Regresión múltiple y otras técnicas multivariadas

Especialización en Estadística Aplicada

Javier Santibáñez

IIMAS, UNAM

jsantibanez@sigma.iimas.unam.mx

Semestre 2017-2

Outline

Unidad 1

Introducción

Introducción

- La humanidad siempre ha intentado explicar el mundo en que vivimos.
- Los primeros intentos atribuían los fenómenos físicos a causas sobrenaturales (magia y religión).
- Con el desarrollo de la ciencia, se lograron explicaciones objetivas y más precisas.
- Explicar la realidad en que vivimos es un proceso que aún continua...

Modelos

En general, se puede representar a los fenómenos que ocurren en la naturaleza como sigue:

La representación anterior es conocida como *modelo*.

Un modelo es una abstracción que simplifica la realidad y permite estudiarla con mayor detalle.

Tipos de fenómenos

Existen dos corrientes científicas para describir cómo ocurren los fenómenos:

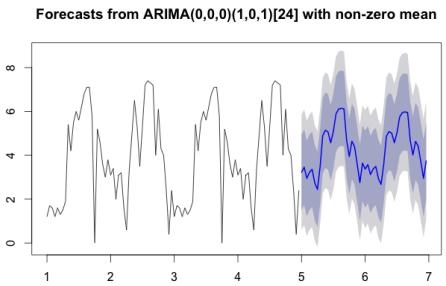
- Determinismo: las condiciones en las que se da el fenómeno determinan completamente su resultado.
- Indeterminismo: se admite la existencia de fenómenos en los que no es posible saber de antemano cuál será su resultado, aún controlando las condiciones en las que se desarrolla.

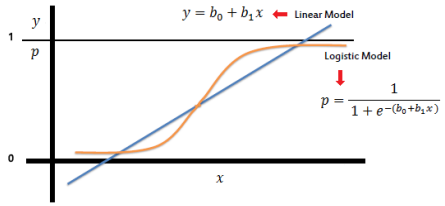
P. S. Laplace

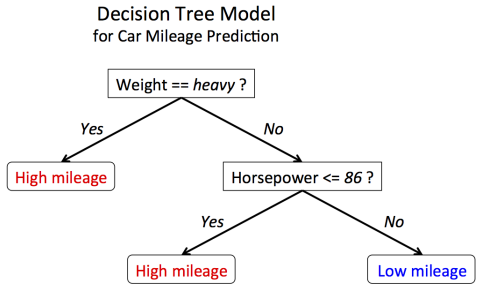
Si fuera posible conocer todas las leyes que rigen todos los procesos que ocurren en el universo y el estado actual de todos los objetos que lo componen además, si fuera posible analizar tales datos, se podría abarcar en una sola fórmula los movimientos de todos los objetos en el universo; nada resultaría incierto y tanto el futuro como el pasado estarían presentes ante nuestros ojos.

Modelación matemática

Desarrollo de expresiones matemáticas para describir, en algún sentido, la relación que existe entre un conjunto de variables que describen las condiciones en que se desarrolla un fenómeno y una variable asociada su resultado.







Modelos deterministas

En el caso determinista, las entradas determinan completamente la salida:

Si $x_1, \ldots, x_p$ son las entradas y y la salida o respuesta, podemos modelar matemáticamente el proceso como:

$$y = f(x_1, \ldots, x_p)$$

El objetivo es determinar una buena elección para f, se pueden utilizar distintos criterios para elegir a un candidato.

Ejemplo. Caída libre

Cuenta la historia que Galileo Galilei descubrió experimentalmente que la distancia que recorre un cuerpo en caída libre, despreciando la resistencia del aire, está dada por:

$$d(t) = \frac{1}{2}gt^2$$

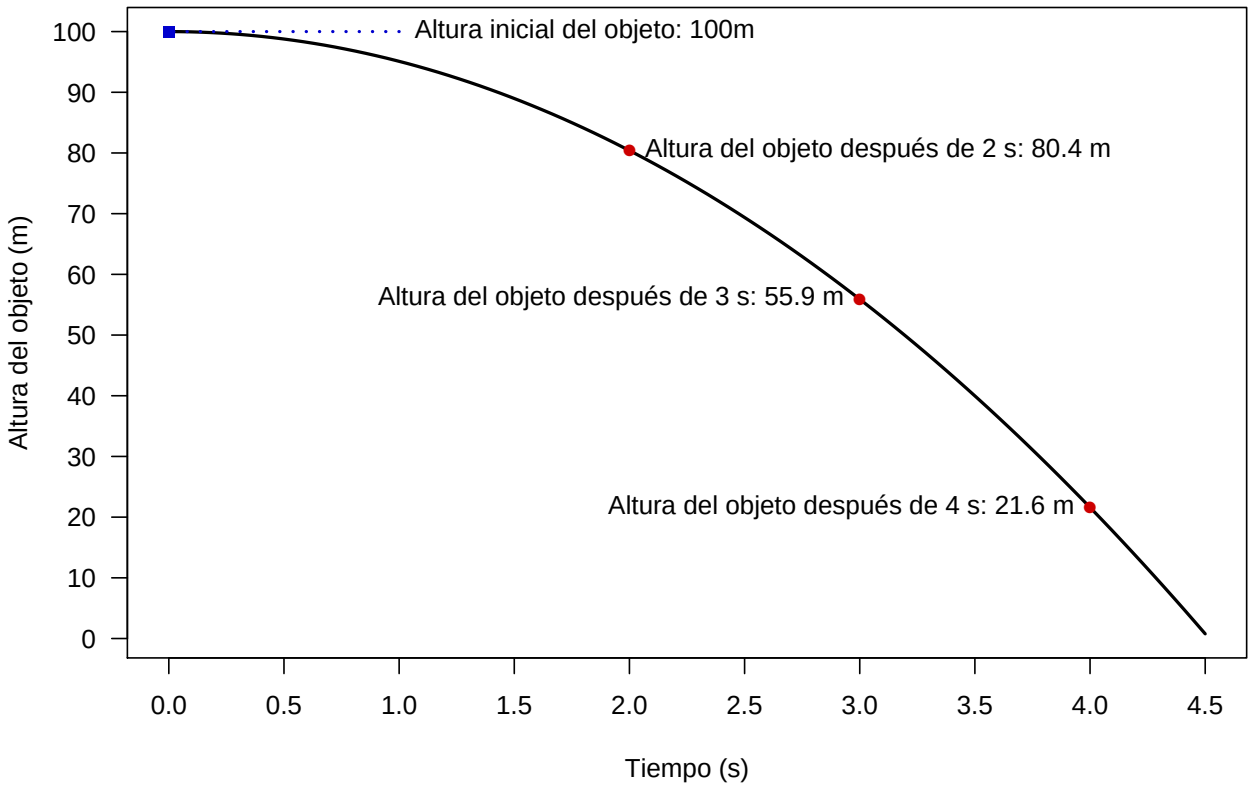
donde:

- t tiempo transcurrido,
- $g = 9.8ms^{-2}$ (aceleración de la gravedad),
- $d(t)$ distancia recorrida al tiempo t.

Caída libre (cont.)

Ejemplo

Se deja caer un objeto desde 100 m altura. ¿Cuál será la altura del objeto después de $t = 2, 3, 4$ s?



Crecimiento poblacional

- El crecimiento exponencial se puede utilizar para modelar el crecimiento o decrecimiento de una población.
- Si el crecimiento/decrecimiento se da a una tasa proporcional al tamaño de la población, entonces se puede modelar el tamaño de la población al tiempo t como

$$N(t) = N_0 \exp^{-\lambda t}$$

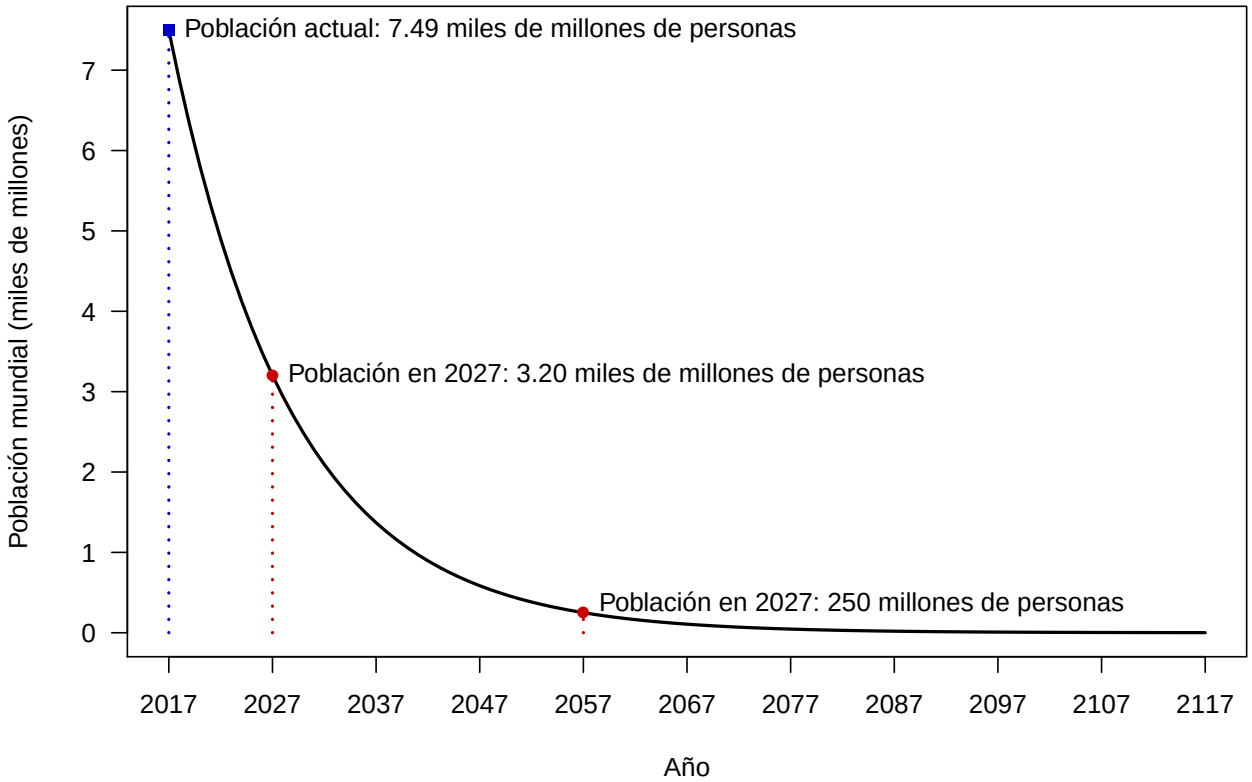
donde:

  - N_0 es la población inicial (tiempo $t = 0$),
  - t es el tiempo transcurrido,
  - λ es la tasa de crecimiento/decrecimiento,
  - $N(t)$ es la población al tiempo t.

Crecimiento poblacional (cont.)

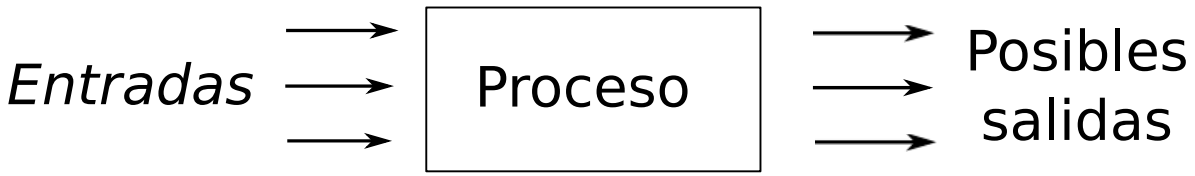
Ejemplo

La población de mundial el 02/02/2017 es aproximadamente 7,494 millones de personas. Si en la población se detuvieran los nacimientos y la tasa se mortalidad se mantuviera constante en 8.5 % anual:



Modelos Estocásticos

En el caso estocástico, el objetivo es modelar matemáticamente la relación entre las condiciones en que se desarrolla un fenómeno y la incertidumbre de sus posibles respuestas.



Si $x_1, \ldots, x_p$ son las entradas y y la salida o respuesta, podemos modelar matemáticamente el proceso como:

$$y \sim F(x_1, \ldots, x_p)$$

El objetivo es encontrar un buen modelo probabilista F, estamos ante un problema típico de inferencia estadística.

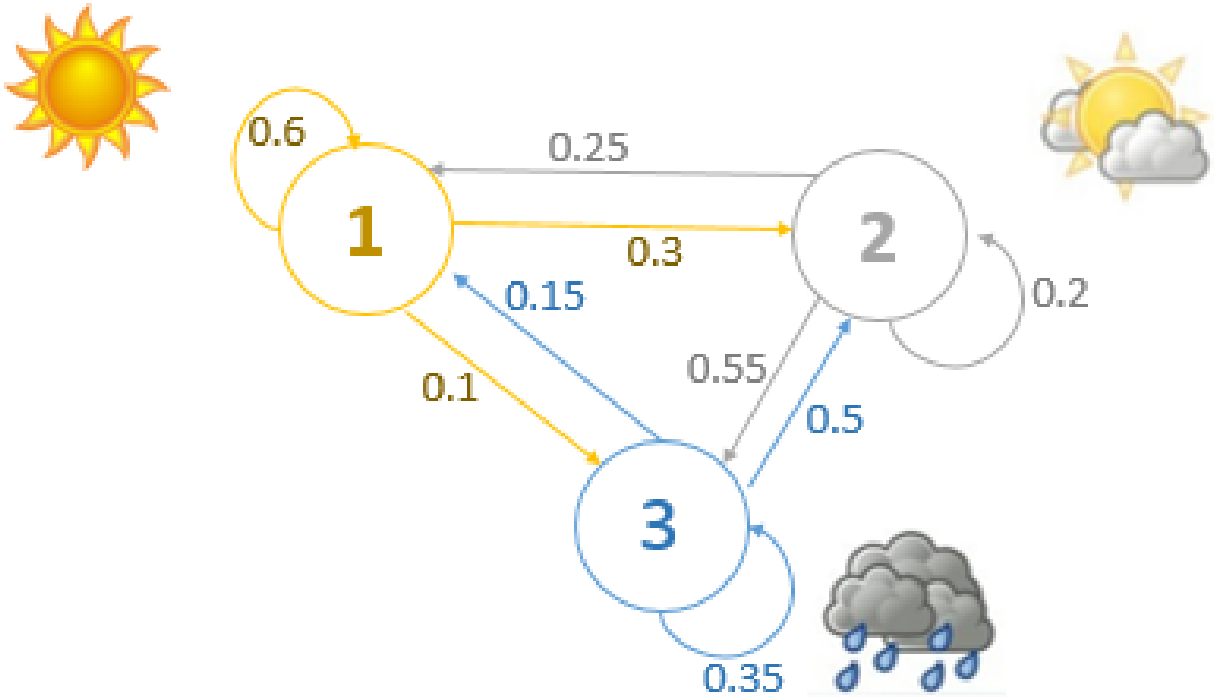
Ejemplo. El estado del tiempo

Es posible modelar el estado del tiempo como una *cadena de Markov*. Consideremos tres estados posibles $S = \{1, 2, 3\}$ donde: 1 es soleado, 2 es nublado, 3 es lluvioso. Suponemos que la matriz de transición P está dada por:

$$P = \begin{bmatrix} 0.60 & 0.30 & 0.10 \\ 0.25 & 0.20 & 0.55 \\ 0.15 & 0.50 & 0.35 \end{bmatrix}$$

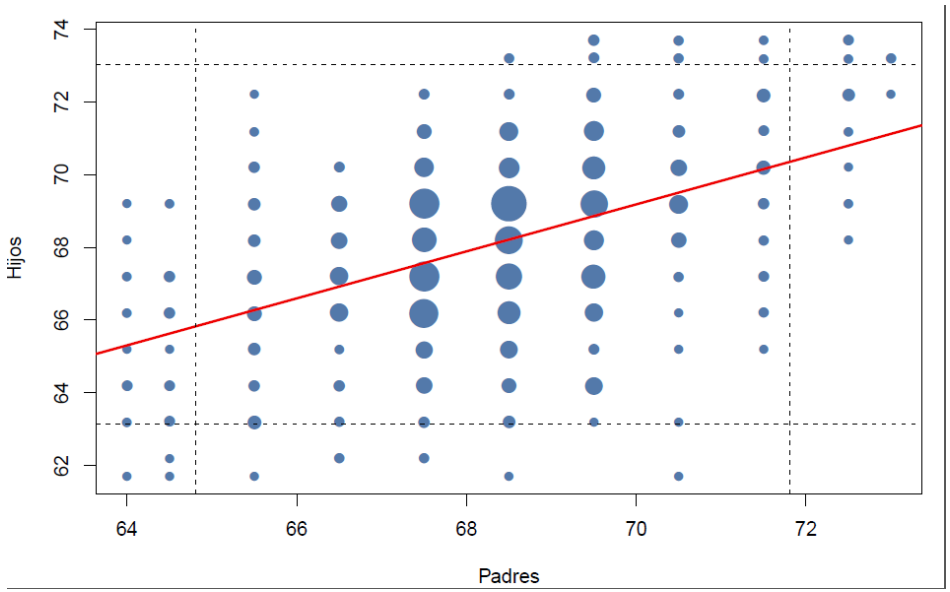
La i, j-ésima entrada de P representa la probabilidad de que el tiempo de mañana sea j dado que hoy es i, $i, j = 1, 2, 3$. Por ejemplo:

$$P(soleado|soleado) = 0.60$$
$$P(nublado|lluvioso) = 0.50$$
$$P(lluvioso|soleado) = 0.10$$


0.6
0.25
1
2
0.3
0.15
0.2
0.55
0.1
0.5
3
0.35

Ejemplo. Regresión de Galton

Francis Galton en el siglo XIX observó la relación que había entre la altura de padres e hijos, notó que aunque existe una tendencia de que un padre alto tenga un hijo alto y padre bajo tenga un hijo bajo, la distribución de las estaturas, no cambia drásticamente de una generación a otra (a este fenómeno se le conoce como regresión a la media).



En resumen..

Los modelos matemáticos se usan para describir la relación funcional entre las variables de entrada $(x_1, x_2, \ldots, x_p)$ y la variable respuesta y .

$$Y = f(x_1, x_2, \ldots, x_p) \qquad \text{(Determinista)}$$
$$Y \sim F(x_1, x_2, \ldots, x_p) \qquad \text{(Probabilistico)}$$

- ¿Qué función f (o F) tomar?
- ¿El modelo es adecuado? ¿Existe el mejor modelo?
- ¿Qué supuestos debemos tomar en cuenta?
- ¿Las variables x_i explican a y?

Los modelos de regresión

Nuestro interés se centrará en los modelos probabilistas y en específico en el modelo de regresión lineal.

$$E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p$$

Donde $\mathbf{X} = (X_1, \ldots, X_p)$.

- El término *regresión* viene la regresión a la media del ejemplo de Galton. El término *lineal* se refiere a que la forma funcional del modelo es lineal en los parámetros β.
- Cuando sólo se tiene una variable de entrada, $p = 1$, el modelo es de regresión lineal simple, cuando hay más de una variable de entrada el modelo es de regresión lineal múltiple.

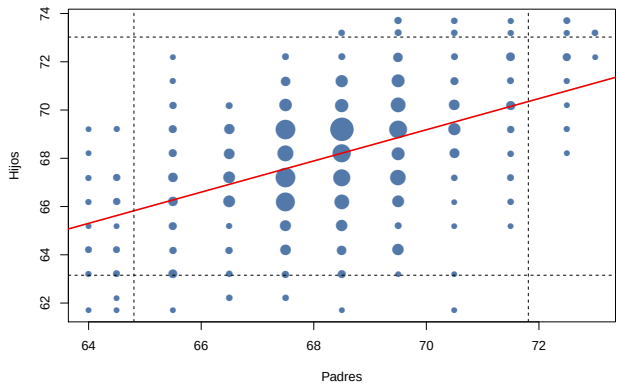
El modelo de regresión lineal simple

En el ejemplo de Galton es posible utilizar un modelo de regresión lineal simple para explicar las alturas de los hijos a partir de las alturas de los padres.

$$E(Y|X) = \beta_0 + \beta_1 X$$

donde:

- Y es la altura del hijo,
- X es la altura del padre,
- β_0 y β_1 son los parámetros del modelo.



El modelo de regresión lineal simple (cont.)

- Podemos notar como para un valor dado de X, no hay un valor único de Y, sino que Y varía alrededor de su valor promedio y lo hace aleatoriamente.
- Denotamos por ϵ al error estadístico que representa la desviación de Y de su media, para un valor de X fijo.

Entonces podemos representar el modelo de regresión lineal simple como sigue

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Lo usual es asumir $\epsilon \sim (0, \sigma^2)$, de manera que

$$Y|X \sim (\beta_0 + \beta_1 X, \sigma^2).$$

- El supuesto anterior es muy fuerte dado que asume una varianza constante para cualquier valor de X.

Interpretación del modelo

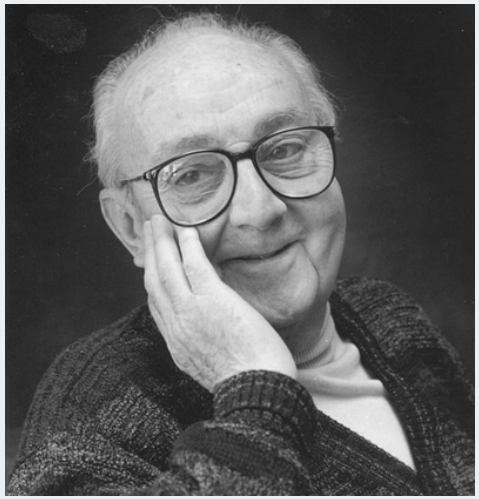
- El objetivo es hacer inferencias sobre los parámetros desconocidos β_0, β_1 y σ^2.
- Para un valor de X fijo, el valor esperado de Y es $\beta_0 + \beta_1 X$.
- El parámetro β_1 representa el incremento en el valor esperado de Y por un incremento unitario en X.
- Para un valor de X fijo, σ^2 es la variablidad de Y alrededor de su valor esperado, es decir, σ^2 es la varianza de Y dado X.
- ¿Qué hay de la interpretación de β_0? Las ecuaciones de regresión solamente son válidas dentro del rango de los valores observados en las entradas.

Aplicaciones de los modelos de regresión

1. Las ventas de un producto pueden ser predichas a partir del gasto en publicidad.
2. El desempeño de un trabajador en un empleo puede ser predicho a partir de las respuestas de una prueba de aptitudes.
3. El tamaño del vocabulario de un niño puede ser predicho a partir de la edad del niño y el grado de escolaridad de sus padres.
4. El salario de un trabajador puede ser predicho a partir de su edad, escolaridad y sector de ocupación.
5. El volumen de madera de un árbol puede ser determinado a partir de variables como el diámetro del tronco y la altura.

George Box

Essentially, all models are wrong, but some are useful.



Unidad 2

Regresion lineal simple

Planteamiento

El modelo de regresión lineal simple relaciona la variable de interés Y, llamada dependiente, con la variable explicativa X. A través de la ecuación

$$E(Y|X) = \beta_0 + \beta_1 X$$

Si consideramos las desviaciones de Y de su media $E(Y|X)$ como errores aleatorios $\epsilon \sim (0, \sigma^2)$, podemos escribir el modelo como

$$Y = \beta_0 + \beta_1 X + \epsilon$$

El objetivo es estimar los parámetros del modelo: β_0, β_1 y σ^2 a partir de un conjunto de observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Estimación

- Como primera aproximación, consideremos el caso en que Y no varia alrededor de su media.
- En este caso, los puntos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ caen exactamente sobre una línea recta.
- Podemos tomar cualquier par de puntos distintos (x_i, y_i) y (x_j, y_j), $i \neq j$, para estimar los parámetros como

$$\hat{\beta}_1 = \frac{y_i - y_j}{x_i - x_j} \qquad \text{y} \qquad \hat{\beta}_0 = y_i - \hat{\beta}_1 x_i$$

- Sin embargo, al considerar que Y no varía, estamos omitiendo la aleatoriedad, que es nuestro principal interés.
- Si Y varía alrededor de su media, por cada par de puntos (x_i, y_i) y (x_j, y_j), $i \neq j$ tendríamos una estimación distinta de β_0 y β_1.
- Entonces, ¿cuál es la mejor estimación que podemos hacer?

Estimación (cont.)

- Si los puntos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ no son colineales, podemos calcular la desviación de y_i con respecto a su media estimada $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_0 x_i$, para $\hat{\beta}_0$ y $\hat{\beta}_1$ dados:

$$e_i := y_i - \hat{y}_i = y_i - \hat{\beta}_0 + \hat{\beta}_0 x_i$$

- Las cantidades $e_1, e_2, \ldots, e_n$ son conocidos como *residuos* y están asociados a estimaciones de β_0 y β_1.
- De alguna forma los residuos nos dicen qué tan bien estamos estimando a los parámetros del modelo. Si los residuos son *grandes*, la estimación es *mala* y si los residuos son *chicos*, la estimación es *buena*.
- ¿Qué criterio utilizar para decidir si los residuos son grandes?
- Dado el criterio anterior, ¿cómo hacer la *mejor* estimación posible?

Minimos Cuadrados (MCO)

- El método de MCO propone utilizar la función *suma de cuadrados de los residuos*

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_0 x_i)^2$$

- $Q(\beta_0, \beta_1)$ es una *medida* de qué tan buena es la estimación del valor esperado de Y con la recta $\beta_0 + \beta_1 X$. Este es un criterio para decidir si los residuales son grandes.
- El método de MCO propone estimar los parámetros con los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ tales que $Q(\beta_0, \beta_1)$ es mínima.

A los estimadores encontrados bajo esta metodología se les conoce como *estimadores de mínimos cuadrados*.

Resultados teóricos

Definiciones

Sea $f : D \subset \mathbb{R}^n \to \mathbb{R}$ y $z \in D$.

1. Si existe una vecindad $N_z \subset D$ tal que $f(x) \leq f(z)$, para todo $x \in N_z$, se dice que f tiene en z un máximo local.
2. Si existe una vecindad $N_z \subset D$ tal que $f(x) \geq f(z)$, para todo $x \in N_z$, se dice que f tiene en z un mínimo local.
3. Si $f(x) \leq f(z)$, para todo $x \in D$, se dice que f tiene un máximo absoluto en z.
4. Si $f(x) \geq f(z)$, para todo $x \in D$, se dice que f tiene un mínimo absoluto en z.

Resultados teóricos (cont.)

Definición (puntos críticos)

Sea $f : D \subset \mathbb{R}^n \to \mathbb{R}$, con derivadas parciales en $D' \subset D$ tal que

$$\frac{\partial f(x)}{\partial x_i}\Big|_{x=z} = 0, \qquad i = 1, \ldots, n.$$

con $z \in D'$. Se dice que z es un punto critico (o estacionario) de f.

Teorema (condiciones necesarias)

Sea f como en la definición anterior. Si f tiene un valor extremo en $z \in D'$, entonces z es un punto crítico de f.

El reciproco del teorema anterior no es verdadero, pero se pueden dar condiciones en las que en punto crítico una función tiene un valor extremo.

Resultados teóricos (cont.)

Definición (matriz hessiana)

Sea $f : D \subset \mathbb{R}^n \to \mathbb{R}$ con segundas derivadas parciales en $D' \subset D$. Se define la *matriz hessiana* de f como

$$H(x) = \left\{ \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right\}, \qquad i, j = 1, \ldots, n.$$

Teorema (criterio de las segundas derivadas parciales)

Sea f como en la definición anterior y $z \in D'$ un punto critico de f. Entonces:

1. Si $H(z)$ en positiva definida, entonces f tiene un mínimo en z.
2. Si $H(z)$ es negativa definida, entonces f tiene un máximo en z.
3. Si $H(z)$ es indefinida, entonces f tiene un punto silla (*saddle point*) en z.

Definición

Sea A una matriz real simétrica de dimensión n. Si para todo x vector real (columna) de dimensión n y $x \neq 0$:

1. $x'Ax > 0$, entonces se dice que A es positiva definida;
2. $x'Ax < 0$, entonces A es negativa definida;

Si $x'Ax$ es positivo para algunos valores de x y negativo para otros, se dice que A es indefinida.

Proposición

Sea $A = [a_{ij}]$ una matriz real simétrica de dimensión 2.

1. Si $a_{11} > 0$ y $det(A) > 0$, entonces A es positiva definida.
2. Si $a_{11} < 0$ y $det(A) > 0$, entonces A es negativa definida.
3. En otro caso, A es indefinida.

Solución analítica

En el problema de minimizar $Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$, es fácil mostrar que:

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = -2n\bar{y}_n + 2n\beta_0 + 2n\beta_1 \bar{x}_n$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i y_i + 2n\beta_0 \bar{x}_n + 2\beta_1 \sum_{i=1}^{n} xi^2$$

con $\bar{x}_n = n^{-1}\sum_{i=1}^{n} x_i$ y $\bar{y}_n = n^{-1}\sum_{i=1}^{n} y_i$. Para encontrar los puntos críticos de $Q(\beta_0, \beta_1)$ se debe resolver el sistema de ecuaciones:

$$\beta_0 + \bar{x}_n \beta_1 = \bar{y}_n$$

$$n\bar{x}_n \beta_0 + \beta_1 \left(\sum_{i=1}^{n} x_i^2 \right) = \sum_{i=1}^{n} x_i y_i$$

Solución analítica (cont.)

Se puede mostrar que la solución al sistema anterior es:

$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n \qquad \text{y} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

con $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ y $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)$.
Para determinar si Q tiene un mínimo en $(\hat{\beta}_0, \hat{\beta}_1)$ se aplica el criterio de las segundas derivadas parciales. Nuevamente, es fácil mostrar que

$$H(\beta_0, \beta_1) = \begin{bmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

Se debe observar que $H(\beta_0, \beta_1)$ es constante para β_0 y β_1.

Solución analítica (cont.)

$H_{11} = 2n > 0$ y

$$\det(H) = 4n \sum_{i=1}^{n} x_i^2 - 4 \left(\sum_{i=1}^{n} x_i \right)^2 > 0$$

Donde se concluye que $Q(\beta_0, \beta_1)$ tiene un mínimo en el punto

$$(\hat{\beta}_0, \hat{\beta}_1) = \left(\bar{y}_n - \frac{S_{xy}}{S_{xx}} \bar{x}_n, \frac{S_{xy}}{S_{xx}} \right)$$

Estimación de la varianza

El método de MCO no proporciona una estimación σ^2 pero, una estimación razonable es

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Entonces, los estimadores de MCO del modelo RLS son

$$\hat{\beta}_0 = \bar{y}_n - \frac{S_{xy}}{S_{xx}}\bar{x}_n$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

El modelo RLS

1. El valor esperado Y_i, condicional al valor de la variable explicativa X, está dado por:
$$E(Y|X_i) = \beta_0 + \beta_1 X_i, \qquad i = 1, \ldots, n.$$
2. Si suponemos que Y, dado X, varía aleatoriamente alrededor de su valor esperado, podemos representar tal desviación como $\epsilon_i|X_i \sim (0, \sigma^2)$. Entonces, podemos escribir el modelo como
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, \ldots, n.$$
3. Además, suponemos que condicional a los respectivos valores de X_i y X_j, la variación de Y_i alrededor de su media no está correlacionada con la respectiva variación de Y_j, es decir, $Cov(Y_i.Y_j|X_i, X_j) = 0$, $i, j = 1, \ldots, n$ e $i \neq j$.

El modelo RLS (cont.)

A parir de la igualdad $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, es fácil deducir las siguientes propiedades de Y_i:

- $V(Y_i|X) = \sigma^2$, $i = 1, \ldots, n$.
- $Cov(Y_i, Y_j) = 0$, $i, j = 1, \ldots, n$ e $i \neq j$.

También es importante señalar que los estimadores de MCO son combinaciones lineales de las $Y's$:

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \left(\frac{x_i - \bar{x}_n}{S_{xx}} \right) \bar{x}_n \right) y_i$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}_n}{S_{xx}} \right) y_i$$

Propiedades de los Estimadores de MCO

Las supuestos anteriores y la linealidad de los estimadores de MCO nos permiten mostrar el siguiente resultado.

Proposición

Los estimadores de MCO de β_0 y β_1 en el modelo de MCO son insesgados, es decir:

$$E(\hat{\beta}_0|\mathbf{X}) = \beta_0 \qquad \text{y} \qquad E(\hat{\beta}_1|\mathbf{X}) = \beta_1$$

Además, la varianza de los estimadores es:

$$V(\hat{\beta}_0|\mathbf{X}) = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2$$

$$V(\hat{\beta}_1|\mathbf{X}) = \frac{\sigma^2}{S_{xx}}$$

Propiedades de los estimadores de MCO (cont.)

Teorema (Gauss-Markov)

Bajo los supuestos del modelo RLS, los estimadores de MCO de β_0 y β_1 son los Mejores Estimadores Lineales Insesgados (MELI, o BLUE, *Best Linear Unbiased Estimators*). Es decir, para cualesquiera $\tilde{\beta}_0$ y $\tilde{\beta}_1$ estimadores lineales insesgados de β_0 y β_1 se tiene

$$V(\hat{\beta}_0) \leq V(\tilde{\beta}_0) \qquad \text{y} \qquad V(\hat{\beta}_1) \leq V(\tilde{\beta}_1)$$

Limitaciones y alcance de los EMCO

Con los resultados obtenidos hasta el momento, somos capaces de ajustar un modelo RLS y usarlo para hacer predicciones de valores futuros. Además, el TGM nos garantiza que los estimadores MCO son los MELI. Pero:

1. ¿Cómo hacer estimación por intervalos?
2. ¿Cómo hacer pruebas de hipótesis?
3. ¿Cómo cuantificar el error de nuestras predicciones?
4. ¿Cómo saber si el modelo está ajustando bien?

Para dar una solución aceptable a los planteamientos anteriores debemos incluir un supuesto adicional sobre la distribución de los errores, sobre la forma en que Y varía alrededor de su valor esperado.

Modelo RLS con errores normales

A los supuestos anteriores del modelo RLS agregamos ahora el supuesto que la distribución conjunta de los errores, desviaciones de Y con respecto a su valor esperado, tienen una distribución normal conjunta $\boldsymbol{\epsilon} \sim N_n\left(\mathbf{0}, \sigma^2 I_n\right)$.

Proposición,

Con el supuesto de normalidad conjunta en los errores, se cumple que:

- ϵ_i es independiente de ϵ_j, $i, j = 1, \ldots, n$ y $i \neq j$.
- $Y_i | X_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$
- $Y_1, \ldots, Y_n$ son independientes , pero no son idénticamente distribuidos.

Función de verosimilitud

Con el supuesto de normalidad podemos expresar la verosimilitud de $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$ como:

$$L(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \left\{ \exp -\frac{1}{2\sigma^2} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

$$= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}$$

$$\ell(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Estimación por Máxima Verosimilitud

Al maximizar la log-verosimilitud usando técnicas de cálculo de varias variables

$$\left(\frac{\partial \ell}{\partial \beta_0}, \frac{\partial \ell}{\partial \beta_1}, \frac{\partial \ell}{\partial \sigma^2} \right) = (0, 0, 0)$$

Nos lleva a resolver las siguientes ecuaciones:

$$n\beta_0 + n\beta_1 \bar{x}_n = n\bar{y}_n$$

$$n\beta_0 \bar{x}_n + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$n\sigma^2 - \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

Estimación por Máxima Verosimilitud (cont.)

Las primeras dos ecuaciones, de donde concluimos que los EMV son los de MCO:

$$\hat{\beta}_0 = \bar{y}_n - \frac{S_{xy}}{S_{xx}}\bar{x}_n \qquad \text{y} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

De la tercera ecuación despejamos el estimador para σ^2:

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

Donde e_i es el i-ésimo residuo del modelo.

Propiedades de los EMV

- Como los EMV de β_0 y β_1 coinciden con los de MCO, sabemos que son los MELI y

$$V(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}}\right)\sigma^2 \qquad \text{y} \qquad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

- Es fácil mostrar que

$$E\hat{\sigma}^2_{MV} = \frac{(n-2)}{n}\sigma^2$$

Por lo que el estimador $\hat{\sigma}^2_{MCO} = \frac{n}{n-2}\hat{\sigma}^2_{MV}$ sí es insesgado, lo que justifica su elección.

Intervalos de confianza

Dado que $\hat{\beta}_0$ y $\hat{\beta}_0$ son combinaciones lineales de v.a. normal es independientes, ambos siguen una distribución normal y de resultados anteriores conocemos su media y varianza:

$$\hat{\beta}_0 \sim N\left(\beta_0, (n^{-1} + \bar{x}_n^2 S_{xx}^{-1})\sigma^2\right) \qquad \hat{\beta}_1 \sim N\left(\beta_1, S_{xx}^{-1}\sigma^2\right)$$

Los estadísticos

$$T_i = \frac{\hat{\beta}_i - \beta_i}{EE(\hat{\beta}_i)}, \qquad i = 0, 1.$$

con $EE(\hat{\beta}_i) = \sqrt{V(\hat{\beta}_i)}$, son cantidades pivotales para β_0 y β_1, respectivamente, ya que $T_i \sim N(0, 1)$. Por lo que pueden ser utilizadas para construir intervalos de confianza. El problema es que dependen de σ^2.

Resultados auxiliares

Definición (distribución t de Student)

Una variable aleatoria continua X tiene distribución t de Student con paramétros μ, λ y α, denotado por $X \sim Stu(\mu, \lambda, \alpha)$ si su función de densidad está dada por

$$f(x) = \frac{\Gamma\left(\frac{\alpha+1}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \left(\frac{\lambda}{\alpha\pi}\right)^{\frac{1}{2}} \left[1 + \frac{\lambda(x-\mu)^2}{\alpha}\right]^{-\frac{\alpha+1}{2}}, \quad -\infty < x < \infty$$

con $-\infty < \mu < \infty$, $\lambda > 0$ y $\alpha > 0$. μ es el parámetro de localidad, λ es el parámetro de escala y α es el parámetro de forma.

Cuando $\mu = 0$, $\lambda = 1$ y α es entero y positivo, $Stu(0, 1, n)$ es la distribución t con n grados de libertad de los cursos de estadística básica.

Resultados auxiliares (cont.)

Proposición

En el modelo de RLS con errores normales se cumple:

1. $(n-2)\hat{\sigma}^2_{MCO}/\sigma^2 \sim \chi^2_{n-2}$.
2. $\hat{\beta}_0, \hat{\beta}_1 \perp \hat{\sigma}^2$.

Proposición

Si $X \sim N(0,1)$, $Y \sim \chi^2_n$ y $X \perp Y$, entonces

$$T = \frac{X}{\sqrt{Y/n}} \sim Stu(0,1,n)$$

Resultados auxiliares (cont.)

Proposición

Si $\hat{EE}(\hat{\beta}_i)$ se obtiene de $EE(\hat{\beta}_i)$ al reemplazar σ^2 por $\hat{\sigma}^2_{MCO}$, entonces

$$T_i^* = \frac{\hat{\beta}_i - \beta_i}{\hat{EE}(\hat{\beta}_i)} \sim Stu(0, 1, n-2), \qquad i = 0, 1.$$

Los estadísticos T_i^* no dependen de los parámetros, por lo que podemos utilizarlos como pivotales para construir intervalos de confianza para β_i, $i = 0, 1$.

Intervalos de confianza para β_0 y β_1

Si $t_{n-2}^{\alpha/2}$ representa el cuantil superior $\alpha/2$, $\alpha \in (0, 0.5)$, de la distribución $Stu(0, 1, n-2)$, es decir

$$P\left(X > t_{n-2}^{\alpha/2}\right) = \frac{\alpha}{2}, \qquad \text{con } X \sim Stu(0, 1, n-2);$$

entonces

$$P\left(-t_{n-2}^{\alpha/2} \leq \frac{\hat{\beta}_i - \beta_i}{\hat{EE}(\hat{\beta}_i)} \leq t_{n-2}^{\alpha/2}\right) = 1 - \alpha, \qquad i = 0, 1.$$

Por lo tanto un intervalo de confianza $100(1-\alpha)\,\%$ para β_i es:

$$\left(\hat{\beta}_i - t_{n-2}^{\alpha/2}\hat{EE}(\hat{\beta}_i), \hat{\beta}_i + t_{n-2}^{\alpha/2}\hat{EE}(\hat{\beta}_i)\right), \qquad i = 0, 1.$$

Intervalo de confianza para σ^2

En el caso de la varianza σ^2, la cantidad pivotal es

$$\frac{(n-2)\hat{\sigma}^2_{MCO}}{\sigma^2} \sim \chi^2_n.$$

Si γ_1^α y γ_2^α son tales que

$$P\left(\gamma_1^\alpha \leq X \leq \gamma_2^\alpha\right) = 1-\alpha, \qquad \text{con } X \sim \chi^2_{n-2},$$

entonces un intervalo de confianza $100(1-\alpha)\,\%$ para σ^2 es

$$\left(\frac{(n-2)\hat{\sigma}^2_{MCO}}{\gamma_2^\alpha}, \frac{(n-2)\hat{\sigma}^2_{MCO}}{\gamma_1^\alpha}\right)$$

Intervalo de confianza para σ^2 (cont.)

- En el caso de los intervalos de confianza para β_0 y β_1, al tomar tomar los cuantiles $-t_{n-2}^{\alpha/2}$ y $t_{n-2}^{\alpha/2}$, se garantiza que los intervalos tienen longitud mínima, debido a que la distribución $Stu(0, 1, n-2)$ es simétrica.
- Sin embargo, en el caso de la varianza σ^2, la distribución χ^2_{n-2} es asimétrica, por lo que al tomar $\gamma_1^\alpha = \chi^2_{n-2}(\alpha/2)$ y $\gamma_2^\alpha = \chi^2_{n-2}(1-\alpha/2)$, no se garantiza que el intervalo tenga longitud mínima.
- Para cada caso en particular, α y n, se pueden encontrar numéricamente los valores de γ_1^α y γ_2^α para los que la longitud del intervalo es mínima.

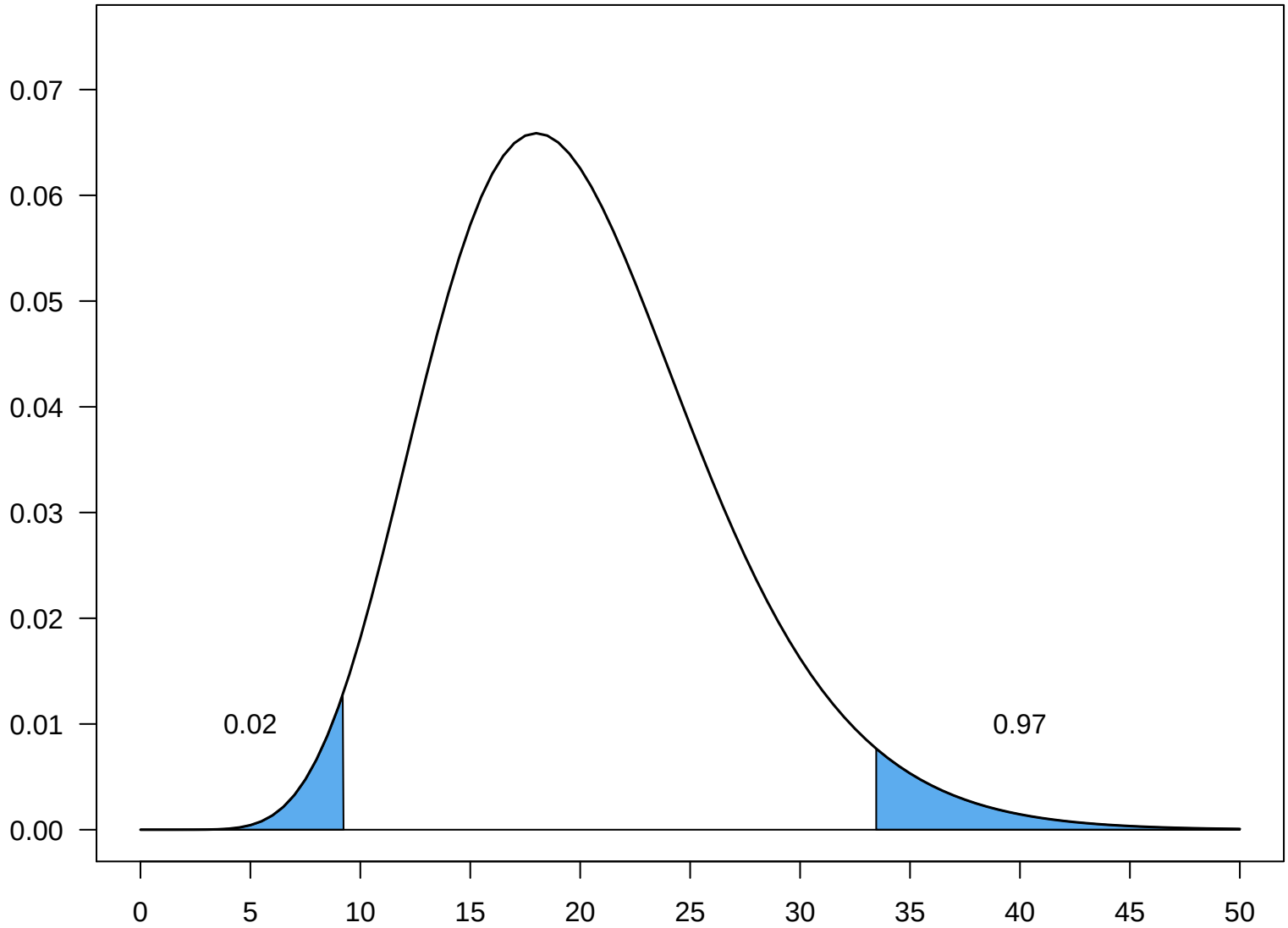


Figura: Densidad χ^2_{50}.

Propiedades de $\hat{\sigma^2}_{MCO}$ y $\hat{\sigma^2}_{MV}$

A partir del resultado

$$\frac{(n-2)\,\hat{\sigma}^2_{MCO}}{\sigma^2} \sim \chi^2_{n-2}$$

se pueden calcular fácilmente los momentos de $\hat{\sigma}^2_{MCO}$ y $\hat{\sigma}^2_{MV}$. Basta recordar que si $X \sim \chi^2_n$ entonces $E(X) = n$ y $V(X) = 2n$. Entonces:

$$E\left(\hat{\sigma}^2_{MV}\right) = \frac{\sigma^2}{n} E\left(\frac{n\hat{\sigma}^2_{MV}}{\sigma^2}\right) = \frac{n-2}{n}\sigma^2$$

$$E\left(\hat{\sigma}^2_{MCO}\right) = \frac{\sigma^2}{n-2} E\left(\frac{(n-2)\hat{\sigma}^2_{MCO}}{\sigma^2}\right) = \sigma^2$$

De igual manera se puede mostrar que

$$V\left(\hat{\sigma}^2_{MV}\right) = 2(n-2)n^{-2}\sigma^4 \quad \text{y} \quad V\left(\hat{\sigma}^2_{MCO}\right) = 2(n-2)^{-1}\sigma^4.$$

Intervalos de confianza para el valor esperado de Y

- También es posible calcular intervalos de confianza para el valor esperado de y para un valor dado de x, al que denotaremos por μ_x.
- Por las propiedades de $\hat{\beta}_0$ y $\hat{\beta}_1$, $\hat{\mu}_x = \hat{\beta}_0 + \hat{\beta}_1 x$ es un estimador insesgado de μ_x, además por ser combinación lineal de las y_i tiene una distribución normal.
- Es fácil mostrar que:

$$V\left(\hat{\mu}_x\right) = \left(\frac{1}{n} + \frac{\left(x - \bar{x}_n\right)^2}{S_{xx}}\right)\sigma^2 =: \sigma_x^2$$

- Entonces

$$\hat{\mu}_x \sim N\left(\mu_x, \sigma_x^2\right).$$

IC para el valor esperado de Y (cont.)

- De lo anterior se sigue que

$$\frac{\hat{\mu}_x - \mu_x}{\sqrt{\sigma_x^2}} \sim N(0,1)$$

- Si $\hat{\sigma}_x^2$ se obtiene de reemplazar σ_{MCO}^2 por σ^2 en σ_x^2, entonces

$$\frac{\hat{\mu}_x - \mu_x}{\sqrt{\hat{\sigma}_x^2}} \sim Stu(0,1,n-2)$$

- Luego, un intervalo de confianza $100(1-\alpha)\,\%$ para μ_x está dado por

$$\left(\hat{\mu}_x - t_{n-2}^{\alpha/2}\sqrt{\hat{\sigma}_x^2}, \hat{\mu}_x - t_{n-2}^{\alpha/2}\sqrt{\hat{\sigma}_x^2}\right)$$

Intervalos de predicción

- También podemos hacer inferencias sobre nuevas observaciones de Y para un valor de X dado, que denotaremos por Y_x
- Como Y_x es una variable aleatoria utilizamos predictores e intervalos de predicción.
- Recordemos que $Y_x \sim N(\beta_0 + \beta_1 X, \sigma^2)$.
- Si los parámetros del modelo fueran conocidos, un predictor puntual de Y_x es $\beta_0 + \beta_1 x$ y un intervalo de predicción para Y_x estaría dado por

$$\left(\beta_0 + \beta_1 x - z^{\alpha/2}\sigma, \beta_0 + \beta_1 x + z^{\alpha/2}\sigma\right)$$

  donde $z^{\alpha/2}$ es el cuantil superior $\alpha/2$ de una distribución $N(0,1)$.
- Como los parámetros del modelo RLS son desconocidos, debemos estimarlos, esto hace que la varianza del la predicción crezca.

Intervalos de predicción (cont.)

- La varianza de la predicción tiene dos componentes: una debida a la estimación de los parámetros del modelo y otra debida a la variabilidad de Y_x.

$$V\left(\hat{Y}_x\right) = V\left(\hat{\mu}_x\right) + V\left(Y_x\right) = \left(1 + \frac{1}{n} + \frac{\left(x - \bar{x}_n\right)^2}{S_{xx}}\right)\sigma^2$$

- Utilizamos como pivotal a

$$\frac{\hat{\mu}_x - Y_x}{\sqrt{\hat{V}(Y_x)}} \sim Stu\left(0, 1, n-2\right)$$

- Luego, un intervalo de predicción para Y_x es

$$\left(\hat{\mu}_x - t_{n-2}^{\alpha/2}\sqrt{\hat{V}(Y_x)}, \hat{\mu}_x + t_{n-2}^{\alpha/2}\sqrt{\hat{V}(Y_x)}\right)$$

Intervalos simultáneos

- Los intervalos de confianza y predicción que se mostraron anteriormente son individuales.
- Las conclusiones que se hagan sobre varios intervalos simultáneamente no tienen necesariamente la misma confianza.
- Se debe hacer algún ajuste en la construcción para obtener una significancia conjunta dada.
- Existen dos métodos: Bonferroni y Working-Hottelling-Scheffé ambos con propiedades diferentes.

Intervalos simultáneos (continuación))

Método de Bonferroni

- Se basa en el principio de inclusión exclusión.
- Para construir k intervalos simultáneos propone usar $t_{n-2}^{\alpha/2k}$ en lugar de $t_{n-2}^{\alpha/2}$.
- Es recomendable para valores de k pequeños.

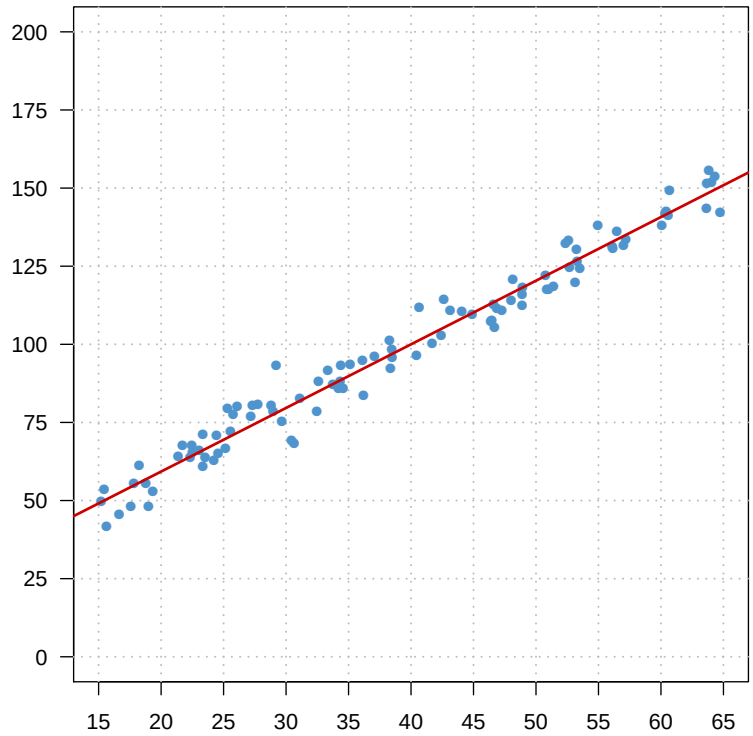
Método de Working-Hotelling-Scheffé

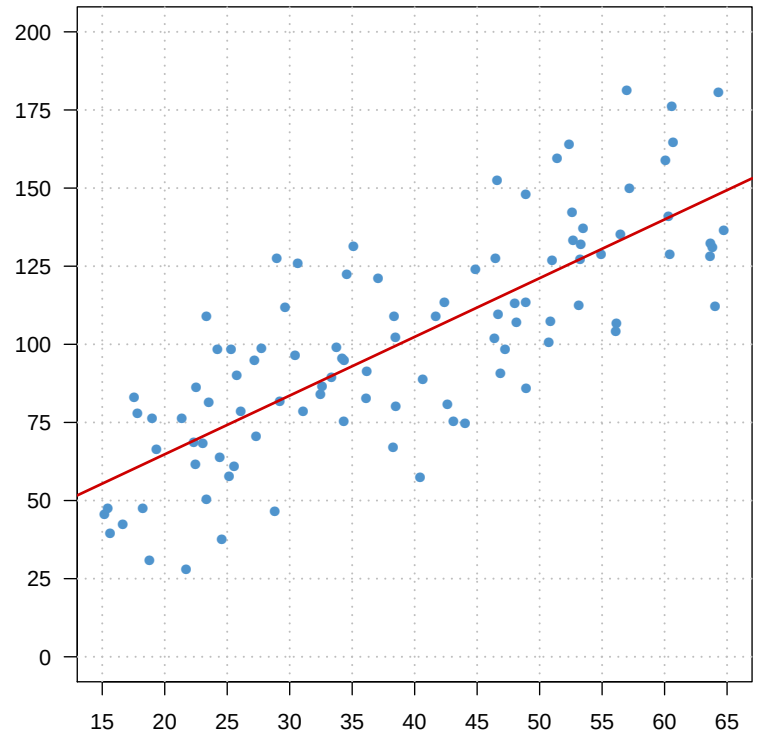
- Se basa en la distribución conjunta de $\hat{\beta}_0$ y $\hat{\beta}_1$ y utiliza un criterio de optimización.
- Es recomendable para construir muchos intervalos simultáneos.
- Para cualquier número de intervalos propone utilizar $2F_{2,n-2}^{\alpha}$ en lugar de $t_{n-2}^{\alpha/2}$.

Pruebas de hipótesis

- Con los intervalos de confianza anteriores podemos probar hipótesis sobre los parámetros del modelo RLS.
- Una de las hipótesis más importantes es $H_0 : \beta_1 = 0$. Recordemos que el modelo RLS establece que el valor esperado de Y depende de X. Si H_0 es cierta, significa que no la media de Y no se ve afectada por X.
- La hipótesis $H_0 : \beta_0 = 0$ no tiene una interpretación tan relevante como la hipótesis anterior, sin embargo, puede servir para determinar si utilizar un modelo RLS con o sin intercepto.
- En general, las inferencias sobre σ^2 son de utilidad para realizar predicciones con el modelo. Recordemos que la amplitud de los intervalos de predicción son más amplios debido a la variabilidad intrinseca de Y. Si la varianza de Y es grande, las predicciones que se hagan no serán precisas.

Pruebas de hipótesis (cont.)





Pruebas de hipótesis para β_1

Recordamos que

$$\hat{\beta}_1 \sim N\left(\beta_1, S_{xx}^{-1}\sigma^2\right)$$

y que

$$T^* = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{S_{xx}^{-1}\hat{\sigma}^2_{MCO}}} \sim Stu\left(0, 1, n-2\right)$$

T^* así definida *no* es una estadística (Pues depende de β_1 que es desconocida) sin embargo al fijar β_1 en una prueba de hipótesis ya puede ser utilizada para construir la region de rechazo.

Pruebas de hipótesis para β_1 (cont.)

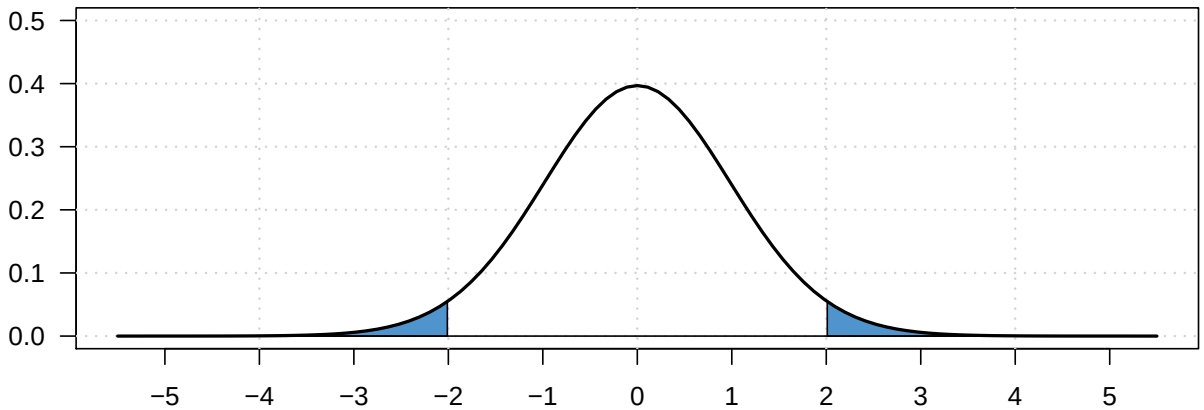
Para contrastar las hipótesis

$$H_0 : \beta_1 = b_1 \quad vs \quad H_1 : \beta_1 \neq b_1$$

La regla de decisión es: rechazar H_0 con una significancia α si

$$|t^*| > t_{n-2}^{\alpha/2}$$

donde: t^* es el valor de T^* con calculado con los datos observados y $\beta_1 = b_1$.



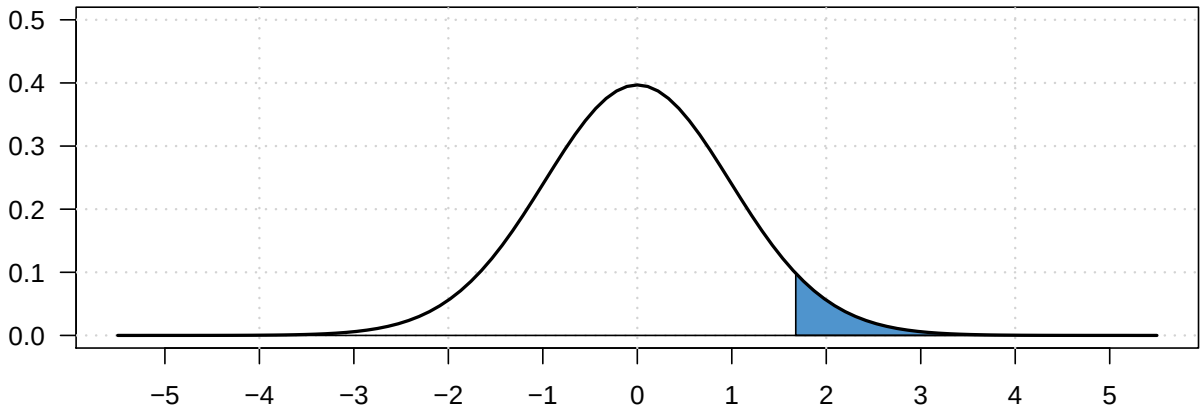
Pruebas de hipótesis para β_1 (cont.)

Para contrastar las hipótesis

$$H_0 : \beta_1 \leq b_1 \quad vs \quad H_1 : \beta_1 > b_1$$

La regla de decisión es: rechazar H_0 con una significancia α si

$$t > t_{n-2}^{\alpha}$$



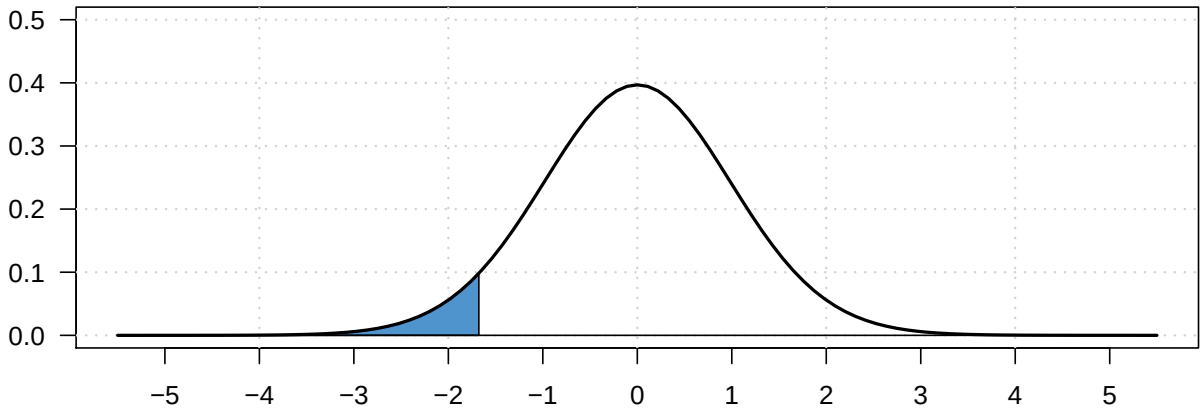
Pruebas de hipótesis para β_1 (cont.)

Para contrastar las hipótesis

$$H_0 : \beta_1 \geq b_1 \quad vs \quad H_1 : \beta_1 < b_1$$

La regla de decisión es: rechazar H_0 con una significancia α si

$$t < t_{n-2}^{1-\alpha}$$



Pruebas de hipótesis para β_0

Hipótesis	Región de rechazo
$H_0 : \beta_0 = b_0$ vs. $H_1 : \beta_0 \neq b_0$	$\lvert t^* \rvert > t_{n-2}^{\alpha/2}$
$H_0 : \beta_0 \leq b_0$ vs. $H_1 : \beta_0 > b_0$	$t^* > t_{n-2}^{\alpha}$
$H_0 : \beta_0 \geq b_0$ vs. $H_1 : \beta_0 < b_0$	$t^* < t_{n-2}^{1-\alpha}$

donde:

$$t^* = \frac{\hat{\beta}_0 - b_0}{\sqrt{\hat{\sigma}^2_{MCO}\left(\frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}}\right)}}$$

Pruebas de hipótesis para σ^2

Hipótesis	Región de rechazo
$H_0 : \sigma^2 = s$ vs $H_1 : \sigma^2 \neq s$	$t < \chi^{2\,(1-\alpha/2)}_{n-2}$ o $t > \chi^{2\,(\alpha/2)}_{n-2}$
$H_0 : \sigma^2 \leq s$ vs $H_1 : \sigma^2 > s$	$t > \chi^{2\,(\alpha)}_{n-2}$
$H_0 : \sigma^2 \geq s$ vs $H_1 : \sigma^2 < s$	$t < \chi^{2\,(1-\alpha)}_{n-2}$

donde:

$$t = \frac{(n-2)\hat{\sigma}^2_{MCO}}{s}$$

Prueba de razón de verosimilitudes para β_1

- La region de rechazo construida para la prueba $H_0 : \beta_1 = b_1 \quad vs \quad H_1 : \beta_1 \neq b_1$ se obtuvo a partir de la distribución de $\hat{\beta}_1$.
- Ahora se obtendrá una prueba a partir del cociente de verosimilitudes generalizadas.
- La prueba basada en el cociente de verosimilitudes nos indica rechazar H_0 si:

$$\Lambda = \frac{L_0}{L_1} = \frac{\sup_{\theta \in \Theta_0} L(\theta); \mathbf{x}, \mathbf{y})}{\sup_{\theta \in \Theta} L(\theta; \mathbf{x}, \mathbf{y})} < k$$

con k elegida para un nivel de significancia dado y

$$\Theta_0 = \left\{ \left(\beta_0, \beta_1, \sigma^2 \right) : \beta_0 \in \mathbb{R}, \beta_1 = b_1, 0 < \sigma^2 < \infty \right\}$$
$$\Theta = \left\{ \left(\beta_0, \beta_1, \sigma^2 \right) : \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}, 0 < \sigma^2 < \infty \right\}$$

Prueba de razón de verosimilitudes para β_1 (cont.)

- Para encontrar L_0 debemos encontrar los EMV bajo $H_0 : \beta_1 = b_1$.
- Bajo H_0 la verosimilitud es:

$$L(\theta_0; \mathbf{x}, \mathbf{y}) = (2\pi)^{-n/2} \left(\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - b_1 x_i)^2\right)$$

- Al maximizar con respecto a β_0 y σ^2 obtenemos:

$$\tilde{\beta}_0 = \bar{y}_n - b_1 \bar{x}_n \qquad \text{y} \qquad \tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \tilde{\beta}_0 - b_1 x_i\right)^2$$

- Entonces:

$$L_0 = (2\pi)^{-n/2} \left(\tilde{\sigma}^2\right)^{-n/2} e^{-n/2}$$

Prueba de razón de verosimilitudes para β_1 (cont.)

- L_1 es la verosimilitud del modelo RLS evaluada en los estimadores de máxima verosimilitud.
- Es fácil mostrar que

$$L_1 = (2\pi)^{-n/2} \left(\hat{\sigma}^2_{MV}\right)^{-n/2} e^{-n/2}$$

- Entonces:

$$\Lambda = \frac{(2\pi)^{-n/2} \left(\tilde{\sigma}^2\right)^{-n/2} e^{-n/2}}{(2\pi)^{-n/2} \left(\hat{\sigma}^2_{MV}\right)^{-n/2} e^{-n/2}} = \left(\frac{\hat{\sigma}^2_{MV}}{\tilde{\sigma}^2}\right)^{n/2}$$

- Por lo que la prueba de razón de verosimilitudes tiene región de rechazo

$$\frac{\hat{\sigma}^2_{MV}}{\tilde{\sigma}^2} < k$$

Prueba de razón de verosimilitudes para β_1 (cont.)

- Para determinar por completo la región de rechazo se debe elegir k de manera que la prueba cumpla con la significancia específicada.
- Para esto se debe trabajar un poco más con el cociente

$$\frac{\hat{\sigma}^2_{MV}}{\tilde{\sigma}^2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} \left(y_i - \tilde{\beta}_0 - b_1 x_i \right)^2}$$

- Es sencillo verificar que

$$\sum_{i=1}^{n} \left(y_i - \tilde{\beta}_0 - b_1 x_i \right)^2 = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 + b_1^2 S_{xx} - 2 b_1 S_{xy}$$

- También es fácil verificar que:

$$\sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Prueba de razón de verosimilitudes para β_1 (cont.)

- Una vez más, es fácil mostrar que:

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2 = \hat{\beta}_1^2 S_{xx}$$

- Al combinar los resultados anteriores obtenemos:

$$\frac{\hat{\sigma}^2_{MV}}{\tilde{\sigma}^2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + S_{xx} \left(\hat{\beta}_1 - b_1\right)^2}$$

- Entonces

$$\frac{\hat{\sigma}^2_{MV}}{\tilde{\sigma}^2} > k \quad \Leftrightarrow \quad \frac{S_{xx} \left(\hat{\beta}_1 - b_1\right)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} > c \quad \Leftrightarrow \quad \frac{\left|\hat{\beta}_1 - b_1\right|}{\sqrt{\hat{V}\left(\hat{\beta}_1\right)}} > c^*$$

Análisis de varianza (ANOVA)

- ¿Hay algún efecto de la variable X sobre el valor esperado de Y, es decir $\beta_1 = 0$?
- La prueba de razón de verosimilitudes nos da como región de rechazo:

$$\frac{\hat{\sigma}^2_{MV}}{\tilde{\sigma}^2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_n)^2} < k \quad \Leftrightarrow \quad \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} > c$$

- La igualdad

$$\sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

relaciona tres *sumas de cuadrados* llamadas: SC del total corregido por la media (SC_{TC}), SC de regresión (SC_{reg}) y SC residual o del error (SC_{error}).

ANOVA (cont.)

Proposición

En el modelo RLS con errores normales y bajo $H_0 : \beta_1 = 0$:

a) $SC_{reg}/\sigma^2 \sim \chi^2_1$.

b) $SC_{error}/\sigma^2 \sim \chi^2_{n-2}$.

c) $SC_{reg} \perp SC_{error}$.

Distribución F

Se dice que una variable aleatoria X tiene una distribución F de Snedecor con parámetros $n_1 > 0$ y $n_2 > 0$ si su densidad está dada por

$$f(x; n_1, n_2) = \frac{1}{Beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \left(\frac{n_1 x}{n_1 x + n_2}\right)^{\frac{n_1}{2}} \left(1 - \frac{n_1 x}{n_1 x + n_2}\right)^{\frac{n_2}{2}} x^{-1}$$

para $x > 0$. Lo anterior se denota como $X \sim F_{n_1, n_2}$. Los parámetros se llaman grados de libertad del numerador y denominador, respectivamente.

ANOVA (cont.)

Proposición

Si $X \sim \chi^2_{n_1}$, $Y \sim \chi^2_{n_2}$ y $X \perp Y$ entonces

$$\frac{X/n_1}{Y/n_2} \sim F_{n_1,n_2}$$

De lo anterior se sigue que

$$\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \propto F \sim F_{1,n-2}$$

de donde podemos especificar la prueba F que se usa en el análisis de varianza.

ANOVA (cont.)

Todo lo anterior se resume en la siguiente tabla, que recibe el nombre de *tabla ANOVA*:

FV	GL	SC	CM	F
Regresión	1	$\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2$	$\frac{SC_{reg}}{1}$	$\frac{CM_{reg}}{CM_{error}}$
Error	$n-2$	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$\frac{SC_{error}}{n-2}$	
Total	$n-1$	$\sum_{i=1}^{n} (y_i - \bar{y}_n)^2$		

- FV: fuente de variación.
- GL: grados de libertad.
- SC: suma de cuadrados.
- CM: cuadrado medio.
- F: cociente de cuadrados medios de regresión y del error.

Relación entre la prueba t y la prueba F

- ¿Qué relación tienen las pruebas t y F para contrastar la hipótesis $H_0 : \beta_1 = 0$.
- En el modelo RLS, no hay diferencias.
- La prueba de t para probar $H_0 : \beta_1 = 0$ rechaza si

$$\frac{\left|\hat{\beta}_1\right|}{\sqrt{\hat{V}\left(\hat{\beta}_1\right)}} > t_{n-2}^{\alpha/2} \quad \Leftrightarrow \quad \frac{S_{xx}\hat{\beta}_1^2}{SC_{error}/(n-2)} > \left(t_{n-2}^{\alpha/2}\right)^2$$

- Como tendrán que mostrar en la tarea

$$SC_{reg} = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}_n\right)^2 = S_{xx}\hat{\beta}_1^2$$

Relación entre la prueba t y la prueba F (cont.)

- Entonces la región de rechazo de la prueba de t es equivalente a

$$\frac{SC_{reg}}{SC_{error}/(n-2)} > \left(t_{n-2}^{\alpha/2}\right)^2$$

- El estadístico de prueba es entonces el mismo que el de la prueba F, ¿qué hay de las constantes que determinan la región de rechazo?
- De resultados anteriores es fácil mostrar que si $T \sim t_{n-1}$, entonces $T^2 \sim F_{1,n_2}$.
- En conclusión, en el modelo RLS las pruebas t y F son equivalentes para contrastar $H_0 : \beta_1 = 0$.

El coeficiente de determinación R^2

- Se define el coeficiente de determinación del modelo de regresión como

$$R^2 = \frac{SC_{reg}}{SC_{TC}} = 1 - \frac{SC_{error}}{SC_{TC}}$$

- El coeficiente R^2 y el cual sirve como una medida del ajuste del modelo.
- SC_{TC} es la variabilidad total de Y.
- SC_{error} es la variabilidad residual, es decir, lo que el modelo lo logra explicar.
- Entonces, R^2 es la proporción de la variabilidad total que se logra explicar con el modelo.

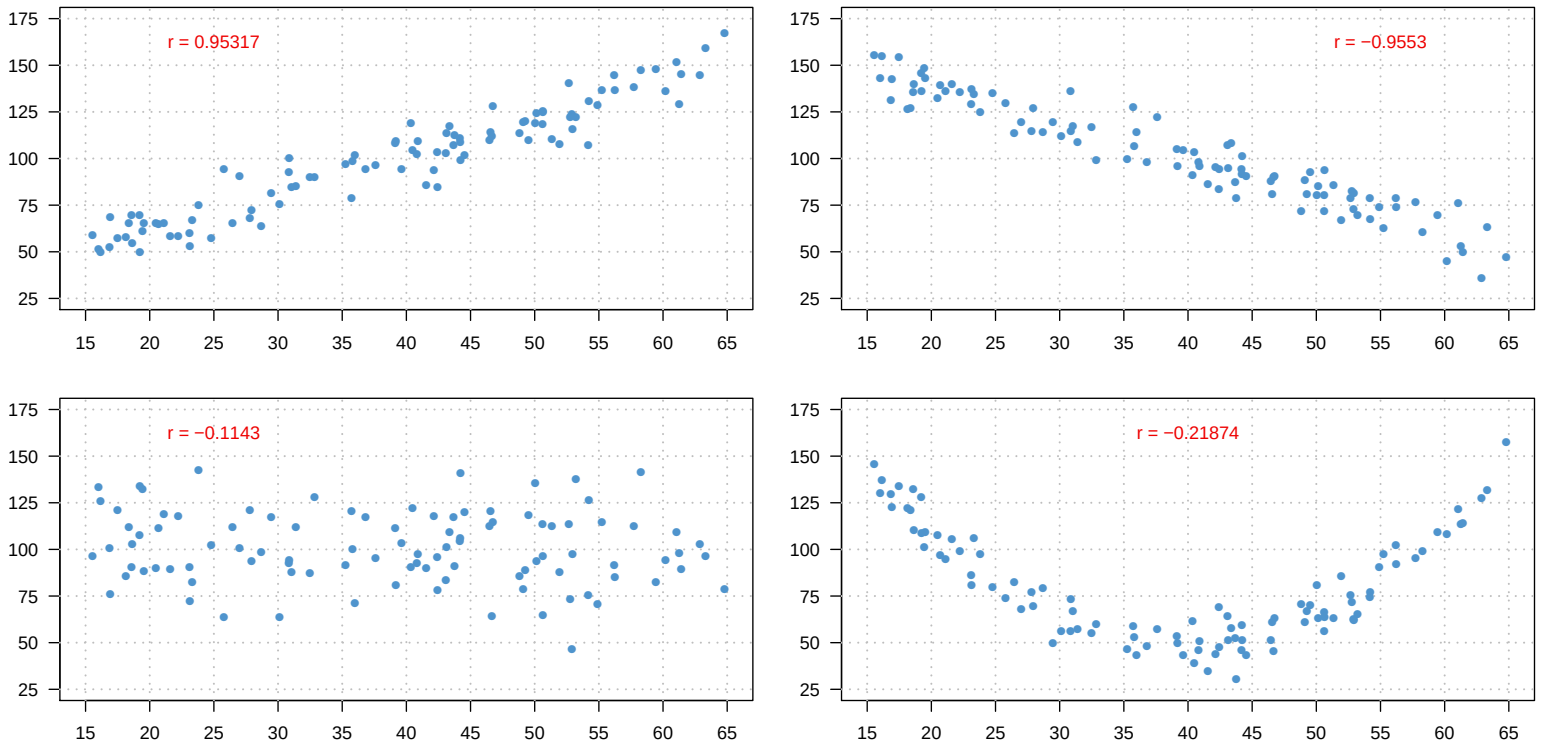
Relación del R^2 y la correlación de Pearson

- El coeficiente de correlación de Pearson entre $x_1, \ldots, x_n$ y $y_1, \ldots, y_n$ se define como

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

- r es una medida de asociación lineal entre las variables x y y.
- Se puede mostrar que $r \in (-1, 1)$.
- $r = 1$ indica una relación lineal directa $x - a \propto y$. $r = -1$ indica una relación lineal inversa $x - a \propto -y$. $r = 0$ indica que no hay una relación **lineal** entre las variables.
- Se puede mostrar que $r^2 = R^2$, donde R^2 es el coeficiente de determinación de la regresión de y sobre x.

Coeficiente de correlación de Pearson



Validación de los supuestos del modelo

- Por el Teorema de Gauss-Markov, las estimaciones puntuales son MELI.
- Los supuestos del teorema son:
  - Correcta especificación del modelo.
  - Errores no correlacionados.
  - Varianza constante.
- Los resultados estimación por intervalos y pruebas de hipótesis descansan sobre el supuesto de normalidad en los errores.

¡La validez de nuestras inferencias depende de que los supuestos se cumplan!

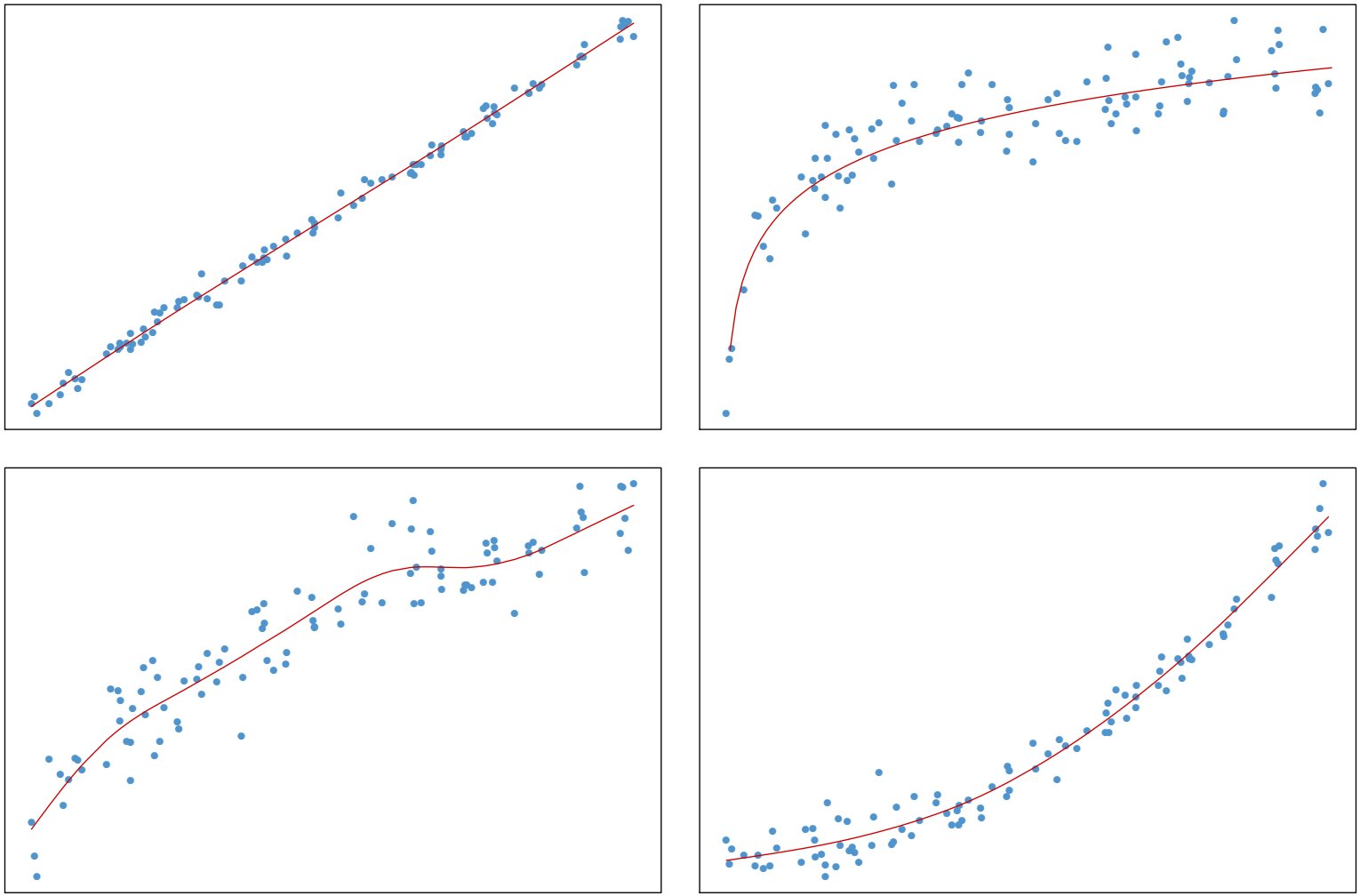
No linealidad en el modelo

- El modelo RLS establece que:

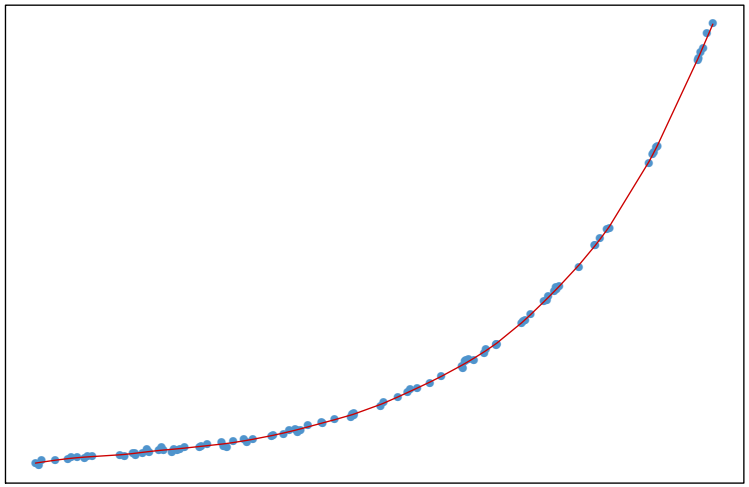
$$E(Y|X) = \beta_0 + \beta_1 X$$

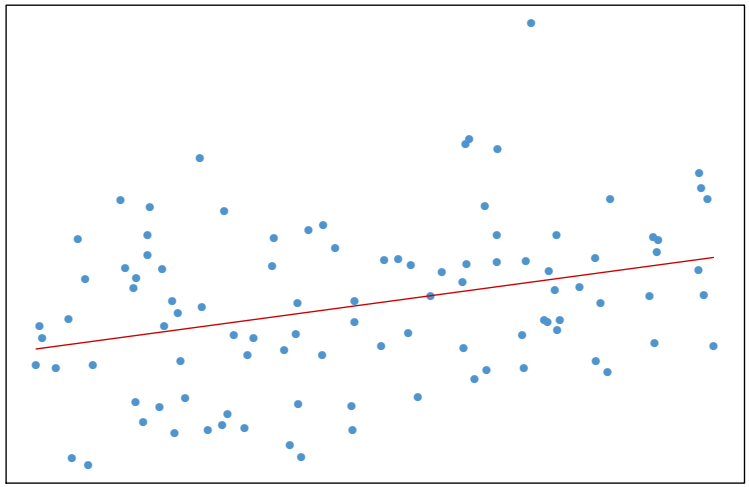
- Entonces, si hacemos un gráfico de dispersión de los datos, deberíamos poder observar que las observaciones se distribuyen aproximadamente alrededor de una recta.
- También se pueden detectar desviaciones al supuesto de linealidad a partir de la gráfica de residuos contra la variable explicativa.
- ¿Qué hacer si lo que observamos no es patrón lineal?
- Una solución es aplicar transformaciones a los datos, de acuerdo a los patrones observados.
  - $x^* = x^k$, con $k \neq 0$.
  - $x^* = log(x)$, si $x > 0$.
  - $y^* = log(y)$, si $x > 0$.
  - $y^* = \log(y)$ y $x^* = \log(x)$, si $x, y > 0$.
- Otras soluciones son ajustar un modelo polinomial o incorporar más variables al modelo.

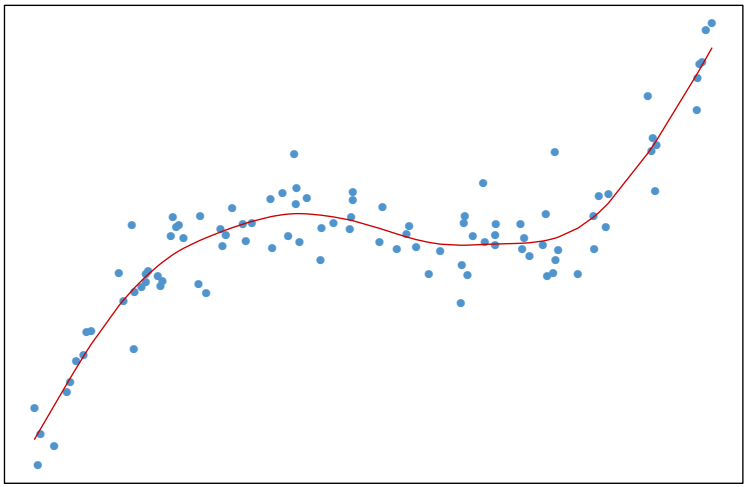
Ejemplos de no linealidad



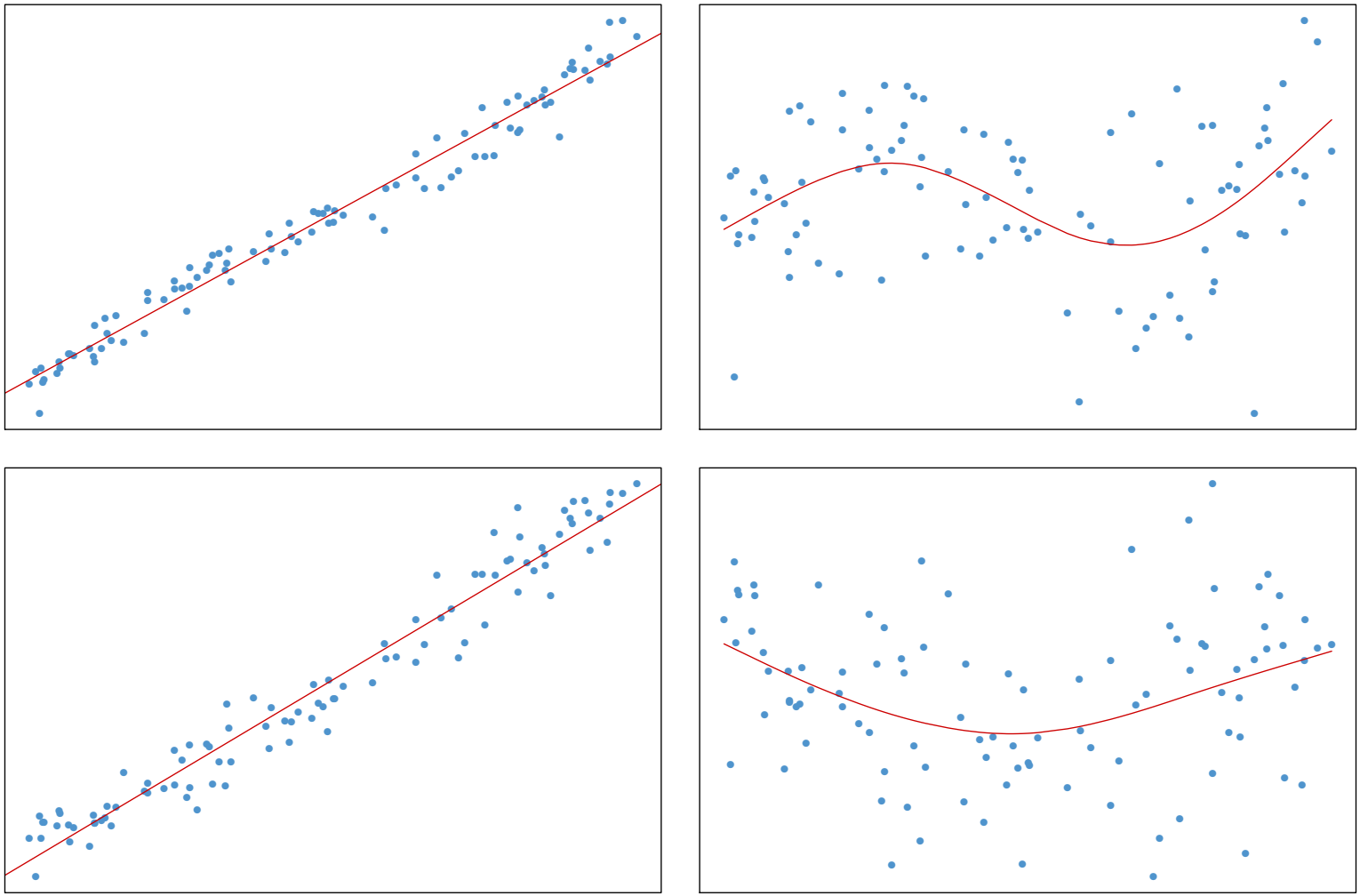
Ejemplos de no linealidad (cont.)







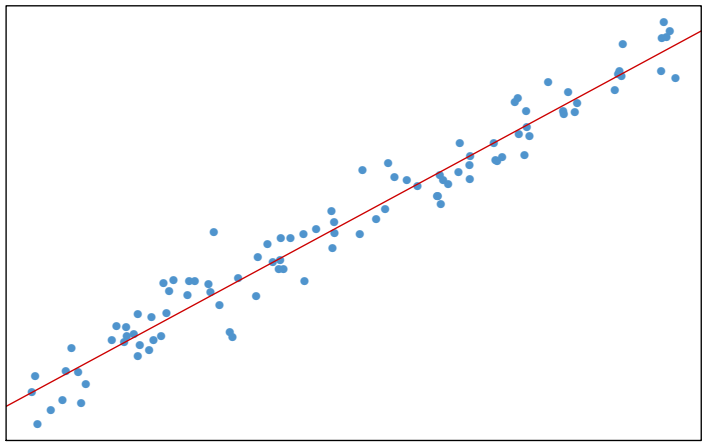
Ejemplos de no linealidad (cont.)

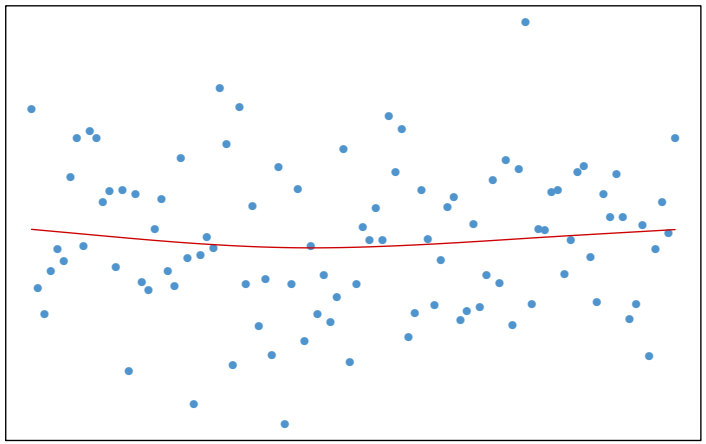


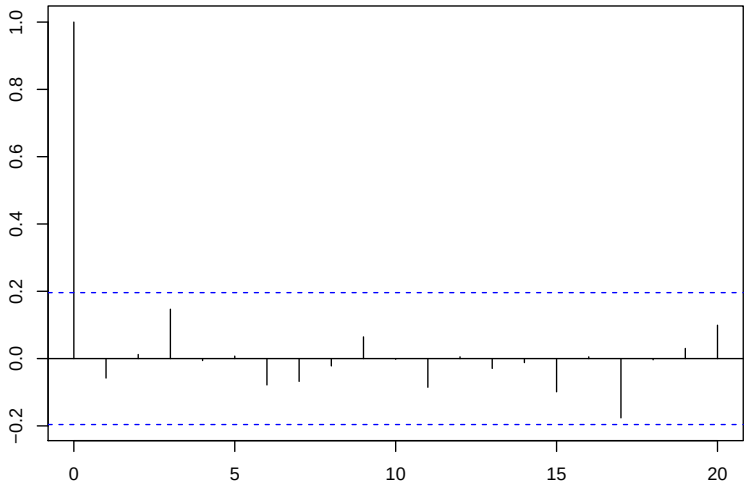
Errores no correlacionados

- El modelo RLS establece que $Cov\left(\epsilon_i, \epsilon_j\right) = 0$.
- Lo anterior implica que $Corr\left(\epsilon_i, \epsilon_j\right) = 0$.
- $e_i = y_i - \hat{y}_i$ es un *estimador* de ϵ_i.
- Se pueden utilizar los residuos $e_1, \ldots, e_n$ para verificar si este supuesto se cumple.
- La idea es básica es considerar los residuos ordenados según el número de observación, en el orden en que fueron obtenidos lo datos o bien, según el orden inducido por x.
- Se puede utilizar la función de autocorrelación para identificar alguna dependencia en los resiudos.
- También se puede utilizar una prueba no paramétrica, conocida como *prueba de rachas*, para identificar *rachas* en los signos de las residuos.

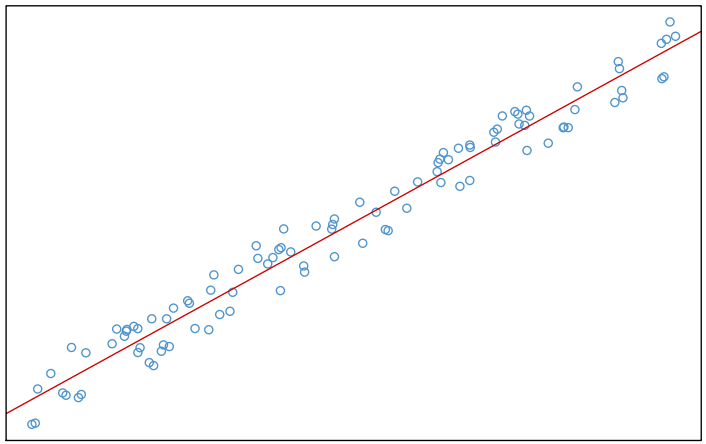
Ejemplos errores no correlacionados

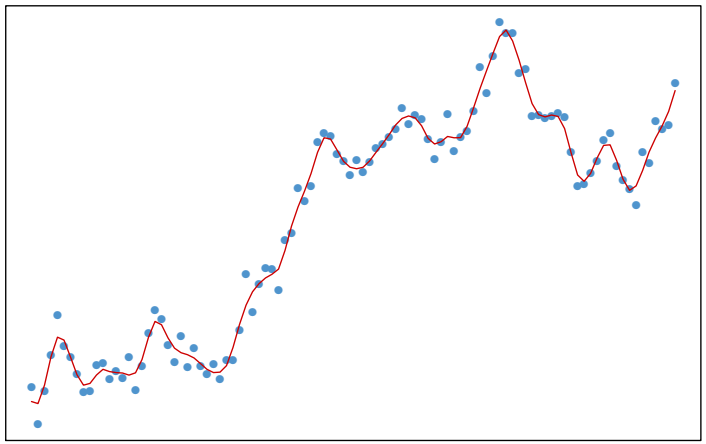


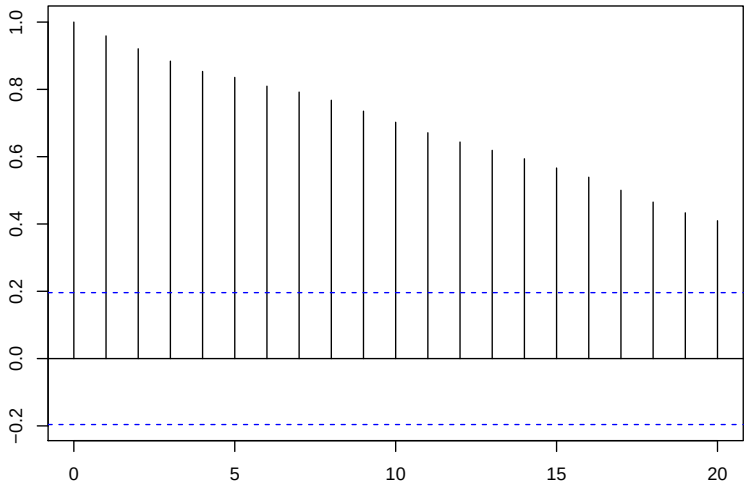




Ejemplos errores correlacionados







Autocorrelación

- La autocorrelación de una serie de datos discreta de un proceso X_t es más que el coeficiente de correlación de dicho proceso con una versión desplazada de la propia serie.
- La forma en como se calcula es la siguiente: suponiendo que tenemos a la serie de datos $(e_1, e_2, \ldots, e_n)$ entonces la función de autocorrelación con rezago k se obtiene como:

$$r_k = \frac{\sum_{i=1}^{n-k} (e_i - \bar{e})(e_{i-k} - \bar{e})}{\sum_{i=1}^{n} (e_i - \bar{e})^2}$$

- Si r_k se aleja del valor 0 hay evidencia de que cada k observaciones hay un patron de los residuales pues dichas observaciones están muy correlacionadas lo que implicaría dependencia de los residuales.
- En `R` la función `acf` se usa para obtener la autocorrelación de una serie de observaciones.

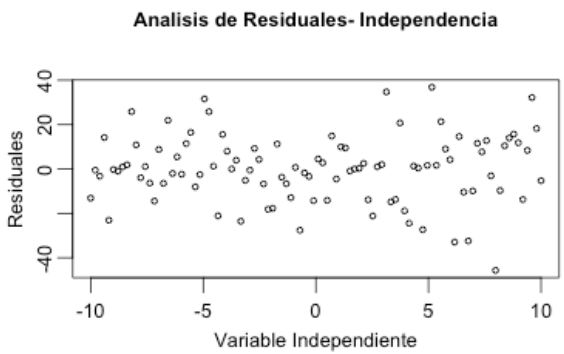
Rachas

- Otra forma que tenemos de verificar que los residuales son independientes es probar la aleatoreidad con la que van cambiando de signo los errores, si los residuales son independientes se esperaría que el cambio de signo del residual conforme se va obteniendo la muestra es aleatorio.
- Se considera una Racha de tamaño k a la secuencia de k de valores consecutivos de un mismo signo siempre y cuando estos sean precedidos y seguidos por valores con signo opuesto a la de la Racha.
- Por ejemplo, la cadena $++-+-$ tiene sólo una racha de $+$. La cadena $++---+---+--++++$ tiene distintas rachas de $+$ y $-$.
- La idea de la prueba es contar el número de rachas en la muestra, luego un número reducido o grande de rachas es indicio de que las observaciones no se han obtenido de forma aleatoria.
- Si la muestra es grande y la hipótesis de aleatoriedad es cierta la distribución muestral del número de rachas R, puede aproximarse mediante una distribución Normal de parámetros:

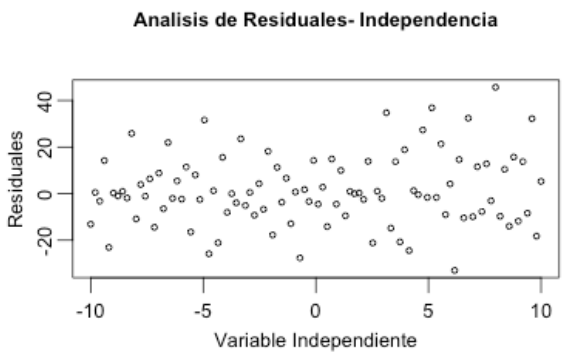
$$\mu_R = \frac{2n_1n_2}{n} \quad \sigma_R^2 = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}$$

Ejemplos de errores con rachas

La prueba de Rachas se encuentra programada de R y se encuentra dentro de la librería *tseries* bajo el nombre *runs.test*, esta función recibe un vector con valores binarios indicando el signo del residual.
Ejemplo:

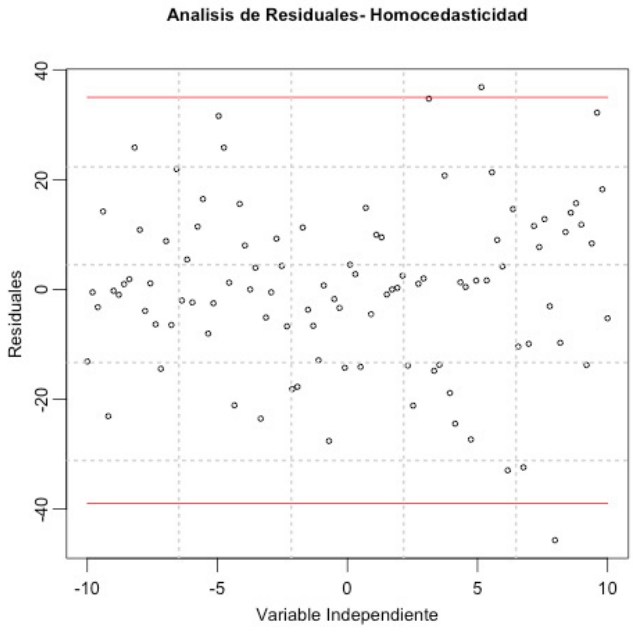


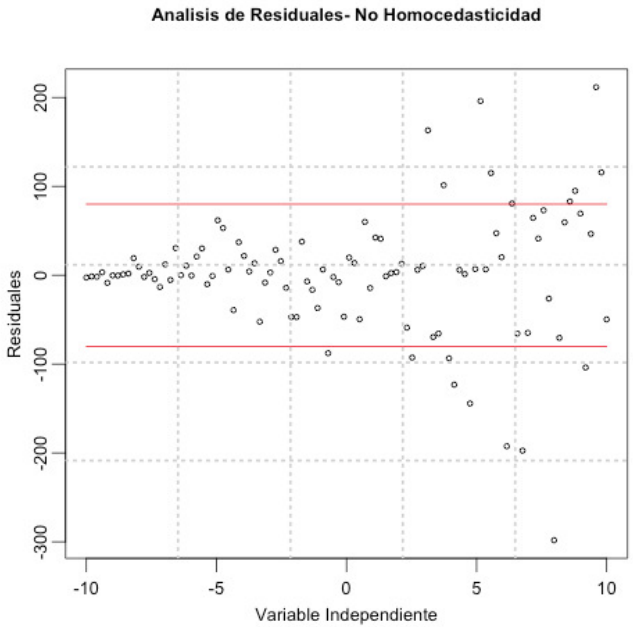

```
library(tseries)
runs.test(as.factor(res>0))
Runs Test
data: as.factor(res > 0)
Standard Normal = 0.004, p-value = 0.9968
alternative hypothesis: two.sided
```



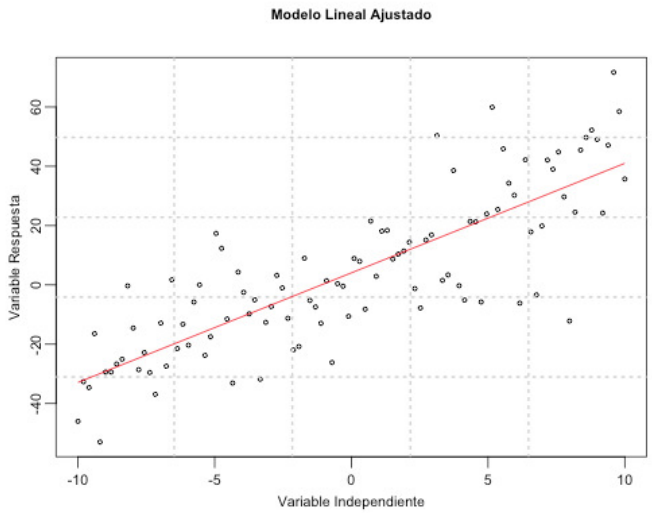

```
library(tseries)
runs.test(as.factor(res2>0))
Runs Test
data: as.factor(res2 > 0)
Standard Normal = 9.8499, p-value < 2.2e-16
alternative hypothesis: two.sided
```

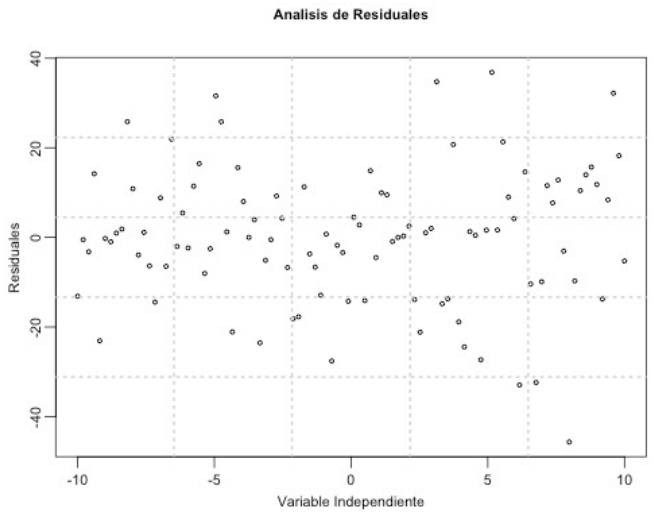
Homocedasticidad

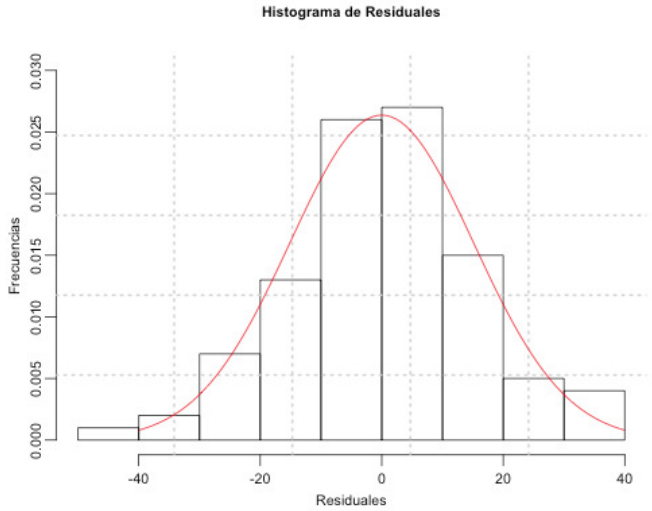


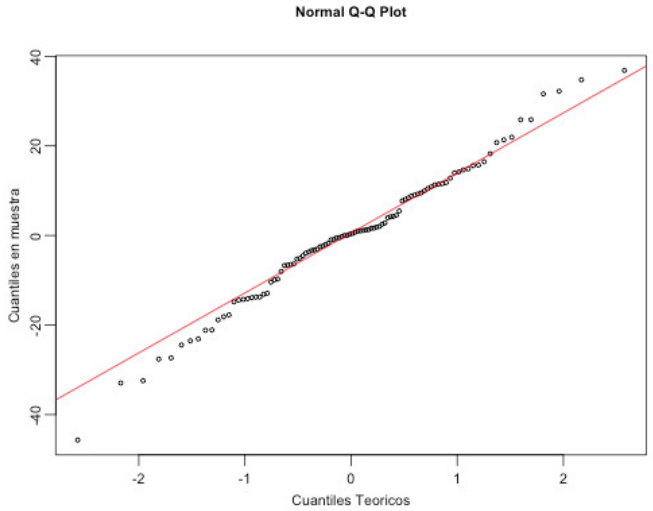


Normalidad


Modelo Lineal Ajustado
Variable Respuesta
Variable Independiente
-40
-20
0
20
40
60
-10
-5
0
5
10


Analisis de Residuales
Residuales
Variable Independiente
-40
-20
0
20
40
-10
-5
0
5
10


Histograma de Residuales
Frecuencias
Residuales
0.000
0.005
0.010
0.015
0.020
0.025
0.030
-40
-20
0
20
40


Normal Q-Q Plot
Cuantiles en muestra
Cuantiles Teoricos
-40
-20
0
20
40
-2
-1
0
1
2

Unidad 3

Conceptos de álgebra lineal y probabilidad

Definición (Matriz)

Una **matriz** *es un arreglo (bidimensional) de números o elementos algebraicos.*

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}_{n \times m}$$

Se dice que la matriz es cuadrada cuando $n = m$.
El espacio de las matrices con entradas de números reales se le denota como $\mathbb{R}^{(n \times m)}$, por lo que aveces se denota $\mathbf{A} \in \mathbb{R}^{(n \times m)}$, o bien $\left(a_{ij}\right) \in \mathbb{R}^{(n \times m)}$
Ejemplos:

- Matriz Identidad:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}_{n \times n}$$

Breve introducción al algebra matricial: Matrices

Operaciones Básicas

- **Suma de Matrices.** Sea $\mathbf{A}$ y $\mathbf{B} \in \mathbb{R}^{(n\times m)}$ se define la suma de matrices $\mathbf{A}+\mathbf{B}=\mathbf{C}$ como:

$$(c_{ij}) = (a_{ij}) + (b_{ij}) \quad i \in \{1,\ldots,n\} \quad j \in \{1,\ldots,m\}$$

- **Multiplicación de Matrices.** Sea $\mathbf{A} \in \mathbb{R}^{(n\times m)}$ y $\mathbf{B} \in \mathbb{R}^{(m\times k)}$ se define la multiplicación de matrices $\mathbf{AB}=\mathbf{C} \in \mathbb{R}^{(n\times k)}$ como:

$$(c_{ij}) = \sum_{q=1}^{m} a_{iq}b_{qj} \quad i \in \{1,\ldots,n\} \quad j \in \{1,\ldots,k\}$$

  Ejercicio: (*Expresar el SCE como producto matricial, *Expresar una combinación lineal como producto matricial)

Matriz inversa:
Sea $\mathbf{A} \in \mathbb{R}^{n\times n}$ una matriz cuadrada. La inversa de $\mathbf{A}$, denotada por $\mathbf{A}^{-1}$, es otra matriz $n \times n$ tal que:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Si la inversa existe, es única.

- **Propiedad Distributiva:** Sea $\mathbf{A}$; $\mathbf{B} \in \mathbb{R}^{(n \times m)}$ y $\mathbf{C} \in \mathbb{R}^{(k \times n)}$ entonces:

$$\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$$

Por ejemplo, supongamos $\mathbf{A}$; $\mathbf{B}$; $\mathbf{C} \in \mathbb{R}^{(n \times n)}$ entonces:

$$\left(\mathbf{A}^2 + \mathbf{ABA}\right) = (\mathbf{A} + \mathbf{AB})\,\mathbf{A} = \mathbf{A}(\mathbf{I} + \mathbf{B})\,\mathbf{A}$$

- **Transpuesta de una Matriz**. Sea $\mathbf{A} \in \mathbb{R}^{(n \times m)}$. La matriz transpuesta de $\mathbf{A}$ denotada por $\mathbf{A}^T$, es una matriz en $\mathbb{R}^{(m \times n)}$ tal que sus columnas son los renglones de $\mathbf{A}$ es decr:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} \Rightarrow \mathbf{A}^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \dots & a_{nm} \end{pmatrix}$$

  Observación: Si $\mathbf{A}$ es una matriz $(n \times m)$ y $\mathbf{B}$ una matriz $(m \times k)$ entonces:

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$$

  Observación: Si $\mathbf{A}$ es una matriz $(n \times m)$ y $\mathbf{B}$ un matriz $(n \times m)$ entonces:

$$(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$$

- **Matriz Simétrica**. Sea $\mathbf{A} \in \mathbb{R}^{(n \times n)}$. Se dice que $\mathbf{A}$ se llama simétrica si $\mathbf{A} = \mathbf{A}^T$
- **Matriz Idempotente**. Sea $\mathbf{A} \in \mathbb{R}^{(n \times n)}$. Se dice que $\mathbf{A}$ es idempotente si $\mathbf{A} = \mathbf{AA} = \mathbf{A}^2$
  Observación: Si $\mathbf{A}$ es simétrica e idempotente, entonces $\mathbf{I} - \mathbf{A}$ es tambien simétrica e idempotente.

- **Matriz Ortogonal.** Sea $\mathbf{A} \in \mathbb{R}^{(n \times n)}$. Se dice que $\mathbf{A}$ es ortogonal si $\mathbf{A}^T\mathbf{A} = \mathsf{I}$, en consecuencia, si $\mathbf{A}$ es ortogonal entonces $\mathbf{A}^{-1} = \mathbf{A}^T$
- **Forma Cuadrática** Sea $\underline{y}$ un vector $(n \times 1)$ y sea $\mathbf{A} \in \mathbb{R}^{(n \times n)}$ entonces la función $f : \mathbb{R}^n \to \mathbb{R}$ con regla de correspondencia:

$$f(\underline{y}) = \underline{y}^T\mathbf{A}\underline{y} = \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}y_iy_j$$

  es llamada una forma cuadrática. $\mathbf{A}$ es llamada la matriz de la forma cuadrática. (Ejercicio: Expresar a SCE como una forma cuadrática en función de $\underline{Y}$)
- **Matriz Definida Positiva y Semidefinida Positiva** Sea $\mathbf{A} \in \mathbb{R}^{(n \times n)}$. Se dice que $\mathbf{A}$ es positiva definida si las siguientes condiciones se cumplen:
  - $\mathbf{A} = \mathbf{A}^T$ (Es simétrica)
  - $\underline{y}^T\mathbf{A}\underline{y} > 0 \quad \forall \underline{y} \in \mathbb{R}^n \quad \underline{y} \neq 0$ (Positiva Definida)
  - $\underline{y}^T\mathbf{A}\underline{y} \geq 0 \quad \forall \underline{y} \in \mathbb{R}^n \quad \underline{y} \neq 0$ (Positiva Semidefinida)

- **Traza de una Matriz**. Sea $\mathbf{A} \in \mathbb{R}^{(n \times n)}$. La traza de $\mathbf{A}$, denotada por $tr(\mathbf{A})$, es la suma de los elementos en la diagonal de $\mathbf{A}$

$$tr(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

Algunas propiedades de la traza son:

- Si $\mathbf{A} \in \mathbb{R}^{(m \times n)}$ y $\mathbf{B} \in \mathbb{R}^{(n \times m)}$ entonces:

$$tr(\mathbf{AB}) = tr(\mathbf{BA})$$

- Si $\mathbf{A} \in \mathbb{R}^{(m \times k)}$; $\mathbf{B} \in \mathbb{R}^{(k \times n)}$ y $\mathbf{C} \in \mathbb{R}^{(n \times m)}$ entonces:

$$tr(\mathbf{ABC}) = tr(\mathbf{CAB})$$

- Si $\mathbf{A} \in \mathbb{R}^{(n \times n)}$ y $\mathbf{B} \in \mathbb{R}^{(n \times n)}$ y a, b dos escalares, entonces:

$$tr(a\mathbf{A} + b\mathbf{B}) = a(tr(\mathbf{A})) + b(tr(\mathbf{B}))$$

- **Rango de Matriz** Sea $\mathbf{A} \in \mathbb{R}^{(m \times n)}$, se define el rango de $\mathbf{A}$ como el número de columnas linealmente independientes, equivalentemente es el número de renglones linealmente independientes. Ejemplo:

$$Rank\left(\mathbf{I}_{n \times n}\right) = n$$

- **Rango de Matriz Idempotente**. Sea $\mathbf{A} \in \mathbb{R}^{(n \times n)}$ una matriz idempotente, entonces:

$$Rank\left(\mathbf{A}\right) = tr\left(\mathbf{A}\right)$$

Cálculo diferencial matricial

Una función real de varias variables es aquella que asocia a un vector $\underline{x} \in \mathbb{R}^n$ a un único valor real $x \in \mathbb{R}$, es decir:

$$f : \mathbb{R}^n \to \mathbb{R} \quad f(\underline{x}) = x$$

La derivada de una función de variables variables se puede calcular derivando parcialmente respecto a cada componente formando a si un vector columna al que se le denomina gradiente de f es decir:

$$\bigtriangledown f = \frac{\partial}{\partial \underline{x}} f = \left(\frac{\partial}{\partial x_1} f(\underline{x}), \frac{\partial}{\partial x_2} f(\underline{x}), \ldots, \frac{\partial}{\partial x_n} f(\underline{x}) \right)^T$$

Supongamos ahora que tenemos $\mathbf{A} \in \mathbb{R}^{(n \times n)}$, $\underline{a}$ un vector columna, es decir $\underline{a} \in \mathbb{R}^{(n \times 1)}$ y $\underline{x}$ un vector columna de variables. Entonces:

- Si $f(\underline{x}) = \underline{a}^T \underline{x}$ o si $f(\underline{x}) = \underline{x}^T \underline{a}$ entonces:

$$\bigtriangledown f(\underline{x}) = \frac{\partial}{\partial \underline{x}} f = \underline{a}$$

Cálculo diferencial matricial

- (Regla de la suma). Supongamos que tenemos dos funciones de varias variables $f : \mathbb{R}^n \to \mathbb{R}$ y $g : \mathbb{R}^n \to \mathbb{R}$. Definamos, la función $h : \mathbb{R}^n \to \mathbb{R}$ como $h(\underline{x}) = f(\underline{x}) + g(\underline{x})$ entonces:

$$\bigtriangledown h(\underline{x}) = \bigtriangledown f(\underline{x}) + \bigtriangledown g(\underline{x})$$

- Si $f(\underline{x}) = \underline{x}^T \underline{x}$ entonces:

$$\bigtriangledown f(\underline{x}) = \frac{\partial}{\partial \underline{x}} f = 2\underline{x}$$

- Si $f(\underline{x}) = \underline{a}^T \mathbf{A} \underline{x}$ entonces:

$$\bigtriangledown f(\underline{x}) = \frac{\partial}{\partial \underline{x}} f = \mathbf{A}^T \underline{a}$$

- (Regla del producto). Supongamos que tenemos dos funciones de varias variables $f : \mathbb{R}^n \to \mathbb{R}$ y $g : \mathbb{R}^n \to \mathbb{R}$. Definamos, la función $h : \mathbb{R}^n \to \mathbb{R}$ como $h(\underline{x}) = f(\underline{x})g(\underline{x})$ entonces:

$$\bigtriangledown h(\underline{x}) = f(\underline{x}) \bigtriangledown g(\underline{x}) + g(\underline{x}) \bigtriangledown f(\underline{x})$$

Cálculo diferencial matricial

- Si $f(\underline{x}) = \underline{x}^T \mathbf{A} \underline{x}$ entonces:

$$\bigtriangledown f(\underline{x}) = \frac{\partial}{\partial \underline{x}} f = \mathbf{A}\underline{x} + \mathbf{A}^T \underline{x}$$

- Si $f(\underline{x}) = \underline{x}^T \mathbf{A} \underline{x}$ y $\mathbf{A}$ es simétrica entonces:

$$\bigtriangledown f(\underline{x}) = \frac{\partial}{\partial \underline{x}} f = 2\mathbf{A}\underline{x}$$

Cálculo diferencial matricial

Puntos Críticos de una función de varias variables:

Teorema (Condiciones necesarias de extremo relativo)

Sea $f : \mathbb{R}^n \to \mathbb{R}$ *una función real definida en conjunto un abierto de* $\mathbb{R}^n$ *y* $\underline{a} \in \mathbb{R}^n$. *Si* f *admite derivadas parciales en* $\underline{a}$ *y alcanza un máximo o mínimo relativo en dicho punto, entonces se verifica:*

$$\bigtriangledown f(\underline{a}) = \underline{0}$$

Luego entonces si queremos encontrar puntos críticos de una función debemos resolver el sistema de ecuaciones

$$\bigtriangledown f(\underline{x}) = \underline{0}$$

Vectores Aleatorios y sus Propiedades

A lo largo de este curso se ha venido trabajando con variables aleatorias reales, es decir, variables que modelan una características numérica de algún fenómeno aleatorio. Por ejemplo $X =$ Número que se obtendrá en el lanzamiento de un dado. Debe observarse entonces que X es una variable que toma valores reales a saber $\{1, 2, 3, 4, 5, 6\}$.
Suponga ahora que queremos modelar varias características numéricas a un fenómeno aleatorio, por ejemplo, se lanza un dardo en un plano cartesiano, sea $X_1 =$ La posición en el eje de las abscisas donde cae el dardo y $X_2 =$ La posición en el eje de las ordenadas donde cae el dardo. Este fenómeno aleatorio se modela entonces por dos variables simultáneamente (X_1, X_2). Cuando tenemos arreglos de variables aleatorias decimos que tenemos un **vector aleatorio.**

Definición (Vector Aleatorio)

Diremos que $\underline{X} := (X_1, X_2, \ldots, X_n)^T$ *es un vector aleatorio en* $\mathbb{R}^n$ *si cada componente de este vector* X_i *es una variable aleatoria real*

Vectores Aleatorios y sus Propiedades

Definición (Distribución de un Vector Aleatorio)

Sea $\underline{X} := (X_1, X_2, \ldots, X_n)^T$ *un vector aleatorio en* $\mathbb{R}^n$*, definimos la función de distribución de* $\underline{X}$ *como:*

$$F_{\underline{X}}(x_1, \ldots . x_n) := \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

Definición (Densidad de un Vector Aleatorio caso Discreto)

Sea $\underline{X} := (X_1, X_2, \ldots, X_n)^T$ *un vector aleatorio en* $\mathbb{R}^n$*, tal que cada* X_i *es una v.a. Discreta. Definimos la densidad de* $\underline{X}$ *como :*

$$f_{\underline{X}}(x_1, \ldots . x_n) := \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

Definición (Densidad de un Vector Aleatorio caso absolutamente continuo)

Sea $\underline{X} := (X_1, X_2, \ldots, X_n)^T$ *un vector aleatorio en* $\mathbb{R}^n$*, tal que cada* X_i *es una v.a. absolutamente continua. Definimos la densidad de* $\underline{X}$ *como aquella función* $f_{\underline{X}} : \mathbb{R}^n \to \mathbb{R}$ *tal que toda región* $D \in \mathbb{R}^n$ *se tiene que :*

$$\mathbb{P}(\underline{X} \in D) = \int_D f_{\underline{X}}(x_1, x_2, \ldots, x_n)$$

Vectores Aleatorios y sus Propiedades

Definición (Esperanza de un Vector Aleatorio)

Sea $\underline{X} := (X_1, X_2, \ldots, X_n)^T$ *un vector aleatorio en* $\mathbb{R}^n$. *Se define la esperanza del vector* $\underline{X}$ *como:*

$$\mathbb{E}(\underline{X}) := (\mathbb{E}(X_1), \mathbb{E}(X_2), \ldots . \mathbb{E}(X_n))^T$$

Definición (Varianza de un Vector Aleatorio)

Sea $\underline{X} := (X_1, X_2, \ldots, X_n)$ *un vector aleatorio en* $\mathbb{R}^n$. *Se define la varianza del vector* $\underline{X}$ *como la siguiente matriz:*

$$Var(\underline{X}) := \Sigma_{\underline{X}} := \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) & \ldots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & Cov(X_2, X_3) & \ldots & Cov(X_2, X_n) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var(X_3) & \ldots & Cov(X_3, X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & Cov(X_n, X_3) & \ldots & Var(X_n) \end{pmatrix}$$

$$Var(\underline{X}) := \mathbb{E}\Big((\underline{X} - \mathbb{E}(\underline{X}))(\underline{X} - \mathbb{E}(\underline{X}))^T\Big)$$

Obs: Se puede probar que $\Sigma_{\underline{X}}$ es una matriz simétrica y definida positiva.

Vectores Aleatorios y sus Propiedades

Propiedades importantes de los vectores aleatorios:
Sea $\mathbf{A} \in \mathbb{R}^{(n \times n)}$ un matriz cuadrada de constantes, $\underline{a}$ un vector columna de constantes $(k \times 1)$ y $\underline{X}$ un vector aleatorio en $\mathbb{R}^n$ tal que $\mathbb{E}(\underline{X}) = \underline{\mu}$ y $\text{Var}(\underline{X}) = \Sigma_{\underline{X}}$

- $\mathbb{E}(\underline{a}^T \underline{X}) = \underline{a}^T \underline{\mu}$
- $\mathbb{E}(\mathbf{A}\underline{X}) = \mathbf{A}\underline{\mu}$
- $\text{Var}(\underline{a}^T \underline{X}) = \underline{a}^T \Sigma_{\underline{X}} \underline{a}$
- $\text{Var}(\mathbf{A}\underline{X}) = \mathbf{A}\Sigma_{\underline{X}}\mathbf{A}^T$
- Si $\Sigma_{\underline{X}} = \sigma^2\mathbf{I}$ entonces $\text{Var}(\mathbf{A}\underline{X}) = \sigma^2\mathbf{A}\mathbf{A}^T$
- $\mathbb{E}(\underline{X}^T\mathbf{A}\underline{X}) = tr\left(\mathbf{A}\Sigma_{\underline{X}}\right) + \underline{\mu}^T\mathbf{A}\underline{\mu}$
- Si $\Sigma_{\underline{X}} = \sigma^2\mathbf{I}$ entonces $\mathbb{E}(\underline{X}^T\mathbf{A}\underline{X}) = \sigma^2 tr\left(\mathbf{A}\right) + \underline{\mu}^T\mathbf{A}\mu$

Normal Multivariada y sus propiedades

Normal Univariante:
Sabemos que si $X \sim N_1(\mu, \sigma^2)$ entonces $\mathbb{E}(X) = \mu$ y $\text{Var}(X) = \sigma^2 > 0$, y tiene por densidad:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad x \in \mathbb{R}$$

Normal Multivariante:
La generalización de la normal univariante al caso multivariado es la siguiente:
Decimos que el vector aleatorio $\underline{X}$ sigue una distribución Normal Multivariada de orden p o p-variada denotada por $\underline{X} \sim N_p\left(\underline{\mu}, \Sigma_{p\times p}\right)$, donde $\underline{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^T$ y Σ es una matriz simétrica definida positiva, si la función de densidad de $\underline{X}$ está dada por:

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} exp\left(-\frac{1}{2}\left(\underline{x}-\underline{\mu}\right)^T \Sigma^{-1}\left(\underline{x}-\underline{\mu}\right)\right) \quad \underline{x} \in \mathbb{R}^p$$

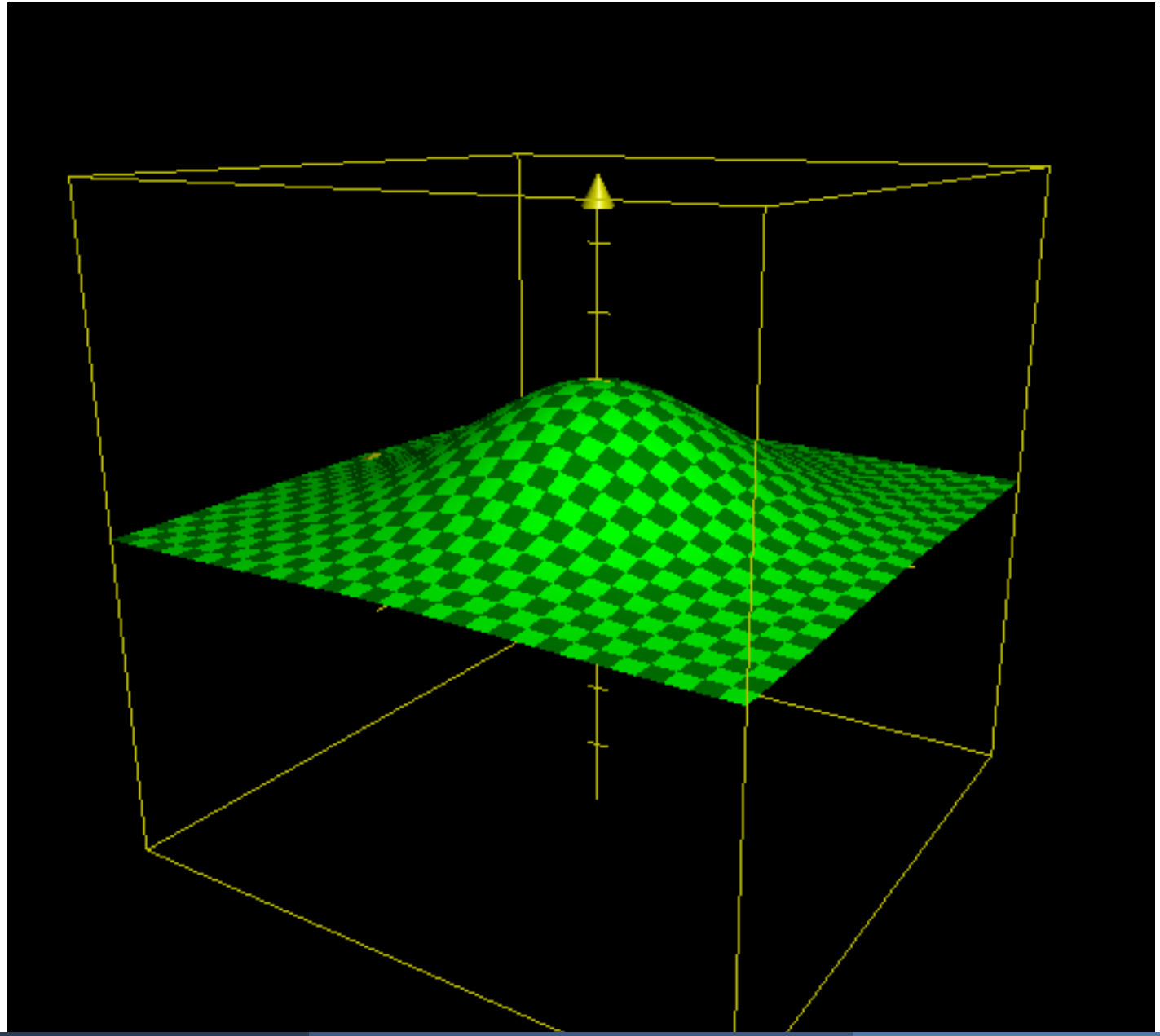
Normal Multivariada y sus propiedades

Ejemplo:
Supongamos que $p = 2$ (Normal Bi-variada) y que $\mu = (0,0)^T$, $\Sigma = \sigma^2 \mathbf{I}_{2\times 2}$ entonces:

$$\begin{aligned} f_{\underline{X}}(x_1, x_2) &= \frac{1}{(2\pi)\sigma^2} exp\left(-\frac{1}{2}(x_1, x_2)^T \left(\sigma^2 \mathbf{I}_{2\times 2}\right)^{-1} (x_1, x_2)\right); \quad (x_1, x_2) \in \mathbb{R}^2 \\ &= \frac{1}{(2\pi)\sigma^2} exp\left(-\frac{1}{2\sigma^2}\left(x_1^2 + x_2^2\right)\right); \quad (x_1, x_2) \in \mathbb{R}^2 \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{x_1^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{x_2^2}{2\sigma^2}\right); \quad (x_1, x_2) \in \mathbb{R}^2 \\ &= f_{X_1}(x_1) * f_{X_2}(x_2) \end{aligned}$$

Es decir, recuperamos la densidad conjunta como el producto de las marginales. Esta es una propiedad que ya sabiamos, pues cuando $x_1, x_2 \sim N_1(0, \sigma^2)$ independientes, entonces la conjunta se obtiene multiplicando las marginales.
Obs: En este caso se verifica el hecho de que si $Cov(x_1, x_2) = 0$ entonces x_1 es independiente de x_2

Normal Multivariada y sus propiedades



Suponga que $\underline{X} \sim N_p\left(\underline{\mu}, \Sigma\right)$ y que $\underline{a} = (a_1, a_2, \ldots, a_p)^T$. Entonces:

- La componente i-ésima del vector $\underline{X}$ sigue una distribución normal con parámetros $\left(\mu_i, \sigma_{ii}^2\right)$ es decir $X_i \sim N\left(\mu_i, \sigma_{ii}^2\right)$
- $\underline{a}^T \underline{X} \sim N_1\left(\underline{a}^T \underline{\mu}, \underline{a}^T \Sigma \underline{a}\right)$
- Si para toda $\underline{a} \in \mathbb{R}^p$ (visto como vector columna) se tiene que $\underline{a}^T \underline{Y}$ es normal entonces $\underline{Y}$ es normal p-variante
- Supongamos que participamos al vector $\underline{X}$ de la siguiente manera $\left(\underline{X_1}, \underline{X_2}\right)$ donde $\underline{X_1} \in \mathbb{R}^q$ y $\underline{X_2} \in \mathbb{R}^{p-q}$ con $1 \leq q \leq p-1$. Entonces, $\underline{X_1}$ sigue una distribución Normal q-variada y $\underline{X_2}$ sigue una distribución Normal $p-q$-variada con los siguientes parámetros.

$$\underline{X_1} \sim N_q\left(\underline{\mu_1}, \Sigma_{11}\right) \qquad \underline{X_2} \sim N_{p-q}\left(\underline{\mu_2}, \Sigma_{22}\right)$$

Donde:

$$\underline{\mu} = \begin{pmatrix} \underline{\mu_1} \\ \underline{\mu_2} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Suponga que $\underline{X} \sim N_p\left(\underline{\mu}, \Sigma\right)$ y que $\mathbf{A} \in \mathbb{R}^{m \times p}$ (De rango completo). Entonces:

$$\mathbf{A}\underline{X} \sim N_m\left(\mathbf{A}\underline{\mu}, \mathbf{A}\Sigma\mathbf{A}^T\right)$$

Cualquier transformación lineal de un vector aleatorio Normal p-variado se distribuye Normal. Ejemplo:

- Si $\mathbf{A} \in \mathbb{R}^{1 \times p}$ (un vector columna en $\mathbb{R}^p$) entonces replicamos la segunda propiedad del slide anterior pues:

$$\mathbf{A}\underline{X} \sim N_1\left(\mathbf{A}\underline{\mu}, \mathbf{A}\Sigma\mathbf{A}^T\right)$$

- Si $\mathbf{A} \in \mathbb{R}^{q \times p}$ con $1 \leq q \leq p-1$ tal que:

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_{q \times q} & 0_{q \times p-q} \end{pmatrix}$$

Entonces

$$\underline{X}_1 = \mathbf{A}\underline{X} \sim N_q\left(\underline{\mu_1}, \Sigma_{11}\right)$$

Es decir replicamos la cuarta propiedad del slide anterior.