del

tórax, mostrando un anillado característico a nivel de tarsos, tibia y fémures de las patas. El mosquito adulto *Aedes Aegypti* se muestra en la Figura 3.1.



Figura 3.1: *Aedes Aegypti*.

Después de las 24 horas de su emergencia el insecto puede iniciar su proceso de reproducción [13].

El ciclo completo del *Aedes Aegypti*, en óptimas condiciones de temperatura y alimentación se completa, en 10 días [19].

La longevidad de los mosquitos adultos depende de la temperatura, precipitación y alimentación, por lo que aproximadamente a 10°C con precipitación y huéspedes para alimentarse viven alrededor de 30 días. A 23°C con precipitación y sin alimento viven aproximadamente cuatro días. La hembra sobrevive más tiempo que el macho ya que es más resistente a los cambios de temperatura y precipitación.

Las hembras hematófagas poseen una probóscis larga, adaptada para succionar sangre a través de la piel. Son las únicas que succionan sangre. Se dice que las horas más comunes para alimentarse son en la mañana entre las 6:00 a 8:00 y en las tardes de 16:00 a 19:00 horas [17]. Por lo que sus hábitos de alimentación son diurnos. Los machos presentan una probóscis larga adaptada para succionar néctar de flores y otras fuentes de azúcar.

Se sabe que el vuelo de *Aedes Aegypti* en especial de una hembra es contrario al viento, no sobrepasa los 100 m de radio durante su vida. Por lo general permanece en la misma casa o lugar donde emergió siempre que disponga de huéspedes y sitios adecuados para su establecimiento. Los mosquitos son atraídos por los olores y gases emitidos por el huésped (dióxido de carbono y ácido láctico principalmente).

Además, es raro, pero han comprobado que una hembra grávida puede volar hasta tres kilómetros en busca de un lugar para poner sus huevos si no encuentra cerca sitios apropiados. Mientras que los machos se dispersan menos que las hembras.

3.2. Aedes Aegypti como vector de virus

Las hembras del mosquito Aedes Aegypti, por ser estas quienes se alimentan de sangre, al picar infectan transmitiendo así los virus a los seres humanos, que a su vez se contaminan al succionar la sangre de quienes tienen el virus.

“El virus infecta el intestino medio del mosquito y luego se extiende hasta las glándulas salivales en un período de entre 8 y 12 días. Tras este período de incubación, el mosquito puede transmitir el virus a las personas al picarlas con fines exploratorios o alimentarios”, OMS.

Al momento de la picadura, los mosquitos inyectan su saliva, la cual puede contener tres tipos de enfermedades: zika, chikungunya y dengue. La enfermedad que transmitan dependerá si el mosquito estaba o no infectado por alguno de estos virus. Si el mosquito no está infectado por ninguno de los virus no habrá transmisión.

Las tres enfermedades que se transmiten a través de la picadura del mosquito Aedes Aegypti que hasta el momento se han detectado se describen en las siguientes secciones.

3.3. Zika

El virus de zika se detectó por primera vez en 1947 en los bosques de zika (por lo cual el virus lleva su nombre) en Uganda país situado en África Oriental. Se descubrió en un mono Rhesus cuando se realizaba un estudio acerca de la transmisión de la fiebre amarilla en la selva [23].

La transmisión ha sido reportada en muchos países y territorios. En México y para ser específicos en los estados de Chiapas y Nuevo León, se confirmaron en noviembre de 2015, dos primeros casos autóctonos del virus del zika, según la OMS [24].

El primero de febrero de 2016, la OMS declaró el virus zika una Emergencia de Salud Pública de Importancia Internacional (ESPII). Como se ha mencionado en secciones anteriores, este virus es transmitido a través de la picadura del mosquito Aedes Aegypti infectado. Es probable que el virus zika continúe extendiéndose a nuevas áreas.

Gran parte de las personas infectadas con el virus zika, no saben que tienen la enfermedad ya que no presentan síntomas, aunque los síntomas comunes son: fiebre, erupción cutánea, dolor en las articulaciones, inflamación que suele concentrarse en manos y pies, conjuntivitis e incluso pueden presentar dolor muscular y dolor de cabeza. No obstante la enfermedad suele ser sutil, los síntomas pueden persistir varios días incluso una semana después de haber sido picado por un mosquito infectado.

Por ser una enfermedad descubierta recientemente no se conoce con exactitud el período de incubación (tiempo desde la presencia de los síntomas) pero se pronostica que es de 3 a 12 días.

Las personas al ignorar la enfermedad o al no presentar los síntomas no acuden al hospital y por consiguiente no se dan cuenta que han sido infectadas por este virus. Además, se dice que es inusual que mueran de zika. El virus por lo general permanece en la sangre de una persona infectada alrededor de una semana. Una vez que una persona ha sido infectada debe evitar las picaduras de mosquito durante la primera semana para evitar futuras infecciones.

Por otra parte, la infección por el virus durante el embarazo es grave ya que causa Microcefalia, así como otros defectos cerebrales fetales [4]. La transmisión sexual del virus zika de un hombre a su pareja es descartada ya que la duración y el patrón de eliminación del virus en el sistema genitourinario o urogenital masculino es limitado. Sin embargo, el uso sistemático y correcto del preservativo y la abstinencia son métodos eficaces para prevenir la transmisión sexual del virus.

3.4. Chikungunya

El virus chikungunya se exhibió por primera vez en 1952, durante un brote ocurrido en el sur de Tanzania [9], país situado en la costa este de África Central. Se trata de un virus ARN de la familia Flaviviridae, “chikungunya”, es una voz del idioma Kimakonde que significa “doblar de dolor”, ya que los infectados presentan un aspecto encorvado debido a los dolores articulares.

El chikungunya también conocida como fiebre chikungunya es una enfermedad viral transmitida al ser humano a través de la picadura del mosquito *Aedes Aegypti* infectado por dicho virus.

El primer caso autóctono de chikungunya en México informado por la Secretaría de Salud fue en el 2014 en Chiapas, así también, reportó en agosto de 2015, 3306 casos de chikungunya distribuidos en 16 estados del país (Campeche, Chiapas, Coahuila, Colima, Estado de México, Guerrero, Jalisco, Michoacán, Morelos, Nayarit, Oaxaca, Quintana Roo, Tabasco, Tamaulipas, Veracruz y Yucatán). Para el nueve de mayo del presente año (2016) reportó que los contagios incrementaron un 17.9 % en una semana [2]. La Situación es similar para otros países de la región afectados por el brote de este virus.

El período de incubación del virus chikungunya es de tres a siete días (aunque el intervalo puede oscilar entre 2 y 12 días). Durante este periodo la persona infectada debe evitar las picaduras del mosquito ya que en este tiempo el virus chikungunya se puede encontrar en la sangre. Así evitar la transmisión de una persona infectada a un mosquito.

La mayoría de las personas infectadas por el virus de chikungunya después de

la picadura del mosquito infectado presentan los siguientes síntomas: fiebre muy alta, picazón, hinchazón de las articulaciones, náuseas, erupciones cutáneas, cansancio, dolor en las articulaciones, cabeza y muscular [8]. En algunos casos se presentan con complicaciones oculares, neurológicas, cardíacas, molestias gastrointestinales y que los dolores articulares pueden durar varios meses, o incluso años. Se dice que la mayor parte de los pacientes se sienten mejor al transcurrir la primer semana.

La OMS manifiesta que gran parte de los pacientes se recuperan completamente, éstas son personas que ya no se vuelven a infectar y no propagarán la enfermedad. Hasta la fecha el virus chikungunya no se ha encontrado en bebés infectados a través de la lactancia, por ser ésta benéfica, se recomienda a las madres a amamantar incluso en zonas donde el virus chikungunya está circulando. Por otra parte, no existen informes de que el virus se propague por medio de una transfusión de sangre.

Para el virus chikungunya al igual que para el virus zika, aún no hay vacuna o medicamentos para prevenir y tratar ninguno de estos dos virus. Sin embargo, para tolerar los síntomas se debe descansar bien, beber líquidos para evitar deshidratarse, tomar medicamentos como el acetaminofeno o paracetamol para reducir la fiebre y el dolor (bajo supervisión médica).

Actualmente se carece de información sobre si el riesgo y gravedad de infección por los virus transmitidos por el *Aedes Aegypti* es diferente para las personas con infección por VIH. Sin embargo, un estudio realizado en Tanzania reveló que las personas con infección por el VIH no tienen ningún riesgo mayor que la población en general para la infección con otros virus que son transmitidos por este mosquito [4].

3.5. Dengue

Las primeras epidemias de dengue reportadas datan de 1770 a 1780 casi simultáneamente en tres continentes Asia, África y América del Norte [6].

Según la OMS, en México en 1941 se reportaron los primeros casos de dengue y se notificaron 6,955 casos en toda la República. Se extinguió en 1963 gracias a la campaña de eliminación del *Aedes Aegypti* que mantuvo al dengue ausente durante 12 años. Sin embargo, en 1978 se reintrodujo en México. Hubo un aumento de casos a partir de 1980 por lo cual el dengue se convirtió en uno de los problemas actuales más importantes de salud pública a nivel nacional debido a su amplia distribución geográfica y su alto grado de susceptibilidad.

El virus del dengue es del tipo ARN y pertenece a la familia Flaviviridae. Este virus está conformado por cuatro serotipos: DEN1, DEN2, DEN3 y DEN4, cada uno corresponde a una población de cepas agrupadas en varios genotipos. A la

enfermedad de dengue se le conoce como la “fiebre rompe huesos”.

El período de incubación del virus del dengue es de cuatro a diez días. Los síntomas que presentan las personas infectadas por este virus son: un cuadro febril agudo de tres a cinco días de duración, dolores articulares, musculares, náuseas, vómitos, diarrea, tos, erupciones cutáneas, también padecen inflamación de los ganglios linfáticos, cefaleas o dolor muy intenso detrás de los globos oculares, y alteraciones gastrointestinales. Ocasionalmente se presentan diferentes manifestaciones hemorrágicas espontáneas como epistaxis o gingivorragias e hipotensión [13]. Si la persona infectada llega a presentar estos últimos síntomas (manifestaciones hemorrágicas) es porque el virus del dengue ha evolucionado a formas más severas como dengue hemorrágico y síndrome de choque de dengue. Las personas infectadas después e incluso durante el padecimiento de los síntomas mencionados poseen una convalecencia duradera, con fatiga y depresión persistentes. La fiebre hemorrágica del dengue se presenta con mayor frecuencia en menores de 15 años y en adultos.

La persona infectada debe aislarse durante el tiempo de incubación para evitar el contagio a mosquitos y con ello evitar que se propague la enfermedad en humanos.

La Secretaría de Salud, a través de la Comisión Federal para la Protección contra Riesgos Sanitarios (COFEPRIS), emitió la primer vacuna para el virus del dengue. Previniendo hospitalizaciones, muertes y generar ahorros económicos por la reducción en costos asociados a la atención médica [7]. La vacuna fue elaborada por el grupo farmacéutico francés Sanofi, el nueve de diciembre del 2015, en México fue aprobada dicha vacuna y hay aprobación comercial en 20 países más.

Es una vacuna quimérica tetravalente (para serotipos DEN1, DEN2, DEN3 y DEN4) elaborada a partir de virus vivos atenuados contra Dengue. Esta vacuna es para la población previamente expuesta al virus del dengue de entre 9 y 45 años de edad. Será aplicada en poblaciones en las que el Dengue sea endémico y su seroprevalencia sea igual o mayor a 60 %. Con una eficacia promedio del 60.5 % para la prevención de dengue y de un 93.2 % para la prevención de dengue hemorrágico.

En general, la transmisión ocurre por la picadura de la hembra del mosquito *Aedes Aegypti*. Se recomienda el uso de mosquiteros, instalación de mallas anti-mosquitos si no se tienen. Utilizar repelentes con icaridina y ropa que cubra las extremidades. Esto para reducir la posibilidad de sufrir una picadura. Otra medida de control suele ser aplicar insecticidas, así como utilizar productos para tratar el agua que almacenan las personas a fin de matar las larvas.

3.6. Importancia del control y factores que influyen en la población del Aedes Aegypti

Se expuso en secciones anteriores que la hembra del mosquito *Aedes Aegypti* es el vector principal que transmite los virus de zika, chikungunya y dengue. De aquí la importancia de controlar la población del mosquito. El riesgo para el país se encuentra en el potencial de transmisión de la enfermedad, que radica en los mosquitos transmisores del virus que viven en la región y en la densidad poblacional de la misma. Al no tener un control de la población del mosquito y al generarse un brote se puede expandir con facilidad y con ello originar una epidemia. Generando muertes y propagando la enfermedad a más países.

Los factores que influyen en el incremento de la población del mosquito y en consecuencia en el desarrollo de epidemias son la lluvia, temperaturas óptimas para su desarrollo y reproducción. En ausencia de lluvia influyen los contenedores para almacenar agua ya que muchas localidades no cuentan con un servicio de agua potable entubada y en caso de contar con el servicio es deficiente. Así como servicio de drenaje y alcantarillado ya que en zonas como éstas suelen ser sitios de reposo. Como se mencionó antes, para que complete el ciclo de vida del mosquito las hembras tienen que alimentarse de sangre, antes de esto busca el sitio donde depositará los huevos. Estos sitios suelen ser habitaciones humanas que son lugares óptimos ya que son oscuros y protegidos. Suelen esconderse bajo de las camas, bajo los muebles o en el peridomicilio en áreas con vegetación como macetas, jardines interiores, o bien en depósitos de aguas artificiales como bebederos de animales domésticos, floreros, canaletas de techos, botes, llantas, en general residuos sólidos que pueden almacenar agua y posteriormente puedan servir de criaderos.

Capítulo 4

Análisis de Efectos Ambientales y Sociales en la Incidencia de Dengue en México

En este capítulo se aplicarán las técnicas estadísticas analizadas en los primeros dos capítulos para el estudio del virus de dengue transmitido por el mosquito *Aedes Aegypti*. Como se mencionó en el capítulo tres, el *Aedes Aegypti* es un vector importante para la transmisión del virus. Por lo cual se realiza el estudio para la incidencia de dengue severo, es decir, fiebre de dengue sin presencia de hemorragia.

El objetivo de dicho análisis es determinar si existe un modelo de regresión lineal que explique de manera apropiada la relación entre la incidencia de dengue en México como variable dependiente y variables independientes los factores sociales y ambientales. Como un caso particular se determinará un modelo de regresión para el estado de Baja California Sur con datos ambientales del año 2014.

Con ayuda del software Python se realizaron los cálculos y gráficas de manera cautelosa.

4.1. Datos de estudio

Para este estudio se usaron datos recopilados y filtrados de incidencia de dengue reportados por el Sistema Nacional de Vigilancia Epidemiológica (SINAVE), la Dirección General de Epidemiología (DGE) [3], datos ambientales reportados por el Servicio Meteorológico Nacional (SMN), la Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT), así como datos sociales reportados por el Instituto Nacional de Estadística y Geografía (INEGI) [1], la Secretaría de Desarrollo Social (SEDESOL) [5] y la Comisión Nacional del Agua (CONAGUA).

Los datos recopilados de incidencia de dengue y factores ambientales corresponden al año 2014 y 2011, los datos recopilados para factores sociales corresponden sólo al año 2011. Los datos sociales son los más actuales hasta el momento, debido a que el censo de población que realiza INEGI se lleva a cabo cada diez años.

4.2. Efectos ambientales en la incidencia de dengue en Baja California Sur (BCS)

El caso particular que se analizó es la incidencia de dengue en el estado de Baja California Sur (BCS), con datos ambientales del año 2014. Ya que, en el año 2014 en BCS se reportaron datos claros debido a una epidemia de dengue. Las variables ambientales independientes que se analizaron son temperatura mínima y precipitación. Durante el año en BCS los gradientes de estas dos variables independientes son grandes, por lo cual, se comparó el efecto de cada una de estas variables independientes, para determinar si se correlaciona el gradiente de estas variables ambientales para estimar MRLS para predecir incidencia de dengue en BCS.

De acuerdo a las suposiciones del capítulo uno, se supuso que el error ϵ del modelo dado en la ecuación $Y = \beta_0 + \beta_1 X + \epsilon$ es que $E[\epsilon] = 0$ y que $V[\epsilon] = \epsilon^2$, independiente de X . Así, las distribuciones de $\hat{\beta}_0$ y $\hat{\beta}_1$ dependen de la distribución del error (ϵ) y si se supone que ϵ está distribuido normalmente con media cero y varianza σ^2 , se deduce que Y_i está distribuida normalmente con media $\beta_0 + \beta_1 X_i$ y varianza σ^2 . Como $\hat{\beta}_0$ y $\hat{\beta}_1$ son funciones lineales de $Y_1, Y_2, \dots, Y_n$, los estimadores están distribuidos normalmente. Es importante señalar que la suposición de normalidad de la distribución del error ϵ y por lo tanto las distribuciones normales resultantes para $\hat{\beta}_0$ y $\hat{\beta}_1$, no son necesarias para el ajuste de mínimos cuadrados. Sin embargo, esta última suposición nos permite desarrollar pruebas de hipótesis basados en la distribución t. Los resultados a utilizar se resumen en el siguiente Teorema.

Teorema 4.1

Propiedades de los estimadores de mínimos cuadrados.

- Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ son insesgados, es decir, $E[\hat{\beta}_i] = \beta_i$, para $i = 0, 1$.
- $V[\hat{\beta}_0] = C_{00}\sigma^2$, en donde $C_{00} = \sum X_i^2 / (nS_{xx})$.
- $V[\hat{\beta}_1] = C_{11}\sigma^2$, en donde $C_{11} = 1/S_{xx}$.
- $Cov(\hat{\beta}_0, \hat{\beta}_1) = C_{01}\sigma^2$, en donde $C_{01} = -\bar{X}/S_{xx}$.

- Un estimador insesgado de σ^2 es $S^2 = SSE/(n-2)$, en donde $SSE = S_{yy} - \hat{\beta}_1 S_{xy}$ y $S_{yy} = (Y_i - \bar{Y})^2$.

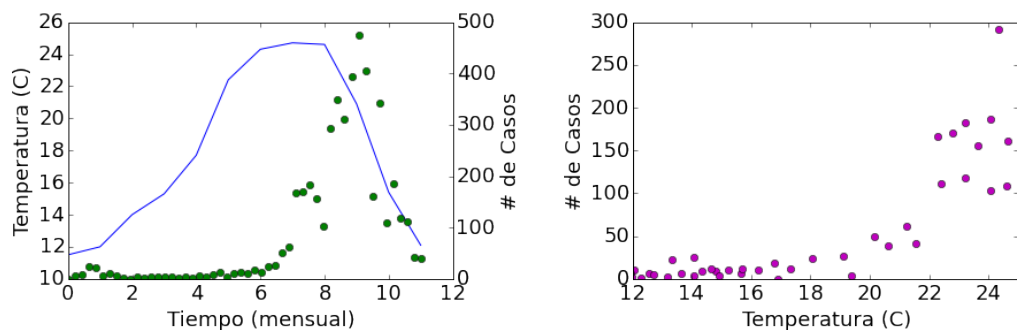
Si, además, el ϵ_i , para $i = 1, 2, \dots, n$ está distribuido normalmente:

- $\hat{\beta}_0$ y $\hat{\beta}_1$ están distribuidas normalmente.
- La variable aleatoria $\frac{(n-2)S^2}{\sigma^2}$ tiene una distribución χ^2 con $n-2$ grados de libertad.
- El estadístico S^2 es independiente de $\hat{\beta}_0$ y $\hat{\beta}_1$.

4.2.1. Efecto de temperatura en la incidencia de dengue en BCS

Se obtuvo un modelo de regresión lineal simple (MRLS) por medio del ajuste de un polinomio para la serie de tiempo de incidencia de dengue en BCS como variable dependiente y temperatura mínima semanal como variable independiente, ya que la temperatura es un factor importante para la reproducción del mosquito *Aedes Aegypti* transmisor del virus del dengue.

En la Figura 4.1.a se muestran los datos de casos de dengue representados por puntos y la línea sólida nos indica las temperaturas registradas mensualmente. En la Figura 4.1.b se muestra el diagrama de dispersión de los datos de temperatura con incidencia de dengue semanal del año 2014. Sobre este conjunto de datos se trabajó para la obtención del modelo de regresión para el estado de BCS. (Ver código en Apéndice B1.1).



(a) Puntos, casos de dengue registrados en 2014. Línea sólida, temperatura mínima semanal. (b) Diagrama de dispersión de temperatura e incidencia para el año 2014.

Figura 4.1: Datos para Baja California Sur de temperatura e incidencia.

Se ajustó un polinomio de grado dos al conjunto de puntos representados en la Figura 4.1.b por medio de regresión lineal simple. Es decir, se propuso un modelo de la forma:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2. \quad (4.1)$$

En donde $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ son coeficientes de regresión que se estiman por el método de mínimos cuadrados y se trata de minimizar la suma de cuadrados entre la ecuación y los puntos de los datos.

Con el método de mínimos cuadrados se obtienen los mejores estimadores insesgados para los coeficientes $\beta_i, i = 0, 1, 2$ los cuales son:

$$\hat{\beta}_0 = 741.195, \hat{\beta}_1 = -99.679, \hat{\beta}_2 = 3.277.$$

Así, el modelo de regresión lineal simple (MRLS) de segundo orden está dado por:

$$\hat{Y} = 741.195 - 99.679X + 3.277X^2. \quad (4.2)$$

El modelo dado por la ecuación (4.2) es representado por la línea azul como se muestra en la figura (4.2.a). (Ver código de figura en apéndice B1.2).

Para obtener un MRLS que ajuste mejor a los datos se hizo como función de prueba un polinomio de grado tres:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \hat{\beta}_3 X^3. \quad (4.3)$$

Al realizar los cálculos se obtienen los siguientes estimadores para este MRLS:

$$\begin{aligned} \hat{\beta}_0 &= -1.902 \times 10^3, \hat{\beta}_1 = 3.729 \times 10^2, \\ \hat{\beta}_2 &= -2.386 \times 10^1, \hat{\beta}_3 = 5.013 \times 10^{-1}, \end{aligned}$$

de tal modo que el MRLS para un polinomio de grado tres está dado por:

$$\begin{aligned} \hat{Y} &= -1.902 \times 10^3 + 3.729 \times 10^2 X - 2.386 \times 10^1 X^2 \\ &\quad + 5.013 \times 10^{-1} X^3. \end{aligned} \quad (4.4)$$

A pesar de que la ecuación (4.4) ajusta mejor a los datos de incidencia influenciada por la temperatura mínima, hay que tomar en cuenta que tiene un parámetro más. Este MRLS se representa gráficamente en la Figura 4.2.b por la línea sólida la cual se superpone a los datos de correlación de incidencia y temperatura representados por los puntos.

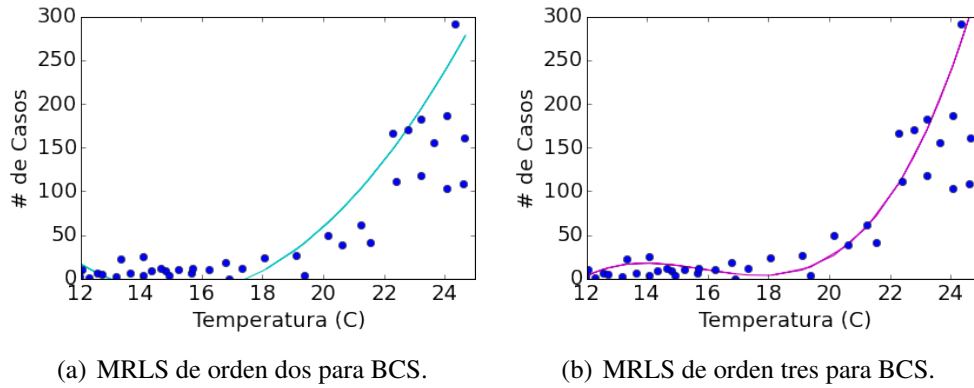


Figura 4.2: Modelos de regresión lineal simple para Baja California Sur (temperatura).

En la Figura 4.3.a se presenta un diagrama que muestra los errores para cada dato ϵ_n contra el error del siguiente dato ϵ_{n+1} . Estos errores corresponden al MRLS de segundo grado encontrado, dado por la ecuación (4.2). Este tipo de gráficas se utilizan para mostrar si existe correlación entre errores consecutivos de la serie de tiempo.

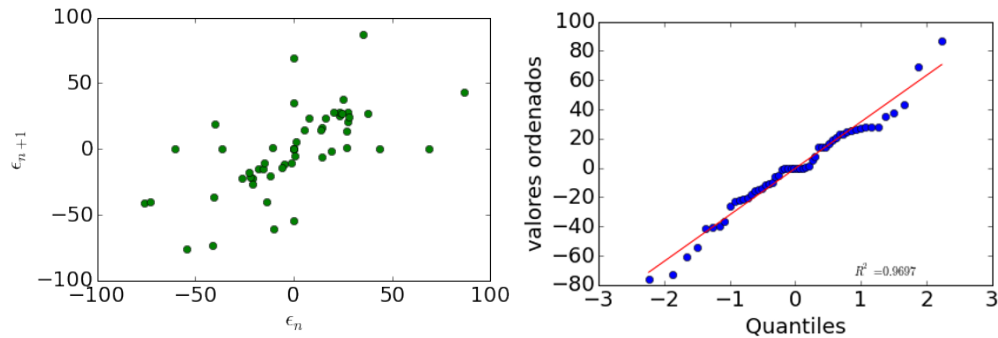
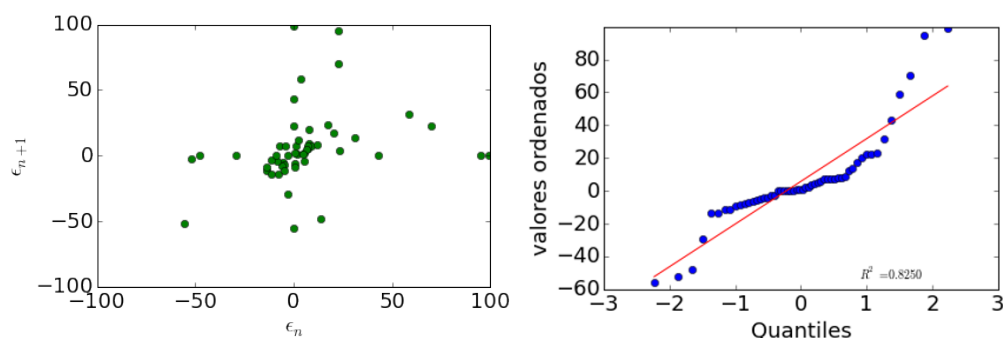


Figura 4.3: Supuestos de normalidad, temperatura segundo grado.

La condición necesaria y que cumplen los datos es que los errores tienen esperanza cero, es decir, $E(\epsilon) = 0$. Además, si los errores están distribuidos normalmente los estimadores de máxima verosimilitud coinciden con los estimadores de mínimos cuadrados ordinarios. Se hizo una gráfica de probabilidad normal (o gráfica cuantilar normal), donde se grafican los puntos (X, Y) en donde cada valor X proviene del conjunto original de datos muestrales, Y son los valores de la puntuación normal correspondiente, es decir es un valor cuantilar esperado de

la distribución normal estándar. Esto se muestra en la Figura 4.3.b, en donde se observa que los puntos se acercan a la línea recta de color rojo, por lo cual los datos tienen una distribución normal.

Nuevamente para mostrar independencia de las mediciones se muestra en la Figura 4.4.a un diagrama $\epsilon_n - \epsilon_{n+1}$ para el MRLS dado por el polinomio de tercer grado correspondiente a la ecuación (4.4).



(a) Diagrama de independencia de las mediciones. Se grafica ϵ_n vs ϵ_{n+1} .

(b) Grafica cuantilar (Normal Q-Q).

Figura 4.4: Supuestos de normalidad, temperatura tercer grado.

La gráfica cuantilar normal, de los puntos (X, Y) en donde cada valor X proviene del conjunto original de datos muestrales, Y son los valores de la puntuación normal correspondiente, es decir es un valor cuantilar esperado de la distribución normal estándar. Esto se muestra en la Figura 4.4.b, en donde se observa que la mayoría de los puntos siguen una línea recta, por lo cual este conjunto de datos tiene una distribución normal, excepto por algunos puntos en los extremos.

4.2.2. Retardo en la aparición de brotes epidemiológicos usando el factor de temperatura para BCS

Como se puede observar en la Figura 4.5 la aparición de brotes de dengue se presentan con un tiempo de retardo después de los cambios de ambientales, es decir, el brote epidémico se presenta unas semanas después de que la temperatura es óptima para la reproducción de los mosquitos que generan los contagios. Esto es, hay un retardo entre el factor de temperatura y su efecto en el brote epidémico (ver apéndice B1.3 para el código de la Figura 4.5).

Para obtener este retardo se calculó la correlación entre la temperatura semanal con los datos reportados de dengue. Esto se repitió sistemáticamente realizando traslaciones semanales de los datos de temperatura, donde se probaron hasta 55 semanas de corrimiento.

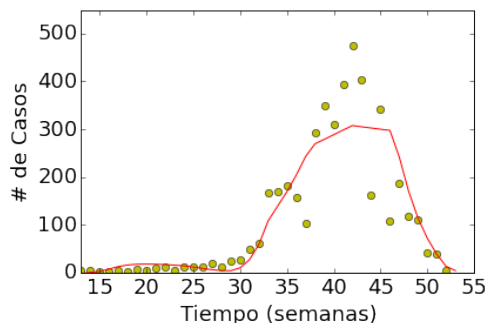


Figura 4.5: Modelo predictivo de incidencia de dengue con temperatura.

El modelo predice con anticipación de un mes. Por lo cual se concluye que el modelo es bueno ya que predice con buen tiempo de anticipación un brote de dengue. Esta información puede ser empleada para tomar medidas preventivas empleando los pronósticos meteorológicos de temperatura.

4.2.3. Efecto de la precipitación en la incidencia de dengue en BCS

También se obtuvo un MRLS para la incidencia de dengue como variable dependiente y la precipitación como variable independiente. Para lograr esto se emplearon los datos de incidencia de dengue y datos de precipitación mensual del año 2014 en BCS. En la Figura 4.6.a, se muestra con puntos de color verdes la incidencia de dengue y mediante la línea sólida de color azul la precipitación media mensual para BCS.

En la Figura 4.6.b, se muestra un diagrama de dispersión entre los casos reportados de dengue y los datos de precipitación. Este conjunto de datos se empleará para obtener el modelo para predecir incidencia de dengue. (Ver código de la Figura 4.6 en apéndice B1.4).

Debido a que la incidencia de dengue está dada en semanas, se estima la precipitación en semanas. Primero se ajustó un polinomio de segundo orden semejante al de la ecuación (4.1).

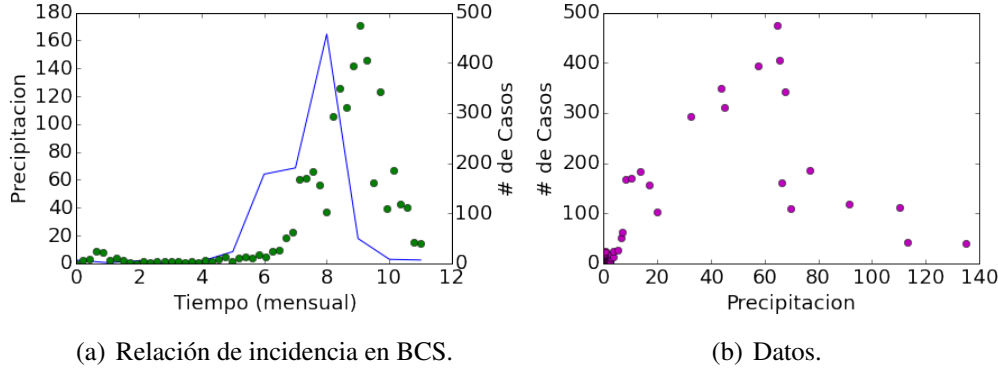


Figura 4.6: Datos para Baja California Sur de precipitación e incidencia.

Ahora, los estimadores de mínimos cuadrados obtenidos son:

$$\hat{\beta}_0 = 9.56412248 \times 10^1, \hat{\beta}_1 = 6.50721545, \hat{\beta}_2 = -4.30948634 \times 10^{-2},$$

así, el MRLS ajustado de segundo orden es:

$$\hat{Y} = 9.56412248 \times 10^1 + 6.50721545X - 4.30948634 \times 10^{-2}X^2. \quad (4.5)$$

El modelo dado por la ecuación (4.5) se muestra en la Figura 4.7.a por la línea sólida de color cian la cual se superpone a los datos de correlación de incidencia y precipitación representados por los puntos de color azules.

También se propuso un MRLS de forma polinomial de orden tres, análogo a la ecuación (4.3). Al calcular los estimadores para los coeficientes de este modelo, se determinaron los siguientes valores:

$$\begin{aligned} \hat{\beta}_0 &= 5.15258160 \times 10^1, \hat{\beta}_1 = 2.01911083 \times 10^1, \\ \hat{\beta}_2 &= -3.24681477 \times 10^{-1}, \hat{\beta}_3 = 1.31887819 \times 10^{-3}, \end{aligned} \quad (4.6)$$

de tal modo que el modelo de orden tres está dado por:

$$\hat{Y} = 5.1525 \times 10^1 + 2.0191 \times 10^1 X - 3.2468 \times 10^{-1} X^2 + 1.3188 \times 10^{-3} X^3. \quad (4.7)$$

La ecuación (4.7) da un buen ajuste a los datos de incidencia influenciada por la precipitación. Este MRLS se representa gráficamente en la Figura 4.7.b, por la línea sólida de color verde, la cual se superpone a los datos de incidencia representados por los puntos de color azul. (Ver código en Apéndice B1.5).

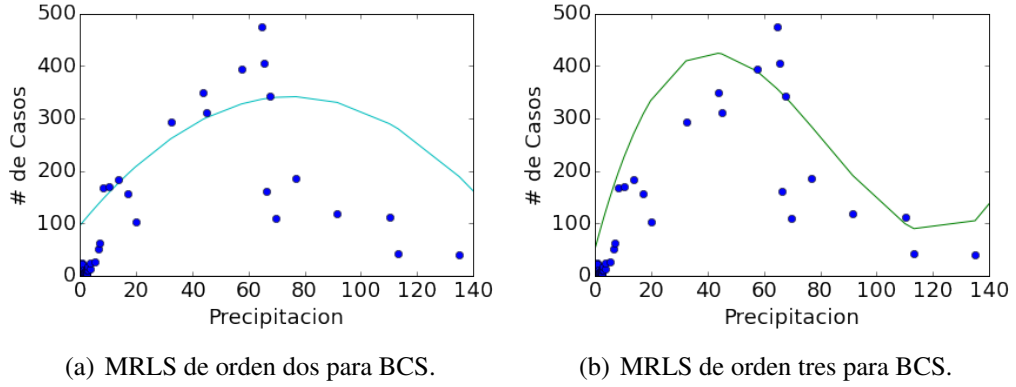
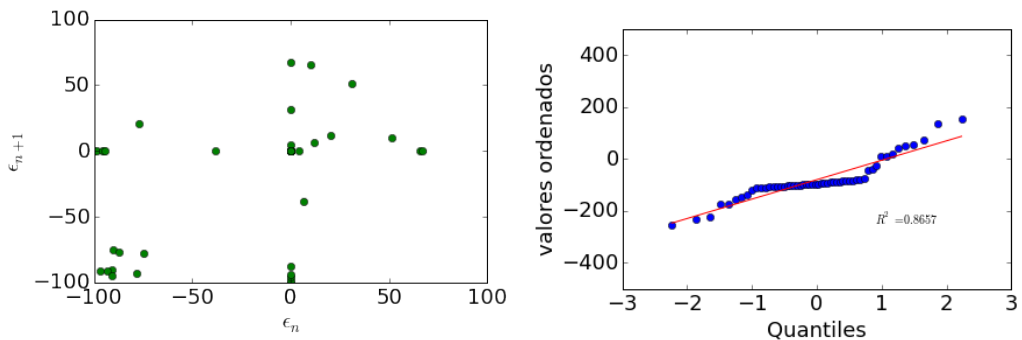


Figura 4.7: Modelo de regresión lineal simple para Baja California Sur (Precipitación).

Al igual que en el análisis del efecto de la temperatura, en la Figura 4.8.a se muestra un diagrama $\epsilon_n - \epsilon_{n+1}$ para el MRLS de segundo grado encontrado, dado por la ecuación (4.5).

Para el supuesto de normalidad se hizo una gráfica cuantilar normal, donde se grafican los puntos (X, Y) en donde cada valor X proviene del conjunto original de datos muestrales, Y son los valores de la puntuación normal correspondiente. Esto se muestra en la Figura 4.8.b, en donde se observa que los puntos se acercan a la línea recta de color rojo, por lo cual los datos tienen una distribución normal.



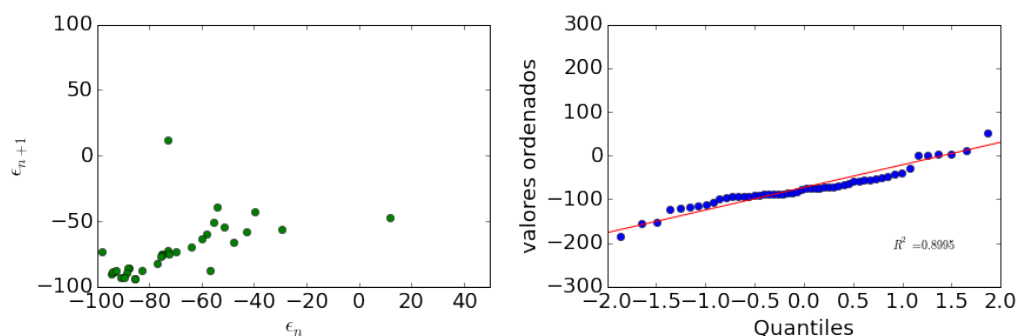
(a) Diagrama de independencia de las mediciones. Se grafica ϵ_n vs ϵ_{n+1} . (b) Gráfica cuantilar (Normal Q-Q).

Figura 4.8: Supuestos de normalidad, precipitación segundo grado.

Nuevamente en la Figura 4.9.a se muestra un diagrama $\epsilon_n - \epsilon_{n+1}$ para los errores del MRLS de tercer grado encontrado dado por la ecuación (4.7). Aquí se puede observar cierta correlación de los errores, lo que nos indica que para este caso el ajuste de segundo orden se comporta mejor con respecto a las hipótesis

necesarias.

Se hizo una gráfica cuantilar normal, donde se grafican los puntos (X, Y) en donde cada valor X proviene del conjunto original de datos muestrales, Y son los valores de la puntuación normal correspondiente. Esto se muestra en la Figura 4.9.b, en donde se observa que los puntos se acercan a la línea recta de color rojo, por lo cual los datos tienen una distribución normal.



(a) Diagrama de independencia de las mediciones. Se grafica ϵ_n vs ϵ_{n+1} .

(b) Grafica cuantilar (Normal Q-Q).

Figura 4.9: Supuestos de normalidad, precipitación tercer grado.

4.2.4. Retardo en la aparición de brotes epidemiológicos usando el factor de precipitación para BCS

Como se puede observar de la Figura 4.10.a la aparición de brotes que se predice mediante el modelo propuesto, se presentan antes que los brotes reales. Esto indica que hay un retardo entre el factor de precipitación y su efecto en el brote epidémico.

La Figura 4.10.b muestra la correlación entre el modelo de regresión con los datos reportados de Dengue a distintos corrimientos temporales del modelo. (Ver código para generar la figura en el Apéndice B1.6).

El modelo de regresión con un retardo de un mes tiene una correlación alta de 0.92 con respecto a la incidencia. Esto indica que el modelo es bueno. Esta información puede ser empleada para tomar medidas preventivas empleando los pronósticos meteorológicos de precipitación.

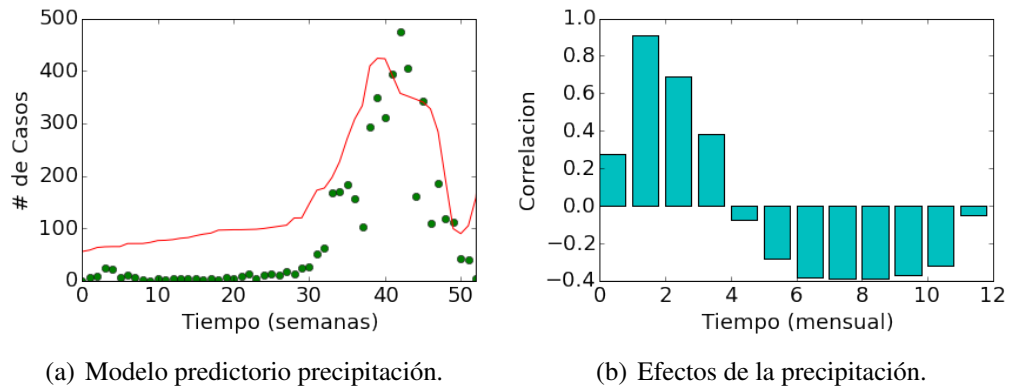


Figura 4.10: Modelo predictivo y correlación de incidencia de dengue con precipitación.

4.3. Efectos ambientales en la incidencia de dengue en México

Se estudiaron los efectos ambientales de temperatura y precipitación en la incidencia de dengue, mediante los datos agregados de los 32 estados de México, en donde se presentaron casos de dengue durante el año 2014. Para este modelo fue necesario escalar la incidencia con la población estatal correspondiente. También se eliminaron los valores estadísticamente atípicos de casos estatales.

4.3.1. Efecto de temperatura en la incidencia de dengue en México

La temperatura es un factor importante para la reproducción del mosquito *Aedes Aegypti* transmisor del virus del dengue. Por lo cual, se obtuvo un modelo de regresión lineal simple (MRLS) por medio del ajuste de un polinomio para la serie de tiempo de incidencia de dengue en México como variable dependiente y temperatura mínima semanal como variable independiente.

En la Figura 4.11 se graficaron los datos mensuales de incidencia de dengue con temperatura para cada estado de México, dado por la nube de puntos de color magenta. Sobre este conjunto de datos se ajustó el MRLS para predecir incidencia de dengue.

Se propuso un polinomio de segundo orden, dada por la ecuación (4.1). Los valores de los coeficientes estimados para este caso son:

$$\hat{\beta}_0 = 907.914, \hat{\beta}_1 = -122.186, \hat{\beta}_2 = 4.479.$$

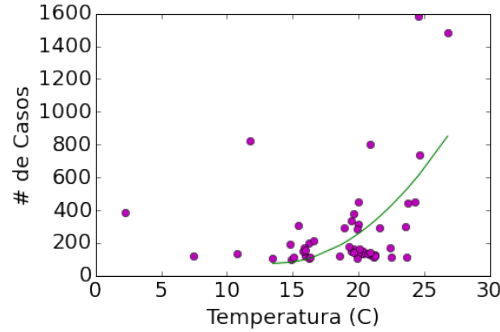


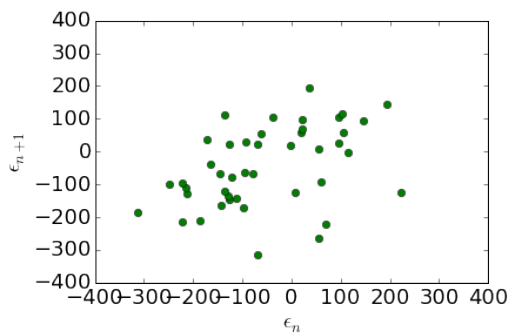
Figura 4.11: Los puntos constituyen el diagrama de dispersión para datos de incidencia con temperatura. La línea sólida es el MRLS del factor temperatura para predecir incidencia de dengue en México.

Siendo así el MRLS para datos de México que relaciona temperatura e incidencia:

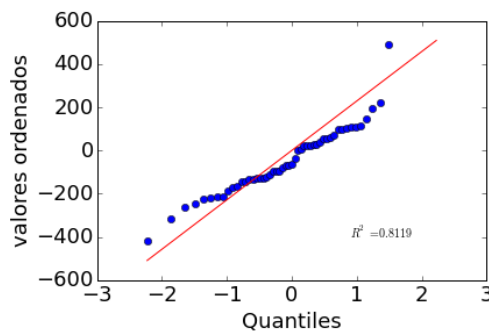
$$\hat{Y} = 907.914 - 122.186X + 4.479X^2. \quad (4.8)$$

Este modelo se aprecia en la Figura 4.11 por la línea de color verde junto con los datos de incidencia y temperatura (Ver código de la Figura 4.11 en Apéndice B1.7).

Ahora para mostrar independencia de las mediciones, se muestra en la Figura 4.12.a un diagrama $\epsilon_n - \epsilon_{n+1}$ del MRLS de segundo grado encontrado dado por la ecuación (4.8). Recordemos que si la nube de puntos en este diagrama está centrada, quiere decir que no hay correlación entre los errores ϵ mostrando su independencia.



(a) Diagrama de independencia de las mediciones. Se grafica ϵ_n vs ϵ_{n+1} .



(b) Grafica cuantilar (Normal Q-Q).

Figura 4.12: Supuestos de normalidad, temperatura segundo grado para México.

Se hizo una gráfica cuantilar normal, donde se grafican los puntos (X, Y) en

donde cada valor X proviene del conjunto original de datos muestrales, Y son los valores de la puntuación normal correspondiente. Esto se muestra en la Figura 4.12.b, en donde se observa que los puntos se acercan a la línea recta de color rojo, por lo cual los datos tienen una distribución normal.

4.3.2. Efecto de precipitación en la incidencia de dengue en México

Se obtuvo un modelo de regresión lineal simple (MRLS) por medio del ajuste de un polinomio para la serie de tiempo de incidencia de dengue en México como variable dependiente y precipitación semanal como variable independiente.

El efecto de la precipitación en la incidencia de dengue en México se analiza de manera semejante que el efecto de temperatura. En la Figura 4.13 se muestran los datos graficados de incidencia de dengue y precipitación mensual del año 2014 para cada estado de México sobre los cuales se ajustó el MRLS.

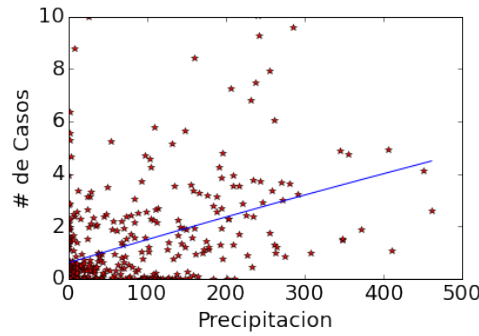


Figura 4.13: Los puntos constituyen el diagrama de dispersión de incidencia de dengue con precipitación. La línea sólida es MRLS estimado para predecir incidencia de dengue en México.

Se obtuvo el modelo que relaciona precipitación con incidencia de dengue ajustando un polinomio de segundo orden, de la forma de la ecuación (4.1). Los valores estimados de los parámetros para este análisis son:

$$\hat{\beta}_0 = 6.018 \times 10^{-1}, \hat{\beta}_1 = 8.965 \times 10^{-3}, \hat{\beta}_2 = -1.201 \times 10^{-6},$$

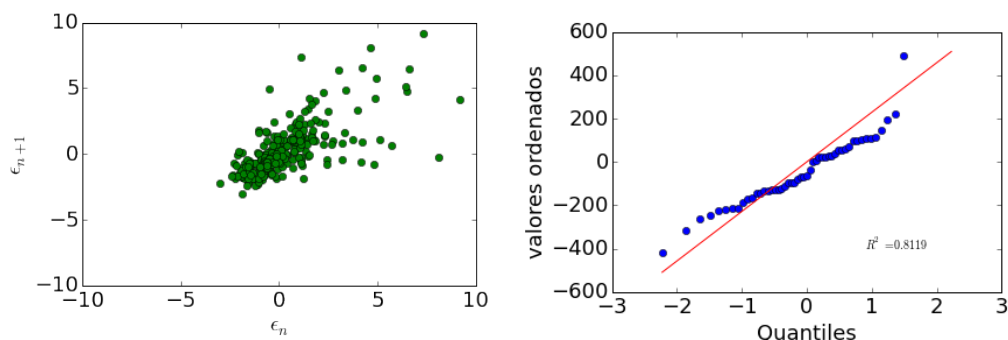
por tanto el MRLS para México está dado por:

$$\hat{Y} = 6.018 \times 10^{-1} + 8.965 \times 10^{-3}X - 1.201 \times 10^{-6}X^2. \quad (4.9)$$

En la Figura 4.13 se aprecian los datos de incidencia con precipitación y el modelo ajustado para predecir incidencia de dengue en México usando la variable

independiente de precipitación. (El código para generar la Figura 4.13 se puede ver en el Apéndice B1.9).

En la Figura 4.14.a se muestra un diagrama $\epsilon_n - \epsilon_{n+1}$ donde se observa poca correlación de los errores en el MRLS de segundo grado encontrado dado por la ecuación (4.9).



(a) Diagrama de independencia de las mediciones. Se grafica ϵ_n vs ϵ_{n+1} .

(b) Grafica cuantilar (Normal Q-Q).

Figura 4.14: Supuestos de normalidad, precipitación segundo grado para México.

Se hizo una gráfica cuantilar normal, donde se grafican los puntos (X, Y) en donde cada valor X proviene del conjunto original de datos muestrales, Y son los valores de la puntuación normal correspondiente. Esto se muestra en la Figura 4.14.b, en donde se observa que los puntos se acercan a la línea recta de color rojo, por lo cual los datos tienen una distribución normal.

4.4. Inclusión de factores sociales en un modelo de regresión lineal múltiple de incidencia de dengue

Se obtuvo un modelo de regresión lineal múltiple que involucra tanto variables sociales como ambientales que influyen en la propagación de dengue. Para calibrar el modelo se usó información del año 2011 de cada uno de los 32 estados de México. En la Figura 4.15 se puede apreciar la incidencia de dengue, donde la difuminación de cada estado está relacionada con la incidencia de dengue, para colores intensos incidencia alta y para colores tenues incidencia de dengue bajos (Ver código de mapa en el Apéndice B1.8).

La información recopilada que se tomará en cuenta en la construcción del modelo es para 27 estados, ya que como se observa en la Figura 4.15, son cinco

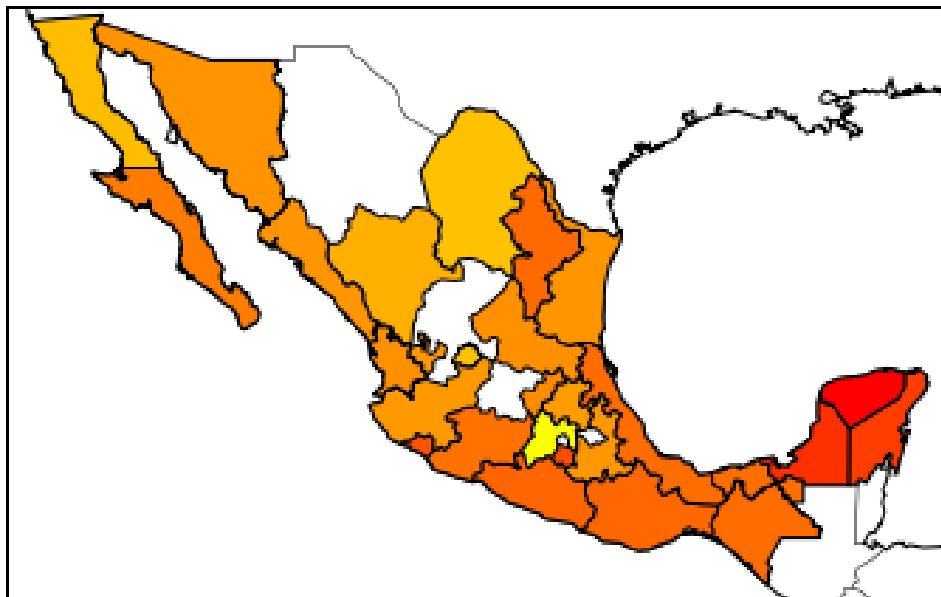


Figura 4.15: Mapa de riesgo de incidencia de dengue en México.

estados en los que no se reportó incidencia de dengue en el 2011. Las variables a tomar en cuenta son:

- Incidencia de dengue por entidad (PID): Está determinada por los casos de dengue del 2011 reportados por DGE entre el número de habitantes por estado.

Como variables sociales independientes se consideran:

- Disposición de residuos sólidos (DRS): Esta variable corresponde a las miles de toneladas de residuos sólidos que son tratadas en sitios controlados, por ejemplo los rellenos sanitarios reportados por INEGI.
- Promedio de habitantes por vivienda (PHV): Es importante considerar esta variable ya que se quiere ver qué tan relacionado está el número de habitantes por vivienda (hacinamiento) con la incidencia de dengue.
- Tomas de agua (TA): Es importante considerar esta variable ya que una vivienda al no contar con una toma de agua se ve en la necesidad de almacenar agua y con ello contribuir el incremento de criaderos del mosquito. Esta variable se determina dividiendo las tomas domiciliarias de agua entre el total de viviendas habitadas (no se consideran negocios ni casas móviles).

- Localidades con agua (LA): La variable está determinada por las localidades con red de distribución de agua entre el número total de localidades. Esta variable es importante para la obtención del MRLM ya que si las localidades no cuentan con una red de distribución de agua almacenan grandes cantidades para su abastecimiento.
- Localidades con drenaje y alcantarillado (LDA): La variable a considerar será realmente la que resulta de dividir LDA entre el número de localidades en el estado.

Como variables ambientales independientes se considerarán:

- Precipitación (P): El agua de lluvia puede almacenarse en recipientes y cacharros generando un gran número de criaderos para los mosquitos.
- Temperatura mínima (TMI): Variable ambiental en la cual el rango registrado para el año 2011 va de 5.975 °C a 21.808 °C.
- Temperatura media (TME): Importante considerar esta otra variable ya que en este rango se encuentra un rango de 14.3 °C a 26.858 °C, que favorecen el desarrollo del mosquito.
- Temperatura máxima (TMA): Una variable que afecta la tasa de desarrollo del mosquito *Aedes Aegypti*, y el rango registrado para esta variable va de 21.966 °C a 32.641 °C.

Nótese que los rangos de las variables de las tres diferentes temperaturas registradas son los promedios anuales para cada uno de los 27 estados considerados.

4.4.1. Análisis de correlación entre los factores Socio-ambientales e incidencia de dengue en México

En esta sección se analiza y se muestran las gráficas de cada una de las variables socio-ambientales que propician el criadero y aumento de la población del *Aedes Aegypti* contra la PID para cada estado.

En la Figura 4.16.a, se muestra la cantidad de generación de residuos sólidos, que hace referencia a la recolección total de residuos sólidos. De acuerdo a la gráfica se tiene que entre menos residuos sólidos se recolecten más incidencia de dengue hay ya que la basura se queda almacenada en las casas y toma la función de criaderos de mosquitos *Aedes Aegypti*. En la Figura 4.16.b, se gráfica la disposición de residuos sólidos contra la probabilidad de incidencia de dengue para cada estado. Los datos de DRS se concentran entre 12 y 2000 toneladas de residuos sólidos. En la gráfica se muestra que entre menos se trate la basura habrá

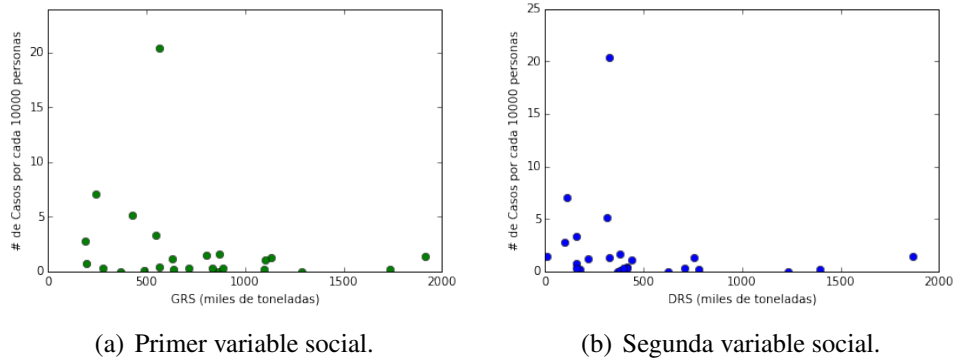


Figura 4.16: Variables sociales GRS y DRS con probabilidad de incidencia.

más incidencia de dengue (Ver código de la Figura 4.16 en el Apéndice B1.10).

En la Figura 4.17.a, se muestra la cantidad total de residuos sólidos, que es el resultado de la diferencia de la generación de residuos sólidos menos la disposición de residuos sólidos. Según la gráfica, esta variable tiene sentido si se toma en cuenta que entre menos basura se recolecte y se trate, más criaderos de mosquitos *Aedes Aegypti* habrá en las casas y por tanto, incrementa el riesgo de incidencia de dengue. La Figura 4.17.b, muestra el número promedio de habitantes por vivienda con la probabilidad de incidencia.

En dicha gráfica se muestra que hay mayor incidencia de dengue cuando hay un promedio de tres a cuatro habitantes por vivienda y que entre más o menos habitantes haya, habrá menos incidencia de dengue, esto es, el resultado de que las viviendas en algunos lugares se encuentran más distribuidas que en otros (Ver código de la Figura 4.17 en el Apéndice B1.11).

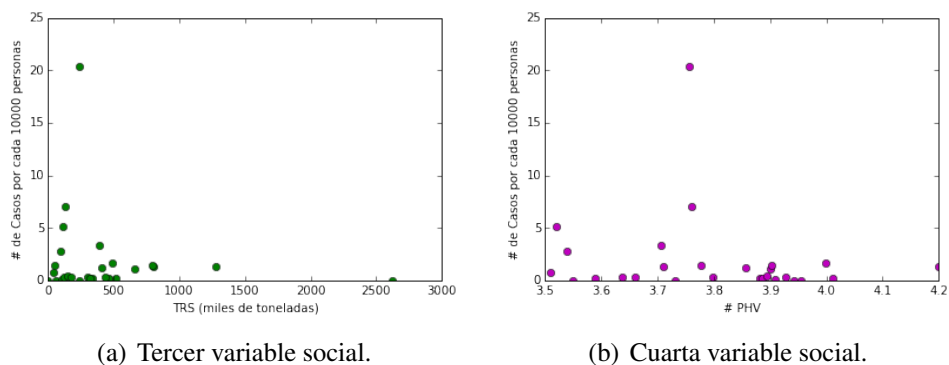


Figura 4.17: Variables sociales TRS y PHV con probabilidad de incidencia.

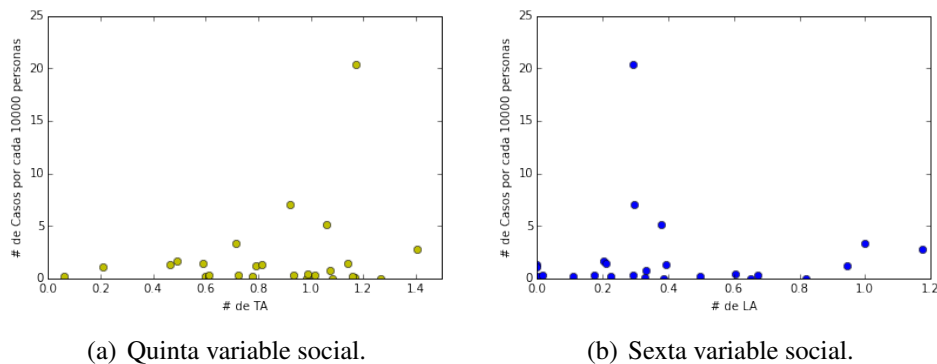


Figura 4.18: Variables sociales TA y LA con Probabilidad de incidencia.

En la Figura 4.18.a se gráfico el número promedio de tomas de agua por vivienda (TA) con la probabilidad de incidencia de dengue. En la gráfica, se observa que entre más tomas de agua, habrá un aumento en el número de incidencia de dengue. En la Figura 4.18.b, se gráfico la variable independiente del número de localidades con servicio de agua (LA) con probabilidad de incidencia. En esta gráfica se muestra que entre menos localidades cuentan con el servicio de agua potable mayor es la incidencia de dengue puesto que se ven en la necesidad de almacenar agua en recipientes para su abastecimiento, los cuales sirven de criaderos del mosquito *Aedes Aegypti* (Ver código de la Figura 4.18 en el Apéndice B1.12).

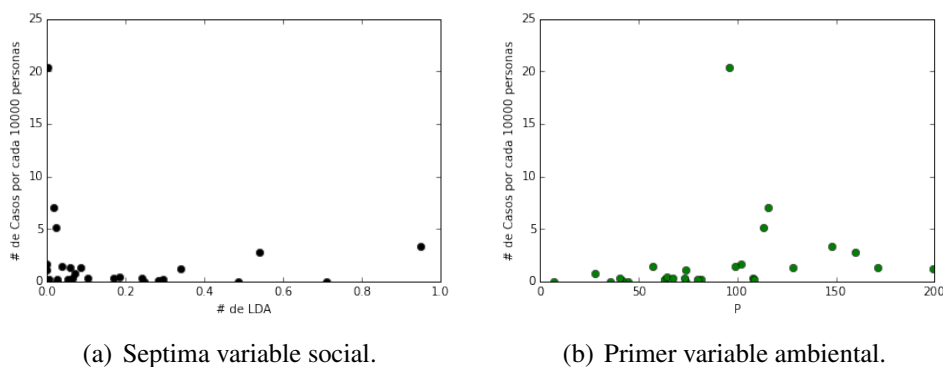


Figura 4.19: Variables sociales LDA y variable ambiental P con probabilidad de incidencia.

En la Figura 4.19.a, se muestra el número de localidades que cuentan con el servicio de drenaje y alcantarillado con la probabilidad de incidencia. La gráfica muestra que entre menos localidades cuentan con el servicio de drenaje y alcantarillado más incidencia hay. En la Figura 4.19.b, se muestran graficados los da-

tos de precipitación (P) con probabilidad de incidencia de dengue. Se observa de esta primer variable ambiental que para precipitaciones considerables; mayores a 200mm, al aumentar la precipitación disminuye el riesgo de incidencia de dengue, esto es debido a que los criaderos son lavados, eliminando a los *Aedes Aegypti* en cualquier etapa de su desarrollo biológico (Ver código de la Figura 4.19 en el apéndice B1.13).

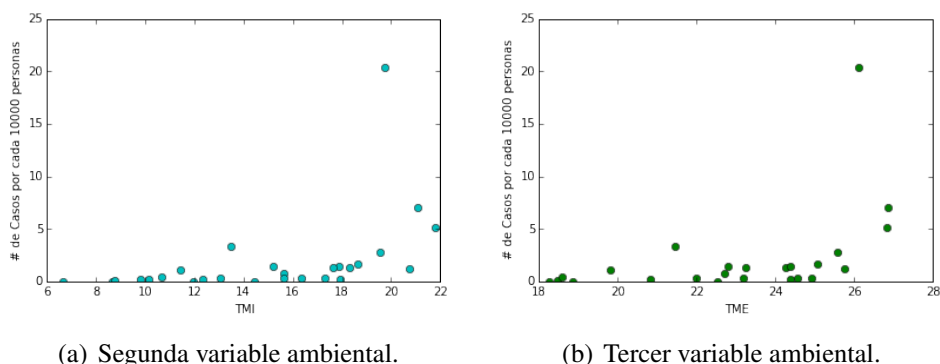


Figura 4.20: Variables ambientales de TMI y TME con probabilidad de incidencia.

En la Figura 4.20.a, se muestran los datos de temperaturas mínimas (TMI) con PID. Note que los casos de incidencia de dengue aumentan cuando la temperatura mínima aumenta, ya que, a bajas temperaturas los mosquitos *Aedes Aegypti* no son capaces de sobrevivir. En la Figura 4.20.b, se graficaron datos de la temperatura media (TME) con probabilidad de incidencia (Ver código de la Figura 4.20 en el apéndice B1.14).

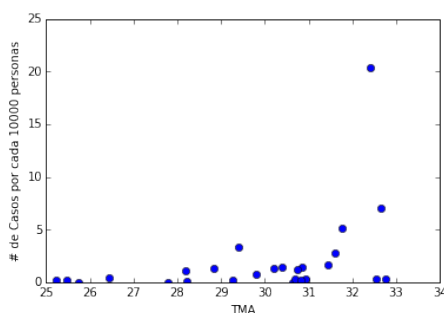


Figura 4.21: Variable ambiental de TMA con probabilidad de incidencia.

En la Figura 4.21, se grafican datos de temperaturas máximas (TMA) con probabilidad de incidencia. Donde la incidencia de dengue aumenta cuando aumenta

la temperatura máxima, siempre que no rebase los 45°C ya que al sobrepasar estos grados el mosquito no es capaz de sobrevivir (Ver código de la Figura 4.21 en el Apéndice B1.15).

Para analizar la fuerza de asociación entre la variable PID en relación con cada una de las variables socio-ambientales se calculó el coeficiente de correlación muestral (R). Además, por el Teorema 4.1 se tiene que, los estimadores $\hat{\beta}_i$ y ϵ_i para $i \leq k$, siguen una distribución normal por lo cual, probaremos el siguiente juego de hipótesis

$$H_0 : R = 0$$

vs

$$H_a : R \neq 0.$$

Lo que nos interesa es saber si $R = 0$, para lo cual la prueba se basa en el estadístico:

$$t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}.$$

Como el estadístico t tiene $n - 2$ grados de libertad, donde n es el tamaño de la muestra y α es el nivel de significancia entonces, la regla de decisión para este juego de hipótesis es rechazar H_0 si:

$$t \geq t_{\alpha/2} \text{ o si } t \leq -t_{\alpha/2}.$$

Para nuestro analisis $n - 2 = 25$ grados de libertad, nivel de significancia $\alpha = 0.5$. El valor t tabulado es $t_{0.5/2} = t_{0.25} = 0.68443$. Por tanto rechazaremos H_0 si $t \geq 0.68443$ o si $t \leq -0.68443$. Los resultados obtenidos se muestran en el Cuadro 4.1.

Variable	R	Estadístico t-student	Correlación
GRS	$-1.9256748 \times 10^{-1}$	-0.981201836	Correlacionado
DRS	$-2.0592639 \times 10^{-1}$	-1.052182856	Correlacionado
TRS	$-1.2503902 \times 10^{-1}$	-0.630140531	No correlacionado
PHV	$-1.4711189 \times 10^{-1}$	-0.743650494	Correlacionado
TA	1.84325×10^{-1}	0.991211160	Correlacionado
LA	-6.31863×10^{-3}	-0.031593795	No correlacionado
LDA	-9.397354×10^{-2}	-0.471956236	No correlacionado
P	2.2022945×10^{-1}	1.128863005	Correlacionado
TMI	4.4878249×10^{-1}	2.510977456	Correlacionado
TME	4.3856151×10^{-1}	2.439974062	Correlacionado
TMA	3.7241246×10^{-1}	2.006387116	Correlacionado

Cuadro 4.1: Tabla de correlación.

Ver código en Apéndice B1.16 para la obtención del coeficiente de correlación y el estadístico t.

Note que las primeras tres variables que se relacionan con residuos sólidos tiene un mayor coeficiente de correlación muestral absoluto, la variable DRS por lo cual es la variable que se tomó en cuenta en conjunto con las ocho variables restantes para la obtención del modelo de regresión lineal múltiple.

De acuerdo a la prueba de hipótesis se obtuvo que no se rechaza la hipótesis nula sólo para tres variables sociales (TRS, LA y LDA), es decir no hay suficiente evidencia para decir que están correlacionadas con la variable de probabilidad de incidencia de dengue (PID), ya que según la prueba de hipótesis el resto de variables (DRS, PHV, TA, P, TMI, TME y TMA) están correlacionadas con la variable PID.

4.4.2. Análisis de componentes principales

Con los datos de las variables socio-ambientales se realizó un análisis de componentes principales para reducir la dimensión del sistema. En el cual, se obtuvieron dos coeficientes principales. Para lo cual se calcularon los eigenvalores y eigenvectores asociados a la matriz de covarianza. Se extrajeron las dos componentes principales correspondientes a los eigenvectores asociados a los dos eigenvalores más grandes (Ver código en Apéndice B1.17 para la obtención de los eigenvalores y eigenvectores).

En el Cuadro 4.2, se pueden observar los coeficientes del primer eigenvector de la primer componente principal para el eigenvalor más grande que es $6.19472048535 \times 10^5$. El Cuadro 4.3, muestra los coeficientes de los eigenvectores de la segunda componente principal para el segundo eigenvalor más grande que en este caso es $2.00981999047 \times 10^3$.

Variable independiente	Valor
Disposición de residuos sólidos	$9.99901537 \times 10^{-1}$
Promedio de habitantes por vivienda	$1.36468760 \times 10^{-2}$
Tomas de agua	$-3.22154425 \times 10^{-3}$
Localidades con agua	$4.55412372 \times 10^{-4}$
Localidades con drenaje y alcantarillado	$-2.79629465 \times 10^{-4}$
Precipitación	$-6.29261920 \times 10^{-6}$
Temperatura mínima	$6.57353866 \times 10^{-5}$
Temperatura media	$-1.04907648 \times 10^{-4}$
Temperatura máxima	$-2.38010261 \times 10^{-6}$

Cuadro 4.2: Primer componente principal.

Variable independiente	Valor
Disposición de residuos sólidos	$3.82087565 \times 10^{-5}$
Promedio de habitantes por vivienda	$8.44427610 \times 10^{-4}$
Tomas de agua	$1.75445653 \times 10^{-2}$
Localidades con agua	$2.87174880 \times 10^{-2}$
Localidades con drenaje y alcantarillado	$-6.20584615 \times 10^{-2}$
Precipitación	$2.47345811 \times 10^{-1}$
Temperatura mínima	$-8.99537811 \times 10^{-1}$
Temperatura media	$-3.53061391 \times 10^{-1}$
Temperatura máxima	$-3.86756373 \times 10^{-3}$

Cuadro 4.3: Segunda componente principal.

De las dos componentes principales que se obtuvieron, se observa que la primer componente principal está relacionada principalmente con las variables sociales. Esta primera componente principal se representa de la siguiente forma:

$$Y^1 = Y_1^1 X_1 + Y_2^1 X_2 + Y_3^1 X_3 + Y_4^1 X_4 + Y_5^1 X_5,$$

en donde X_i para $i = 1, 2, 3, 4, 5$ son las respectivas variables sociales. Las variables Y_j^1 para $j = 1, 2, 3, 4, 5$ son los coeficientes del eigenvector correspondiente. Por lo cual se obtuvo que el coeficiente que más influye en la incidencia de dengue de este primer grupo de variables es la variable de disposición de residuos sólidos (DRS).

La segunda componente principal agrupa de manera mas representativa a las variables ambientales. Esta componente principal se representa de la siguiente forma:

$$Y^2 = Y_1^2 X_1 + Y_2^2 X_2 + Y_3^2 X_3 + Y_4^2 X_4,$$

en donde X_k para $k = 1, 2, 3, 4$ son las respectivas variables ambientales. Las variables Y_l^2 para $l = 1, 2, 3, 4$ son los coeficientes del eigenvector obtenido. Por lo cual se obtuvo que el coeficiente que más influye en la incidencia de dengue de este segundo grupo de variables es la variable de temperatura mínima (TMI), (El código para obtener las componentes principales se puede ver en el Apéndice B1.18).

La gráfica que ilustra este resultado para las primeras dos componentes principales es la Figura 4.22 (ver el código para generar la gráfica en Apéndice B1.19).

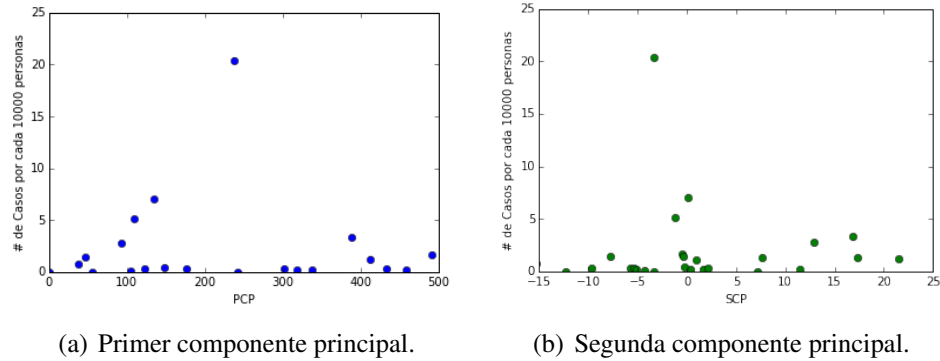


Figura 4.22: Componentes principales con probabilidad de incidencia.

4.4.3. Cálculo de los estimadores de mínimos cuadrados para el MRLM para México

El siguiente análisis se realizó usando las variables de disposición de residuos sólidos (DRS) y temperatura mínima (TMI). Estas son las variables independientes mas influyentes para cada grupo de variables independientes respectivamente que se obtuvieron con el ACP. Para observar la relación que existe entre ellas y la variable dependiente (PID), se graficaron los datos de dichas variables en la Figura 4.23 (Ver código de la Figura 4.23 en el Apéndice B1.20).

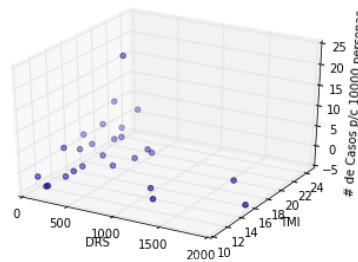


Figura 4.23: Variables de DRS y TMI para el MRLM.

Se propusieron dos modelos que toman en cuenta los factores sociales y ambientales. El primero, es un modelo reducido o lineal de orden uno de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \quad (4.10)$$

en donde X_1 corresponde a temperatura mínima (TMI) y X_2 representa la disposición de residuos sólidos (DRS). Al minimizar la suma de cuadrados entre la

ecuación y los puntos de los datos se obtiene que:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} -4.61986522 \\ 5.27658092 \times 10^{-5} \\ 4.32518778 \times 10^{-1} \end{bmatrix}.$$

Así, el MRLM está dado por

$$\hat{Y} = -4.61986522 + 5.27658092 \times 10^{-5} X_1 + 4.32518778 \times 10^{-1} X_2. \quad (4.11)$$

El modelo dado por la ecuación (4.11) y los datos de incidencia de dengue se muestra en la Figura 4.24.a.

El segundo, es un modelo de orden dos o completo con las mismas variables independientes (DRS y TMI). Dicho modelo es de la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon, \quad (4.12)$$

en donde

$$X_3 = X_1^2, X_4 = X_2^2, X_5 = X_1 X_2.$$

Ahora los nuevos estimadores para el MRLS completo están dados por:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 1.03902203 \times 10^1 \\ -1.50459320 \times 10^{-3} \\ -1.70122342 \\ 9.61475423 \times 10^{-8} \\ 7.08672648 \times 10^{-2} \\ 8.96065492 \times 10^{-5} \end{bmatrix},$$

de tal modo que el modelo completo ajustado es:

$$\begin{aligned} \hat{Y} = & 1.039022 \times 10^1 - 1.504593 \times 10^{-3} X_1 - 1.701223 X_2 \\ & + 9.614754 \times 10^{-8} X_3 + 7.086726 \times 10^{-2} X_4 + 8.960654 \times 10^{-5} X_5. \end{aligned} \quad (4.13)$$

La ecuación (4.13) se representa gráficamente y se superpone a los datos de incidencia como se muestra en la Figura 4.24.b (Ver código para generar Figura 4.24 en el Apéndice B1.21).

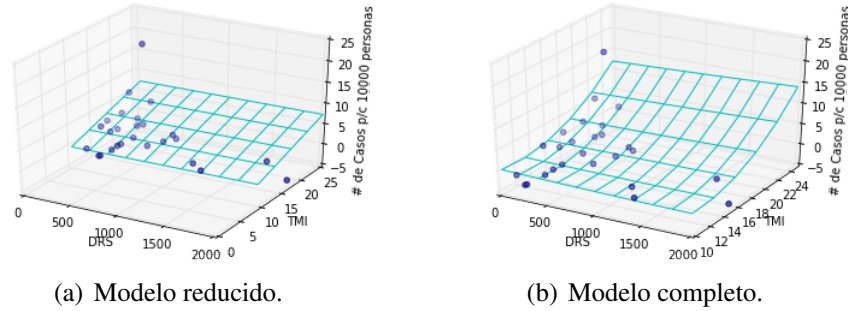


Figura 4.24: MRLM para incidencia de dengue en México.

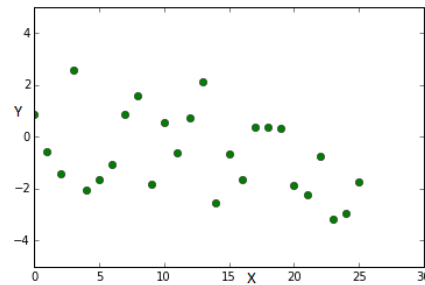
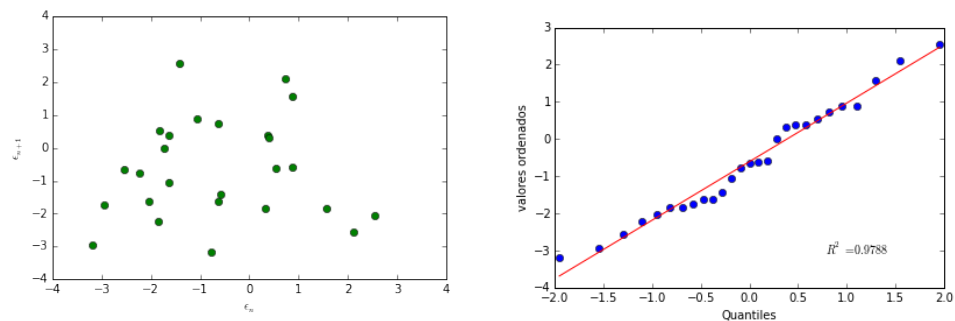


Figura 4.25: Diagrama de dispersión de los errores del MRLM reducido.

La Figura 4.25 es un diagrama de dispersión de los los errores para el MRLM reducido. Para mostrar independencia de las mediciones se muestra en la Figura 4.26.a un diagrama $\epsilon_n - \epsilon_{n+1}$ para los errores en los casos predichos por el MRLM reducido encontrado, dado por la ecuación (4.11).



(a) Diagrama de independencia de las mediciones. Se grafica ϵ_n vs ϵ_{n+1} .

(b) Grafica cuantilar (Normal Q-Q).

Figura 4.26: Supuestos de normalidad, modelo reducido.

La condición necesaria y que cumplen los datos es que los errores tienen valor

esperado cero, como los errores están distribuidos normalmente los estimadores de máxima verosimilitud coinciden con los estimadores de mínimos cuadrados ordinarios. Para esto, se hizo una gráfica cuantilar normal, donde se grafican los puntos (X, Y) en donde cada valor X proviene del conjunto original de datos muestrales, Y son los valores de la puntuación normal correspondiente, es decir es un valor cuantilar esperado de la distribución normal estándar. Esto se muestra en la Figura 4.26.b, en donde se observa que los puntos se acercan a la línea recta de color rojo, por lo cual los datos tienen una distribución normal.

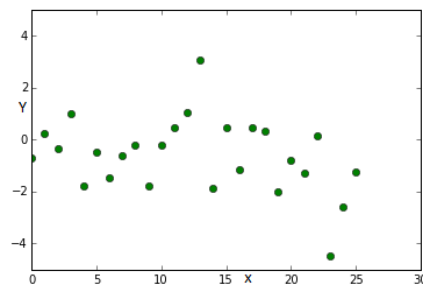
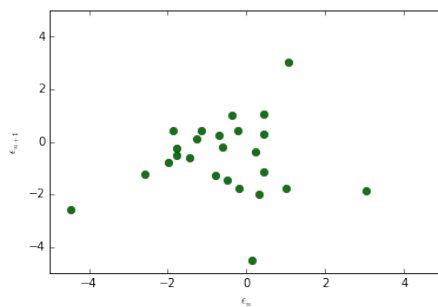
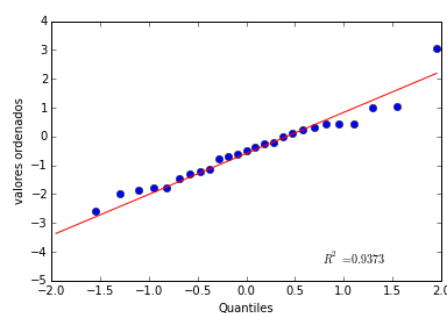


Figura 4.27: Diagrama de dispersión de los errores del MRLM completo.

La Figura 4.27 es un diagrama de dispersión de los los errores para el MRLM completo. Nuevamente para mostrar independencia de las mediciones se muestra en la Figura 4.28.a un diagrama para ϵ_{n+1} contra ϵ_n para el MRLM completo correspondiente a la ecuación (4.13).



(a) Diagrama de independencia de las mediciones. Se grafica ϵ_n vs ϵ_{n+1} .



(b) Gráfica cuantilar (Normal Q-Q).

Figura 4.28: Supuestos de normalidad, modelo completo.

La gráfica cuantilar normal, de los puntos (X, Y) en donde cada valor X proviene del conjunto original de datos muestrales, Y son los valores de la puntuación normal correspondiente, es decir es un valor cuantilar esperado de la distribución

normal estándar. Esto se muestra en la Figura 4.28.b, en donde se observa que los puntos se acercan a la línea recta de color rojo, por lo cual los datos tienen una distribución normal.

4.4.4. Prueba de mejor ajuste entre modelo lineal completo y reducido

A pesar de que la suma de los errores del modelo cuadrático es menor que la del modelo lineal, no significa necesariamente que sea un mejor modelo ya que tiene tres parámetros extra. Es decir, un buen modelo es aquel que nos da la mayor información con el menor número de parámetros.

Para seleccionar el mejor modelo entre el modelo lineal y el cuadrático, se realizó una prueba de hipótesis donde el juego de hipótesis a contrastar es:

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

vs

$$H_a : \beta_j \neq 0, \quad \text{al menos para un } j, j = 3, 4, 5.$$

Los coeficientes β_i corresponden a los coeficientes del modelo dado por la ecuación (4.12). Para esta prueba el estadístico F dado por

$$F = \frac{SSE_A / (k - p)}{SSE_C / (n - [k + 1])},$$

sigue una distribución F en caso de que la hipótesis nula sea verdadera [26]. Donde k es el número de variables independientes del modelo de regresión completo, p es el número de variables independientes del modelo de regresión reducido, n es el tamaño de la muestra y SSE_A , SSE_C denotan la suma de los errores cuadráticos para el modelo reducido y completo respectivamente. En este análisis $k = 5$, $p = 2$, $n = 27$.

Los valores para calcular el estadístico F se muestran en el Cuadro 4.4:

Suma	Valor
SSE_R	$3.40360896015 \times 10^{10}$
SSE_c	$2.50659556082 \times 10^{10}$
SSE_A	$8.97013399333 \times 10^9$
	G.L
$\nu_1 = k - p$	3
$\nu_2 = n - (k + 1)$	21
F	2.505028

Cuadro 4.4: Sumas de cuadrados del error y G.L.

Por tanto el estadístico $F = 2.505028$. El valor F tabulado para $\alpha = 0.1$ con $\nu_1 = 3$ grados de libertad en el numerador y $\nu_2 = 21$ grados de libertad en el denominador es $F_\alpha = 2.364888$. Por tanto, el valor observado del estadístico de prueba cae en la región de rechazo y se concluye que en el nivel de $\alpha = 0.1$ hay suficiente evidencia para apoyar la afirmación de que β_3, β_4 o β_5 , difieren de cero. Esto nos da evidencias para pensar que el MRLM cuadrático en las variables independientes es más adecuado para ajustar a los datos.

Conclusiones

Las técnicas estadísticas de modelos de regresión lineal ya sean simple o múltiple, son muy útiles en un gran número de fenómenos cuando no se conocen a detalle los mecanismos que generan las correlaciones entre variables independientes y la dependiente. En este trabajo se usaron dichas técnicas para encontrar un modelo de regresión lineal simple para predecir incidencia de dengue para el estado de Baja California Sur, por medio de la precipitación y uno más para predecir incidencia con datos de temperatura mínima.

Por otra parte se calculó la correlación que existe entre varios factores socio-ambientales y la incidencia de dengue en México. Con esto, se logró observar que en el conjunto de variables ambientales, es la temperatura mínima la que más influencia tiene en el riesgo de transmisión de la enfermedad. Por otro lado, en el conjunto de factores sociales, se observó que la disposición de residuos sólidos es un gran riesgo de salud pública respecto a potenciales brotes de dengue.

El análisis de componentes principales que se realizó, confirma que la influencia de los factores se agrupa estadísticamente en factores sociales y ambientales. Entre los factores sociales, la disposición de residuos sólidos es la componente más representativa. Mientras que la temperatura mínima es la componente más representativa de factores ambientales. Este resultado está de acuerdo con el análisis de correlaciones que se realizó.

Tomando en cuenta lo anterior, se construyeron dos modelos de regresión lineal múltiple donde las variables independientes fueron la disposición de residuos sólidos y la temperatura mínima. El primer modelo sólo considera dependencia lineal con respecto a las variables independientes, mientras que el segundo modelo considera términos cuadráticos. Se usó una prueba de hipótesis para verificar cuál de los dos modelos es el más adecuado y se concluyó que el modelo cuadrático podría ser el más adecuado.

Este estudio podrá ser de utilidad para el diseño de campañas preventivas dando indicadores sobre los tiempos óptimos de su ejecución mediante pronósticos de clima y mediciones de disposición de residuos sólidos en diversas regiones. Además, puede ser empleado como sistema de alerta temprana que ayude a los servicios de salud. Como un trabajo posterior se emplearán los resultados obtenidos para estimar la incidencia de dengue en zonas de menor escala, es decir en municipios y localidades para los cuales no se tenga información sobre de la incidencia de dengue.

Apéndice

A1.1 Demostración del Teorema 1.5:

Al interpretar la SSE como una variable aleatoria cuyos valores varían si el experimento se repite un número finito de veces, entonces

$$\begin{aligned} SSE &= S_{yy} - BS_{xy} \\ &= S_{yy} - B^2 S_{xx} \quad \text{dado que } (S_{xy} = BS_{xx}) \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - B^2 \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \sum_{i=1}^n Y_i^2 - n\bar{Y}^2 - B^2 \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right). \end{aligned}$$

Ahora bien si calculamos la esperanza de SSE se tiene:

$$\mathbf{E}(SSE) = \sum_{i=1}^n \mathbf{E}(Y_i^2) - n\mathbf{E}(\bar{Y}^2) - \mathbf{E}(B^2) \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right). \quad (4.14)$$

Haciendo uso del siguiente teorema:

Teorema 4.2

La varianza de una variable aleatoria Z , está dado por

$$\sigma^2 = E(Z^2) - \mu^2.$$

Se pueden sustituir las cantidades:

$$\begin{aligned} \mathbf{E}(Y_i^2) &= \sigma_{Y_i}^2 + \mu_{Y_i}^2, \\ \mathbf{E}(\bar{Y}^2) &= \sigma_{\bar{Y}}^2 + \mu_{\bar{Y}}^2, \\ \mathbf{E}(B^2) &= \sigma_{B_i}^2 + \mu_{B_i}^2. \end{aligned}$$

Sustituyendo estos resultados en la ecuación (4.14), se tiene que

$$\begin{aligned} \mathbf{E}(SSE) &= \sum_{i=1}^n (\sigma_{Y_i}^2 + \mu_{Y_i}^2) - n(\sigma_Y^2 + \mu_Y^2) - (\sigma_{B_i}^2 + \mu_{B_i}^2) \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) \\ &= n\sigma^2 + \sum_{i=1}^n (\alpha - \beta X_i)^2 - n \left[\frac{\sigma^2}{n} + (\alpha - \beta \bar{X})^2 \right] \\ &\quad - \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) \left(\frac{\sigma^2}{S_{xx}} + \beta^2 \right). \end{aligned}$$

Con $S_{xx} = \sum_{i=1}^n X_i^2 - n\bar{X}^2$ y al simplificar, se obtiene $\mathbf{E}(SSE) = (n-2)\sigma^2$. Por lo tanto,

$$E(S^2) = \frac{E(SSE)}{n-2} = \frac{(n-2)\sigma^2}{n-2} = \sigma^2$$

y se llega a que el estimador S^2 es insesgado. $\square$

A1.2 Demostración del Teorema 2.6

Para demostrar este resultado tómese primero dos variables independientes es decir $k = 2$

$$\begin{aligned} SSE &= \sum_{i=1}^n (\mathbf{Y} - \hat{\mathbf{Y}}_i)^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2 \\ &= \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})]^2 \\ &= \sum_{i=1}^n [Y_i^2 - Y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) + (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})^2 \\ &\quad - Y_i(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})] \\ &= \sum_{i=1}^n Y_i^2 - \beta_0 \sum_{i=1}^n Y_i - \beta_1 \sum_{i=1}^n Y_i X_{1i} - \beta_2 \sum_{i=1}^n Y_i X_{2i} \\ &\quad + \beta_0 \sum_{i=1}^n (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} - Y_i) \\ &\quad + \beta_1 \sum_{i=1}^n (\beta_0 X_{1i} + \beta_1 X_{1i}^2 + \beta_2 X_{1i} X_{2i} - X_{1i} Y_i) \\ &\quad + \beta_2 \sum_{i=1}^n (\beta_0 X_{2i} + \beta_1 X_{1i} X_{2i} + \beta_2 X_{2i}^2 - X_{2i} Y_i). \end{aligned}$$

Al hacer uso de las ecuaciones normales dadas en (1.8), se tiene que las tres últimas sumas son igual a cero, así

$$SSE = \sum_{i=1}^n Y_i^2 - \beta_0 \sum_{i=1}^n Y_i - \beta_1 \sum_{i=1}^n Y_i X_{1i} - \beta_2 \sum_{i=1}^n Y_i X_{2i}.$$

Al generalizar a k variables independientes, se tiene que:

$$\begin{aligned} SSE &= S_{yy} - \left[\sum_{j=0}^k (\hat{\beta}^t \mathbf{X}^t \mathbf{Y}) - \frac{(\sum_{i=1}^n Y_i)^2}{n} \right] \\ &= SST - SSR. \square \end{aligned}$$

A1.3

Teorema 4.3

Teorema de Roché-Frobenius

Sea el sistema $\mathbf{AX} = \mathbf{B}$ con m ecuaciones lineales sobre un campo $\mathbb{K}$ y con n incógnitas, donde m y n son naturales mayores que cero. Entonces,

1. El sistema $\mathbf{AX} = \mathbf{B}$ es compatible si y solo si $\text{rango}(\mathbf{A}) = \text{rango}(\mathbf{A}|\mathbf{B})$.
2. El sistema $\mathbf{AX} = \mathbf{B}$ es compatible determinado si y sólo si $\text{rango}(\mathbf{A}) = \text{rango}(\mathbf{A}|\mathbf{B}) = n$.

Demostración:

Represéntese como $FER(\mathbf{A})$ a la matriz en forma escalonada reducida de $\mathbf{A}$. Existe una matriz regular $\mathbf{P}$ tal que

$$FER(\mathbf{A}) = \mathbf{PA}.$$

Esta matriz $\mathbf{P}$ es la matriz producto de las matrices elementales que nos permiten obtener la forma escalonada reducida de $\mathbf{A}$. Puesto que $\mathbf{P}$ es regular, el sistema

$$(\mathbf{PA})\mathbf{X} = FER(\mathbf{A})\mathbf{X} = \mathbf{PB}$$

es equivalente a

$$\mathbf{AX} = \mathbf{B}.$$

Como el sistema $FER(\mathbf{A})\mathbf{X} = \mathbf{PB}$ tiene forma escalonada reducida, se le puede aplicar el método de Gauss-Jordan y se tiene que:

1. $\mathbf{AX} = \mathbf{B}$ es compatible si y sólo si $FER(\mathbf{A})\mathbf{X} = \mathbf{PB}$ lo es, es decir que la última columna de la matriz $[FER(\mathbf{A})|\mathbf{PB}]$ no sea principal y, por tanto, que

$$rango(FER(\mathbf{A})|\mathbf{PB}) = rango(FER(\mathbf{A}))$$

de la definición de rango, $FER(\mathbf{A}|\mathbf{B}) = [FER(\mathbf{A})|\mathbf{PB}]$ y de la igualdad anterior, tenemos que

$$rango(\mathbf{A}|\mathbf{B}) = rango(FER(\mathbf{A})) = rango(\mathbf{A}).$$

Así, queda demostrado el primer apartado.

2. Del mismo modo, el sistema $\mathbf{AX} = \mathbf{B}$ es determinado si y sólo si $FER(\mathbf{A}) = \mathbf{B}$ también lo es.
Es decir que en la matriz $FER(\mathbf{A}|\mathbf{PB})$ todas las columnas son principales excepto la última. Esto equivale a

$$rango(FER(\mathbf{A})|\mathbf{PB}) = rango(FER(\mathbf{A})) = n.$$

De donde se obtiene

$$rango(\mathbf{A}) = n = rango(\mathbf{A}|\mathbf{B}). \square$$

A1.4

Teorema 4.4

Eigenvalores de una matriz simétrica.

Sea $\mathbf{M}_{n \times n}$ matriz simétrica entonces los eigenvectores asociados a eigenvalores distintos son ortogonales.

Demostración:

Sean $\mathbf{v}_1$ y $\mathbf{v}_2$ eigenvectores asociados a los eigenvalores λ_1 y λ_2 respectivamente, $\lambda_1 \neq \lambda_2$.

Por demostrar que $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0$
entonces:

$$\begin{aligned} \lambda_1 \langle \mathbf{v}_1, \mathbf{v}_2 \rangle &= \lambda_1 \mathbf{v}_1^t \mathbf{v}_2 = (\lambda_1 \mathbf{v}_1)^t \mathbf{v}_2 = (\mathbf{M} \mathbf{v}_1)^t \mathbf{v}_2 = \mathbf{v}_1^t \mathbf{M}^t \mathbf{v}_2 \\ &= \mathbf{v}_1^t (\mathbf{M} \mathbf{v}_2) = \mathbf{v}_1^t \lambda_2 \mathbf{v}_2 = \lambda_2 \mathbf{v}_1^t \mathbf{v}_2 = \lambda_2 \langle \mathbf{v}_1, \mathbf{v}_2 \rangle \end{aligned}$$

Por lo que $(\lambda_1 - \lambda_2) \langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0$ dado que $(\lambda_1 - \lambda_2) \neq 0$, entonces $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle = 0. \square$

B1.Código-Python

Los acentos en las líneas de código se omiten ya que en el lenguaje de programación de python no están permitidos.

B1.1 Código para la Figura 4.1, temperatura e incidencia de Dengue. Donde **data.cases[region]** son los datos de incidencia de Dengue de la región de Baja California Sur y **temp** indica los datos de temperatura registradas para la misma región.

```
#INCISO (a)
fig1, ax1 = plt.subplots()
ax1.plot(range(0,len(data.cases[region])-1), temp, 'b')
ax1.set_xlabel('Tiempo (semanas)')
ax1.set_ylabel('Temperatura, C')
ax2 = ax1.twinx()
ax2.plot(range(0,len(data.cases[region])-1),data.cases[region][1:], 'go')
ax2.set_ylabel("Casos (semanas)")
ax2.set_ylim([0,500])
fig1.suptitle(data.cases[region][0], fontsize=20)
plt.show()
```

```
#INCISO (b)
fig22, ax1 = plt.subplots()
ax1.plot(Temp,data.cases[region][1:], 'mo')
ax1.set_xlabel('Temperatura (C)',)
ax1.set_ylabel('Casos')
fig22.suptitle(data.cases[region][0], fontsize=20)
ax1.axis([12, 25, 0, 300])
plt.show()
```

B1.2 Código para generar la Figura (4.2), en donde se calcula la correlación entre temperatura y casos de Dengue. Se estima también los modelos de orden dos y tres. Donde **xdata** representa las temperaturas registradas semanalmente y **ydata** indica los casos de Dengue registrados semanalmente.

```
#Ajuste a una cuadratica- inciso (a)
def func(x, a, b, c):
    return a + b*x + c*x*x
xdata = ShiftTemp
ydata = data.cases[region][1:]
x0 = np.array([10.0, 20.0, 1.0])
sigma = np.ones(len(xdata))
print optimization.curve_fit(func, xdata, ydata, x0, sigma)
fig23, ax1 = plt.subplots()
ax1.plot(xdata, func(xdata, 741.195, -99.679, 3.277), 'c')
ax1.plot(xdata, ydata, 'bo')
ax1.set_xlabel('Temperature (C)')
ax1.set_ylabel('Casos')
fig23.suptitle('Ajuste a una funcion cuadratica', fontsize=20)
ax1.axis([12, 25, 0, 300])
plt.show()
```

```
#Ajuste a una cubica- inciso (b)
def func(x, a, b, c, d):
    return a + b*x + c*x*x + d*x*x*x
x0 = np.array([714.195, -99.679, 3.277, 1.0])
sigma = np.ones(len(xdata))
print optimization.curve_fit(func, xdata, ydata, x0, sigma)
fig23, ax1 = plt.subplots()
ax1.plot(xdata, func(xdata, -1.90294525e+03, 3.72969314e+02, -
2.38602639e+01,
5.01366047e-01), 'm')
ax1.plot(xdata, ydata, 'bo')
ax1.set_xlabel('Temperature (C)')
ax1.set_ylabel('Casos')
fig23.suptitle('Ajuste a una funcion cubica', fontsize=20)
ax1.axis([12, 25, 0, 300])
plt.show()
```

B1.3 Código de la Figura 4.5 retardo en la aparición de brotes epidemiológicos para BCS. (Modelo predictorio)

```
# Modelo prediccion temperatura
prediction = np.zeros(len(data.cases[region])-1)
for i in range(0,len(data.cases[region])-1):
    prediction[i]=func(ShiftTemp[i], -1.90294525e+03,3.72969314e+02,
    -2.38602639e+01, 5.01366047e-01)
fig24, ax1 = plt.subplots()
ax1.plot(range(0,len(data.cases[region])-1),data.cases[region][1:], 'yo')
ax1.plot(range(0,len(data.cases[region])-1),prediction, 'r')
fig24.suptitle('Prediccion ', fontsize=20)
ax1.set_xlabel('Tiempo(semanas)')
ax1.set_ylabel('Casos')
ax1.axis([13, 55, 0, 550])
plt.show()
```

B1.4 Código para la Figura 4.6, precipitación e incidencia de Dengue. En el cual **Y** son los datos de incidencia de Dengue de la región de Baja California Sur y **X** indica los datos de precipitación registradas para la misma región.

```
#inciso (a)
X = np.array(precipitacion[region])
Y = np.array(data.cases[region])
fig2, ax1 = plt.subplots()
time = np.linspace(0,11,12)
ax1.plot(time,X,'b')
ax1.set_xlabel('Tiempo')
ax1.set_ylabel('Precipitacion')
time = np.linspace(0,11,52)
ax2 = ax1.twinx()
ax2.plot(time,Y,'go')
ax2.set_ylabel('casos')
fig2.suptitle('Baja California Sur', fontsize=20)
plt.show()
```

```
#inciso (b)
fig2, ax1 = plt.subplots()
xdataToplot = np.sort(X[X < 190])
ax1.plot(xdataToplot, data.cases[region][1:], 'mo')
ax1.set_xlabel('Precipitacion (semanal)',)
ax1.set_ylabel('Casos (semanales)')
fig2.suptitle("Baja California Sur", fontsize=20)
ax1.axis([0, 140, 0, 500])
plt.show()
```

B1.5 Código para la Figura 4.7, en donde se calcula la correlación entre precipitación y casos de Dengue, así como la estimación de los modelos de orden dos y tres. Nótese que **xdatas** indica las precipitaciones y **ydatas** indica los casos de Dengue estimados mensualmente.

```
#Ajuste del modelo cuadratico
def ModelPrecipitation(x,a,b,c):
    return a + b*x + c*x*x
xdatas = precipitacionWhitLag
ydatas = casosMensual
x0 = np.array([1.0,1.0,1.0])
sigma = np.ones(len(xdatas))
fit = optimization.curve_fit(ModelPrecipitation, xdatas, ydatas, x0, sigma)
print fit
fig59,ax1 = plt.subplots()
ax1.plot(xdatas, ydatas, 'ro')
ax1.plot(xdatas, ModelPrecipitation(xdatas,
fit[0][0],fit[0][1],fit[0][2]), 'b')
plt.xlabel("Precipitacion (mensual)")
plt.ylabel("Casos (mensuales)")
fig59.suptitle("Ajuste a una fuincion cuadratica", fontsize=20)
ax1.axis([0, 80, 0, 900])
plt.show()
```

```
#Ajuste a una cubica
def funci(x,a,b,c,d):
    return a + b*x + c*x*x + d*x*x*x
x01 = np.array([9.18434296e+01, 9.12241327e+00, -1.25183361e-03,1.0])
sigma1 = np.ones(len(xdatas))
print optimization.curve_fit(funci, xdatas, ydatas, x01, sigma1)
fig60, ax1 = plt.subplots()
xdataToplot = np.sort(xdatas[xdatas < 190])
ax1.plot(xdataToplot,funci(xdataToplot,-2.88952888e+01,5.07000023e+01,
-8.71281165e-01,3.78497648e-03),'g')
ax1.plot(xdata,ydata,'bo')
ax1.set_xlabel('Precipitacion (mensual)')
ax1.set_ylabel('Casos (mensual)')
fig60.suptitle("Ajuste a una funcion cubica", fontsize=20)
plt.show()
```

B1.6 Código de la Figura 4.10 retardo en la aparición de brotes epidemiológicos para BCS y correlación entre los datos y el modelo a distintos corrimientos temporales. (Modelo predictorio precipitación)

```
#Inciso (a)
# Modelo predicción precipitación
prediction = np.zeros(len(casosSemanales))
for i in range(0,len(casosSemanales)):
    prediction[i]=funci(precipitacionWhitLag[i],-2.88952888e+01,
5.07000023e+01,-8.71281165e-01,3.78497648e-03)
fig10, ax1 = plt.subplots()
ax1.plot(range(0,len(casosSemanales)),casosSemanales,'go')
ax1.plot(range(0,len(casosSemanales)),prediction,'r')
fig10.suptitle("Prediccion", fontsize=20)
plt.xlabel("Precipitacion ")
plt.ylabel("Casos ")
plt.show()
```

```
#Inciso (b)
for tau in range(0,12):
    precipitacionWhitLag = time_lag(X,tau)
    corr[tau]=np.corrcoef(precipitacionWhitLag, casosMensual[0][1])
fig3=plt.figure(3)
plt.bar(range(12),corr, color = 'r')
plt.xlabel("Retardo en correlacion")
plt.ylabel("Correlacion")
fig3.suptitle("Correlacion maxima", fontsize=20)
plt.show()
```

B1.7 El siguiente código genera la Figura 4.11, donde se aprecian los datos de incidencia de Dengue en México y el modelo ajustado para temperatura.

```
#Modelo para Temperatura vs Casos
def ModelTemperature(x, a, b, c):
    return a + b*x + c*x*x
xdata = nacTemperature
ydata = nacCases
x0 = np.array([1.0,1.0,1.0])
sigma = np.ones(len(xdata))
fit = optimization.curve_fit(ModelTemperature, xdata,
ydata, x0, sigma)
print fit
fig8, ax1=plt.subplots()
ax1.plot(xdata,ydata,'mo')
xdataToPlot = np.sort(xdata[xdata < 12])
ax1.plot(xdataToPlot,ModelTemperature(xdataToPlot,fit[0][0],
fit[0][1],fit[0][2]),'g')
plt.xlabel("Temperatura (mensual)")
plt.ylabel("Casos ( mensuales)")
fig8.suptitle("Ajuste a una funcion cuadratica", fontsize=20)
plt.show()
```

B1.8 El siguiente código genera el mapa de riesgo de incidencia de Dengue en México del año 2011.

```
from matplotlib.patches import Polygon
from matplotlib.collections import PatchCollection
from matplotlib.colors import rgb2hex
cmap = plt.cm.autumn
fig = plt.figure()
ax = fig.add_subplot(111)
m = Basemap(projection='mill',
llcrnrlat = 14,
llcrnrlon = -118,
urcrnrlat = 33, urcrnrlon = -86,
resolution='l')
m.readshapefile('./DATA/mexstates', 'mexstates')
m.readshapefile('./DATA/destdvlgw', 'destdvlgw')
patches = []
infos = []
for info, shape in zip(m.mexstates_info, m.mexstates):
patches.append( Polygon(np.array(shape), True) )
infos.append(info[ADMIN_NAME])
info1=sorted(zip(infos, patches))
estates = []
estates.append(info1[0]) #Aguascalientes
estates.append(info1[1]) #Baja California Norte
estates.append(info1[11]) #Baja California Sur
estates.append(info1[16]) #Campeche
estates.append(info1[17]) #Chiapas
estates.append(info1[19]) #Chihuahua
estates.append(info1[20]) #Coahuila
estates.append(info1[21]) #Colima
estates.append(info1[22]) #Distrito Federal
estates.append(info1[23]) #Durango
estates.append(info1[28]) #(Estado) Mexico
```

```

estates.append(info1[24]) #Guanajuato
estates.append(info1[25]) #Guerrero
estates.append(info1[26]) #Hidalgo
estates.append(info1[27]) #Jalisco
estates.append(info1[29]) #Michuacan
estates.append(info1[30]) #Morelos
estates.append(info1[31]) #Nayarit
estates.append(info1[34]) #Nuevo Leon
estates.append(info1[35]) #Oaxaca
estates.append(info1[36]) #Puebla
estates.append(info1[37]) #Queretaro
estates.append(info1[38]) #Quintana Roo
estates.append(info1[40]) #San Luis Potosi
estates.append(info1[41]) #Sinaloa
estates.append(info1[42]) #Sonora
estates.append(info1[45]) #Tabasco
estates.append(info1[46]) #Tamaulipas
estates.append(info1[47]) #Tlaxcala
estates.append(info1[48]) #Veracruz
estates.append(info1[49]) #Yucatan
estates.append(info1[50]) #Zacatecas
ax.add_collection(PatchCollection([estates[3][1]], facecolor= 'r', edgecolor= 'k', linewidths=1., zorder=2))
estates1=[]
Y = p_incidencia.astype(np.float)
y_incidencia = []
for i in range(len(Y)):
    if (Y[i]!=0.0):
        estates1.append(estates[i])
        y_incidencia.append(Y[i])
fig1, ax1 = plt.subplots()
for i in range(len(y_incidencia)):
    inc = (y_incidencia[i] - min(y_incidencia))/(max(y_incidencia)-min(y_incidencia))
    colorRGB = cmap(1.0-pow(inc,0.2))
    colorHex = rgb2hex(colorRGB)
    ax1.add_collection(PatchCollection([estates1[i][1]], facecolor= colorHex, linewidths=1.,zorder=(i+2)))
m.drawcoastlines()
m.drawcountries()
plt.show()

```

B1.9 El siguiente código genera la Figura 4.13, donde se aprecian los datos de incidencia de Dengue en México y el modelo ajustado para precipitación.

```
#Modelo para Precipitacion vs Casos
def ModelPrecipitationNac(x,a,b,c):
    return a + b*x + c*x*x
xdatass = nacPrecipitation[nacCases < 12]
ydatass = nacCases[nacCases < 12]
x0 = np.array([1.0,1.0,1.0])
sigma = np.ones(len(xdatass))
fit = optimization.curve_fit(ModelPrecipitationNac, xdatass, ydatass, x0, sigma)
print fit
fig9=plt.figure(9)
plt.plot(nacPrecipitation,nacCases,'r*')
plt.plot(np.sort(nacPrecipitation),ModelPrecipitationNac(np.sort(nacPrecipitation),fit[0][0],fit[0][1],fit[0][2]),'b')
plt.ylim((0,10))
plt.xlabel("Precipitacion (mensual)")
plt.ylabel("Casos (mensual) ")
fig9.suptitle("Ajuste a una funcion cuadratica", fontsize=20)
plt.show()
```

B1.10 Código para la Figura 4.16 de las variables sociales de GRS y DRS contra Probabilidad de incidencia de Dengue.

```
#Inciso (a)
fig2a, ax2a = plt.subplots()
ax2a.plot(X2a, y_incidencia, 'go')
ax2a.set_ylabel('# de Casos por cada 10000 personas')
ax2a.set_xlabel('GRS (miles de toneladas)')
fig2a.suptitle('Generacion de residuos S. vs PID', fontsize=18)
ax2a.axis([ 0, 2000,0, 24])
plt.show()
```

```
#Inciso (b)
#—Probabilidad de incidencia vs disposicion de residuos
fig2b, ax2b = plt.subplots()
ax2b.plot(X2d, y_incidencia, 'bo')
ax2b.set_ylabel('# de Casos por cada 10000 personas')
ax2b.set_xlabel('DRS (miles de toneladas)')
fig2b.suptitle('Disposicion de residuos S. vs PID', fontsize=18)
ax2b.axis([ 0, 2000,0, 25])
plt.show()
```

B1.11 Código para la Figura 4.17 de las variables sociales de TRS y PHV contra Probabilidad de incidencia de Dengue.

```
#Inciso (a)
#—Probabilidad de incidencia vs total de residuos solidos
fig2, ax2 = plt.subplots()
ax2.plot(X2, y_incidencia, 'go')
ax2.set_ylabel('# de Casos por cada 10000 personas')
ax2.set_xlabel('TRS (miles de toneladas)')
fig2.suptitle('Total de residuos S. vs PID', fontsize=18)
ax2.axis([ 0, 1400,0, 25])
plt.show()
#Inciso (b)
#—Promedio de habitantes por vivienda——
fig3, ax3 = plt.subplots()
ax3.plot(X3, y_incidencia, 'mo')
ax3.set_ylabel('# de Casos por cada 10000 personas')
ax3.set_xlabel('# PHV')
fig3.suptitle('Promedio de HV vs PID', fontsize=18)
plt.show()
```

B1.12 Código para la Figura 4.18 de las variables sociales de TA y LA contra Probabilidad de incidencia de Dengue.

```
#Inciso (a)
#—— Tomas domiciliarias de agua vs PID
fig4, ax4 = plt.subplots()
ax4.plot(X4, y_incidencia, 'yo')
ax4.set_ylabel('# de Casos por cada 10000 personas')
ax4.set_xlabel('# de Tomas de agua')
fig4.suptitle('Tomas de agua vs PID', fontsize=18)
ax4.axis([ 0, 1.5, 0, 25])
plt.show()
```

```
#Inciso (b)
#—— localidades con distribucion de agua vs PID
fig5, ax5 = plt.subplots()
ax5.plot(X5, y_incidencia, 'bo')
ax5.set_ylabel('# de Casos por cada 10000 personas')
ax5.set_xlabel('# de Localidades con agua')
fig5.suptitle('Localidades con agua vs PID', fontsize=18)
ax5.axis([ 0, 1.2, 0, 25])
plt.show()
```

B1.13 Código para la Figura 4.19 de la variable social de LDA y P contra Probabilidad de incidencia de Dengue.

```
#Inciso (a)
#—— Localidades con drenaje y alcantarillado vs PID
fig6, ax6 = plt.subplots()
ax6.plot(X6, y_incidencia, 'ko')
ax6.set_ylabel('# de Casos por cada 10000 personas')
ax6.set_xlabel('# de L. Drenaje-Alcantarillado')
fig6.suptitle('L.Drenaje y Alcantarillado vs PID', fontsize=18)
ax6.axis([0,1, 0, 25])
plt.show()
```

```
# Inciso (b)
# ———precipitacion vs PID
fig7, ax7 = plt.subplots()
ax7.plot(X7, y_incidencia, óg')
ax7.set_ylabel('# de Casos por cada 10000 personas')
ax7.set_xlabel('Precipitacion')
fig7.suptitle('Precipitacion vs PID', fontsize=18)
plt.show()
```

B1.14 Código para la Figura 4.20 de las variables ambientales de TMI y TME con Probabilidad de incidencia de Dengue.

```
#Inciso (a)
# ——— temperatura mínima vs PID
fig8, ax8 = plt.subplots()
ax8.plot(X8, y_incidencia, óc')
ax8.set_ylabel('# de Casos por cada 10000 personas')
ax8.set_xlabel('Temperatura minima')
fig8.suptitle('Temperatura minima vs PID', fontsize=18)
plt.show()
#Inciso (b)
# ——— temperatura media vs PID
fig9, ax9 = plt.subplots()
ax9.plot(X9, y_incidencia, óg')
ax9.set_ylabel('# de Casos por cada 10000 personas')
ax9.set_xlabel('Temperatura media')
fig9.suptitle('Temperatura media vs PID', fontsize=18)
ax9.axis([18, 28, 0, 25])
plt.show()
```

B1.15 Código para la Figura 4.21 de las variable ambiental de TMA con Probabilidad de incidencia de Dengue.

```
# ——— temperatura maxima vs PID
fig10, ax10 = plt.subplots()
ax10.plot(X10, y_incidencia, ób')
ax10.set_ylabel('# de Casos por cada 10000 personas')
ax10.set_xlabel('Temperatura maxima')
fig10.suptitle('Temperatura maxima vs PID', fontsize=18)
ax10.axis([25, 34, 0, 25])
plt.show()
```

B1.16 Código para calcular el coeficiente de correlación muestral y el estadístico t para cada una de las variables socio-ambientales e incidencia de Dengue. Donde $n=1,2,\dots,9$, e indica cada una de las nueve variables socio-ambientales de estudio y m es el tamaño de la muestra.

```
Y=P_incidencia.astype(np.float)
Xn = []
y_incidencia = []
for i in range(len(Y)):
    if (Y[i]!=0.0):
        Xn.append(variable-independiente.astype(np.float)[i])
        y_incidencia.append(Y[i])
R=np.corrcoef(np.array(Xn), np.array(y_incidencia))
print('estadístico t')
t=((R*(math.sqrt(m-2)))/(math.sqrt(1-(pow(R,2)))))
```

B1.17 Código para obtener la matriz de covarianza y los respectivos eigenvalores y eigenvectores del ACP.

```
cov_mat = np.cov([TRS,PHV,TA,LA,LDA,P,TMI,TME,TMA])
eig_valc, eig_vecc=np.linalg.eig(cov_mat)
for i in range(len(eig_valc)):
    eigvec_cov = eig_vecc[:,i].reshape(1,9).T
    print('Eigenvalor {} de matriz de covarianza: {}'.format(i + 1, eig_valc[i]))
    print('Eigenvector {} '.format(eig_vecc[i]))
```

B1.18 Código para obtener las dos coeficientes principales.

```
eig-unit = []
for ev in eig-vecc:
    eig-unit.append(ev/np.linalg.norm(ev))
xx= []
yy=[]
for estado in range(len(X2)):
    xx.append(eig-unit[0][0]*X2[estado]+eig-unit[0][1]*X3[estado]+eig-
unit[0][2]*X4[estado]
+eig-unit[0][3]*X5[estado]+eig-unit[0][4]*X6[estado]+eig-
unit[0][5]*X7[estado]
+eig-unit[0][6]*X8[estado]+eig-unit[0][7]*X9[estado]+eig-
unit[0][8]*X10[estado])
    yy.append(eig-unit[1][0]*X2[estado]+eig-unit[1][1]*X3[estado]+eig-
unit[1][2]*X4[estado]
+eig-unit[1][3]*X5[estado]+eig-unit[1][4]*X6[estado]+eig-
unit[1][5]*X7[estado]
+eig-unit[1][6]*X8[estado]+eig-unit[1][7]*X9[estado]+eig-
unit[1][8]*X10[estado])
```

B1.19 Código para obtener Figura 4.22 de las dos coeficientes principales.

```
#——Grafica 1a componente principal e incidencia
fig11, ax11 = plt.subplots()
ax11.plot(xx, y_incidencia, ób')
ax11.set_ylabel('# de Casos por cada 10000 personas')
ax11.set_xlabel('Primer componente')
ax11.axis([0, 1500,0, 0.00025])
plt.show()
#——Grafica 2a componente principal e incidencia
fig12, ax12 = plt.subplots()
ax12.plot(yy, y_incidencia, óg')
ax12.set_ylabel('# de Casos por cada 10000 personas')
ax12.set_xlabel('Segunda componente ')
ax12.axis([-30, 5,0, 0.00020])
plt.show()
```

B1.20 Código para obtener Figura 4.23 de correlación de las variables mas influyentes e incidencia de dengue.

```
fig13=plt.figure(13)
ax=fig13.add_subplot(111,projection='3d')
ax.scatter(X2,X8,y_incidencia)
ax.set_xlabel('Disposicion del de Residuos Solidos')
ax.set_ylabel('Temperatura minima')
ax.set_zlabel('# de Casos por cada 10000 personas') ax.set_title('Correlacion
componentes principales')
plt.show()
```

B1.21 El siguiente código genera la Figura 4.24, donde se aprecian los datos de incidencia de Dengue en México y los modelos ajustados de primer y segundo orden respectivamente.

```
#—modelo prediccion-superficie primer orden
def fnc(x, g, h, k):
    return g + h*x[:,0] + k*x[:,1]
cpr= scipy.array([X2, X8, y_incidencia]).T
y_dengue= scipy.array(y_incidencia)
x0r= np.array([1.0,1.0,1.0])
fitParams, fitCovariances = curve_fit(fnc, cpr[:,2], cpr[:,2], x0r)
print 'coeficientes fit : \n', fitParams
fig14=plt.figure(14)
ax=fig14.add_subplot(111,projection='3d')
ax.scatter(X2,X8,y_incidencia)
X = np.arange(0,3000,100)
Y = np.arange(0,250,10)
X, Y = np.meshgrid(X, Y)
Z = fitParams[0] + fitParams[1]*X + fitParams[2]*Y
ax.plot_wireframe(X,Y,Z,color='b',)
ax.set_xlabel('Disposicion de Residuos Solidos')
ax.set_ylabel('Tempereatura minima')
ax.set_zlabel('# de Casos por cada 10000 personas')
ax.set_title('Modelo prediccion primer orden')
plt.axis([0, 1200, 0, 200])
plt.show()
```

```
#—modelo prediccion—superficie de segundo orden
def fn(x, a, b, c, d, e, f):
    return a + b*x[:,0] + c*x[:,1] + d*x[:,0]*x[:,0] + e*x[:,1]*x[:,1] +
    f*x[:,0]*x[:,1]
cp= scipy.array([X2, X8, y_incidencia]).T
y_dengue= scipy.array(y_incidencia)
x0 = np.array([1.0,1.0,1.0,1.0,1.0,1.0])
fitParams, fitCovariances = curve_fit(fn, cp[:,2], cp[:,2], x0)
print 'coeficientes fit : \ n', fitParams
fig14=plt.figure(15)
ax=fig15.add_subplot(111,projection='3d')
ax.scatter(X2,X8,y_incidencia)
X = np.arange(0,3000,100)
Y = np.arange(0,250,10)
X, Y = np.meshgrid(X, Y)
Z = fitParams[0] + fitParams[1]*X + fitParams[2]*Y + fitParams[3]*X*X
+ fitParams[4]*Y*Y+fitParams[5]*X*Y
ax.plot_wireframe(X,Y,Z,color='c',)
ax.set_xlabel('Disposicion de Residuos Solidos')
ax.set_ylabel('Temperatura minima')
ax.set_zlabel('# de Casos por cada 10000 personas')
ax.set_title('Modelo prediccion segundo orden ')
plt.axis([0, 1200, 0, 200])
plt.show()
```

Índice de figuras

2.1. Elipse de concentración.	23
3.1. Aedes Aegypti.	34
4.1. Datos para Baja California Sur de temperatura e incidencia.	42
4.2. Modelos de regresión lineal simple para Baja California Sur (temperatura).	44
4.3. Supuestos de normalidad, temperatura segundo grado.	44
4.4. Supuestos de normalidad, temperatura tercer grado.	45
4.5. Modelo predictivo de incidencia de dengue con temperatura.	46
4.6. Datos para Baja California Sur de precipitación e incidencia.	47
4.7. Modelo de regresión lineal simple para Baja California Sur (Precipitación).	48
4.8. Supuestos de normalidad, precipitación segundo grado.	48
4.9. Supuestos de normalidad, precipitación tercer grado.	49
4.10. Modelo predictivo y correlación de incidencia de dengue con precipitación.	50
4.11. Los puntos constituyen el diagrama de dispersión para datos de incidencia con temperatura. La línea solida es el MRLS del factor temperatura para predecir incidencia de dengue en México.	51
4.12. Supuestos de normalidad, temperatura segundo grado para México.	51
4.13. Los puntos constituyen el diagrama de dispersión de incidencia de dengue con precipitación. La línea sólida es MRLS estimado para predecir incidencia de dengue en México.	52
4.14. Supuestos de normalidad, precipitación segundo grado para México.	53
4.15. Mapa de riesgo de incidencia de dengue en México.	54
4.16. Variables sociales GRS y DRS con probabilidad de incidencia.	56
4.17. Variables sociales TRS y PHV con probabilidad de incidencia.	56
4.18. Variables sociales TA y LA con Probabilidad de incidencia.	57
4.19. Variables sociales LDA y variable ambiental P con probabilidad de incidencia.	57

4.20. Variables ambientales de TMI y TME con probabilidad de incidencia.	58
4.21. Variable ambiental de TMA con probabilidad de incidencia.	58
4.22. Componentes principales con probabilidad de incidencia.	62
4.23. Variables de DRS y TMI para el MRLM.	62
4.24. MRLM para incidencia de dengue en México.	64
4.25. Diagrama de dispersión de los errores del MRLM reducido.	64
4.26. Supuestos de normalidad, modelo reducido.	64
4.27. Diagrama de dispersión de los errores del MRLM completo.	65
4.28. Supuestos de normalidad, modelo completo.	65

Índice de cuadros

4.1. Tabla de correlación.	59
4.2. Primer componente principal.	60
4.3. Segunda componente principal.	61
4.4. Sumas de cuadrados del error y G.L.	67

Bibliografía

- [1] Banco de informacion INEGI- Producción. <http://www3.inegi.org.mx/sistemas/biinegi/>. Visitado: 2016-03-03.
- [2] Chikungunya llegó a México en 2014; ya hay más de 3 mil casos - Grupo Milenio. http://www.milenio.com/estados/chikungunya-virus_del_chikungunya_en_Mexico-alerta_en_16_estados_por_virus_0_566943428.html. Visitado: 2016-08-03.
- [3] Dirección General de Epidemiología. http://www.epidemiologia.salud.gob.mx/dgae/panodengue/historicos_dengue.html. Visitado: 2016-08-03.
- [4] Entomology & Ecology | Dengue | CDC. <http://www.cdc.gov/dengue/entomologyEcology/>. Visitado: 2016-08-03.
- [5] Geoestadística. Localidades Geoestadísticas - archivo histórico - consulta. http://www.inegi.org.mx/geo/contenidos/geoestadistica/consulta_localidades.aspx. Visitado: 2016-08-03.
- [6] Hoja de datos sobre el dengue - Dengue - CDC en Español. <http://www.cdc.gov/spanish/enfermedades/dengue/hojadatos.htm>. Visitado: 2016-08-18.
- [7] México aprueba la primera vacuna contra el virus del dengue a nivel mundial. <http://www.cofepris.gob.mx/Documents/NotasPrincipales/09122015.pdf>. Visitado: 2016-09-06.
- [8] OMS | Chikungunya. <http://www.who.int/mediacentre/factsheets/fs327/es/>. Visitado: 2016-08-03.
- [9] Qué es la chikungunya, la enfermedad que avanza sin descanso en América Latina - BBC Mundo. <http://www.bbc.com/mundo/>

noticias/2015/05/150512_salud_que_es_enfermedad_chikungunya_lv. Visitado: 2016-08-03.

- [10] Resúmenes Mensuales de Temperaturas y Lluvia. <http://smn.cna.gob.mx/es/climatologia/temperaturas-y-lluvias/resumenes-mensuales-de-temperaturas-y-lluvias>. Visitado: 2016-03-17.
- [11] AGRELO, R. S. *Aedes aegypti*, *Aedes albopictus* (Diptera, Culicidae) y su papel como vectores em las Américas. La situación de Uruguay. *Rev Med Uruguay* 12 (1996), 28–36.
- [12] BADII, M. H. Ecología e historia del dengue en las Américas (Ecology and history of dengue in Americas) Badii, MH, J. Landeros, E. Cerna y JL Abreu.
- [13] BAR, D. M. E. El *Aedes aegypti* y la transmisión del Dengue.
- [14] FREUND, JOHN E., E. A. *Estadística matemática con aplicaciones*. Pearson Educación, 2000.
- [15] GETIS, ARTHUR, E. A. Characteristics of the spatial pattern of the dengue vector, *Aedes aegypti*, in Iquitos, Peru. *The American Journal of Tropical Medicine and Hygiene* 69, 5 (2003), 494–505.
- [16] GIL, SAID INFANTE, E. A. *Métodos estadísticos: un enfoque interdisciplinario*. Trillas, 1990.
- [17] GONZÁLEZ OLVERA, G. *Resistencia a insecticidas en el mosquito vector del dengue *Aedes aegypti* (L) en dos épocas de transmisión de la enfermedad en Mérida, Yucatán*. PhD thesis, Universidad Autónoma de Nuevo León, 2013.
- [18] GUBLER, D. J. Dengue and Dengue Hemorrhagic Fever. *Clinical Microbiology Reviews* 11, 3 (1998), 480–496.
- [19] LÓPEZ AMAYA, J. L. *Un modelo tipo predador-presa para la dinámica de población del mosquito *Aedes aegypti* con control químico y control biológico*. PhD thesis, Universidad de El Salvador, 2011.
- [20] MATTERN, S. L. D. I. *Inferencia estadística y análisis de datos*. Pearson Educación, 2008.
- [21] MILTON, J. SUSAN, E. A. *Estadística para biología y ciencias de la salud*. McGraw-Hill, Interamericana de España, Madrid, 2007. OCLC: 434480697.

-
- [22] MONTGOMERY, DOUGLAS C., E. A. *Probabilidad y estadística aplicadas a la ingeniería*. Limusa Wiley, 2002.
- [23] MUNDO, B. B. C. Zika: el bosque de Uganda de donde salió el virus que afecta a América Latina. http://www.bbc.com/mundo/noticias/2016/01/160129_salud_uganda_bosque_zika_virus_enfermedad_lv. Visitado: 2016-08-03.
- [24] MUNDO, B. B. C. ¿Qué es el zika, el virus que se está propagando por América Latina? http://www.bbc.com/mundo/noticias/2015/11/151128_salud_nuevo_virus_zika_men.shtml. Visitado: 2016-08-03.
- [25] PEÑA, DANIEL. *Análisis de datos multivariantes*. McGraw-Hill, Aravaca (Madrid), 2002.
- [26] WACKERLY, DENNIS D., E. A. *Estadística Matemática con Aplicaciones*. Cengage Learning, 2009.
- [27] WALPOLE, RONALD E., E. A. *Probabilidad y estadística para ingenieros*. Pearson Educación, 1999.