Modelos no paramétricos y de regresión

Regresión lineal simple

Javier Santibáñez

Facultad de Ciencias, UNAM

jsantibanez@sigma.iimas.unam.mx

Semestre 2019-1

Contenido

Planteamiento

- El modelo de regresión lineal simple (RLS) relaciona una variable aleatoria continua Y con una variable no aleatoria X.
- El modelo RLS queda especificado por las siguientes ecuaciones

$$E(Y \mid X = x) = \beta_0 + \beta_1 x \qquad \text{y} \qquad V(Y \mid X = x) = \sigma^2.$$

  donde $\beta_0, \beta_1 \in \mathbb{R}$ y $\sigma \in \mathbb{R}^+$ son los parámetros del modelo.
- La primera ecuación establece que el valor esperado de Y es una función lineal de X, mientras que la segunda ecuación indica que la variación de Y alrededor de su valor es constante para X.
- El objetivo es estimar β_0, β_1 y σ^2 a partir de un conjunto de observaciones de la población $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$.

Planteamiento

- Dado el conjunto de observaciones $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ es util representar a Y_i como

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

  donde ϵ_i es un error aleatorio que representa la variación que tiene Y_i alrededor de su valor esperado.

- De la representación anterior se sigue que

$$E(\epsilon_i \,|\, x_i) = 0 \qquad \text{y} \qquad V(\epsilon_i \,|\, x_i) = \sigma^2. \qquad i = 1, \ldots, n.$$

- Dado que los parámetros del modelo son desconocidos, los errores ϵ_i son variables aleatorias no observables.

Supuestos del modelo RLS

Dado el conjunto de observaciones observaciones

$$(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n),$$

los primeros tres supuestos del modelo RLS son:

1. Linealidad

$$E(Y_i \,|\, x_i) = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n.$$

2. Homocedasticidad (varianza constante)

$$V(Y_i \,|\, x_i) = \sigma^2, \qquad i = 1, \ldots, n.$$

3. No correlación

$$Cov(Y_i, Y_j \,|\, x_i, x_j) = 0, \qquad i, j = 1, \ldots, n \text{ e } i \neq j.$$

Supuestos del modelo

Los supuestos 2 y 3 se pueden expresar en términos de los errores como sigue

2'. Homocedasticidad (varianza constante).

$$V(\epsilon_i \,|\, x_i) = \sigma^2, \qquad i = 1, \ldots, n.$$

3'. No correlación.

$$Cov(\epsilon_i, \epsilon_j \,|\, x_i, x_j) = 0, \qquad i, j = 1, \ldots, n \text{ e } i \neq j.$$

El supuesto de no correlación implica que la forma en que dos observaciones cualquiera Y_i y Y_j varían alrededor de su valor esperado, no tiene ninguna asociación lineal.

Estimación

Como primera aproximación, consideremos el caso en que Y no varia alrededor de su media, es decir, $V(\epsilon_i \mid x_i) = 0$.
En este caso, los puntos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ caen exactamente sobre la recta

$$y = \beta_0 + \beta_1 x$$

Para determinar los valores de los parámetros basta con tomar cualquier par de puntos distintos (x_i, y_i) y (x_j, y_j), $i \neq j$, para estimar y estimar

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

Sin embargo, al considerar que Y no varía, estamos omitiendo la aleatoriedad, que es nuestro principal interés.

Ejemplo

Se tiene interés en determinar empíricamente la relación entre el precio de un viaje en taxi y la distancia del recorrido, a partir de la siguiente información

$$(2.3, 17.7), \quad (1.6, 14.9), \quad (3.1, 20.9), \quad (4.5, 26.5), \quad (2.7, 19.3)$$

Si se toma el primer par de observaciones se obtiene

$$\beta_1 = \frac{14.9 - 17.7}{1.6 - 2.3} = 4, \qquad \beta_0 = 14.9 - 1.6 \times \beta_1 = 8.5.$$

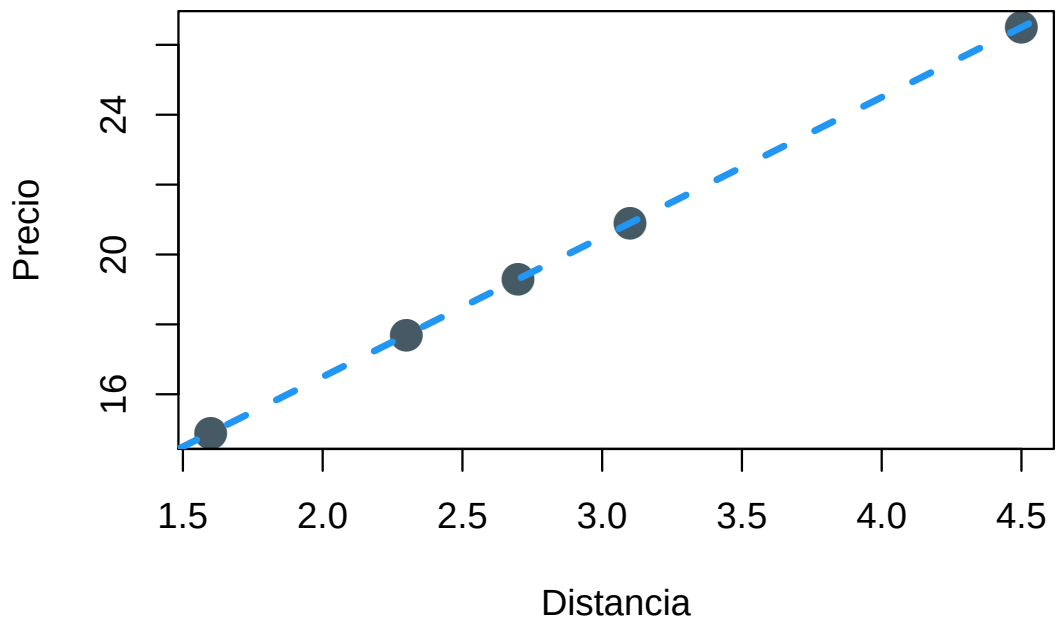
Si se toma el último par de observaciones, los resultados coinciden

$$\beta_1 = \frac{19.3 - 26.5}{2.7 - 4.5} = 4, \qquad \beta_0 = 19.3 - 2.7 \times \beta_1 = 8.5.$$

Ejemplo

En este ejemplo el precio del viaje en taxi está completamente determinado por la distancia del recorrido y la relación es la siguiente

$$\texttt{precio} = 8.5 + 4 \times \texttt{distancia}.$$



Estimación

Si Y varía alrededor de su media, es decir, $V(\epsilon_i \,|\, x_i) > 0$, entonces los puntos (x_i, y_i) no son colineales y al utilizar las expresiones anteriores se tendría una estimación distinta de β_0 y β_1 para cada par de punto.
Para cualquier estimación propuesta $\hat{\beta}_0$ y $\hat{\beta}_1$, es posible calcular la desviación de y_i con respecto a valor ajustado $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Las cantidades $e_1, e_2, \ldots, e_n$ reciben el nombre de *residuos* y están asociados a estimaciones de β_0 y β_1.
Si los residuos son *grandes*, la estimación es *mala* y si los residuos son *chicos*, la estimación es *buena*.

Minimos Cuadrados (MCO)

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

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

La suma de cuadrados de los residuos permite comparar distintas estimaciones de β_0 y β_1.

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.

Los estimadores obtenidos con este criterio son llamados *estimadores de mínimos cuadrados ordinarios.*

Solución analítica

Para minimizar Q se sigue el procedimiento *estándar* utilizando cálculo.

Es fácil verificar que

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

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

donde $\bar{x}_n = n^{-1}\sum_{i=1}^{n} x_i$ y $\bar{y}_n = n^{-1}\sum_{i=1}^{n} y_i$.

Para encontrar los puntos críticos de Q se debe resolver el sistema

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

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

Solución analítica

Se puede mostrar que la solución al sistema anterior es

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

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 que la solución sea válida se debe cumplir que $S_{xx} \neq 0$. En el caso simple, la condición $S_{xx} \neq 0$ es equivalente a que no todas las x_i sean iguales.

Para garantizar que esto se cumple, se agrega como un supuesto más al modelo RLS

Supuestos del modelo RLS

4. $S_{xx} > 0$.

Solución analítica

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

$$H(\beta_0, \beta_1) = \left[\frac{\partial^2}{\partial\beta_i \partial\beta_j} Q\right]_{i,j=0,1} = 2\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

Se debe observar que $H(\beta_0, \beta_1)$ es constante para β_0 y β_1. $H_{11} = 2n > 0$ y

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

De lo anterior se sigue que $H(\hat{\beta}_0, \hat{\beta}_1)$ es positiva definida y por lo tanto 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 la siguiente

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

Por lo tanto, los estimadores de MCO del modelo RLS son

$$\hat{\beta}_0 = \bar{Y}_n - \frac{S_{xY}}{S_{xx}}\bar{x}_n$$

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

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

Propiedades de los EMCO

Proposición

Los EMCO de β_0 y β_1 se pueden escribir como combinaciones lineales de las Y_i

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

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

Propiedades de los EMCO

Proposición

Los estimadores de MCO de β_0 y β_1 en el modelo RLS son insesgados,

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

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

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*). Lo anterior significa que

1. se pueden escribir como combinaciones lineales de las Y_i,
2. son insesgados,
3. para cualesquiera $\tilde{\beta}_0$ y $\tilde{\beta}_1$ estimadores lineales insesgados de β_0 y β_1 se cumple

$$V(\hat{\beta}_0 \,|\, \mathbf{x}) \leq V(\tilde{\beta}_0 \,|\, \mathbf{x}) \qquad \text{y} \qquad V(\hat{\beta}_1 \,|\, \mathbf{x}) \leq V(\tilde{\beta}_1 \,|\, \mathbf{x}).$$

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

Con el propósito de poder hacer estimación por intervalos y contraste de hipótesis, se agrega el siguiente supuesto al modelo RLS

Supuestos del modelo RLS

5. Normalidad: la distribución conjunta de los errores es normal.

Es fácil mostrar a partir de los supuestos anteriores que

$$\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

con lo cual

- 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.

Supuestos del modelo RLS con errores normales

1. Linealidad

$$E(Y_i \,|\, x_i) = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n.$$

2. Homocedasticidad (varianza constante)

$$V(Y_i \,|\, x_i) = \sigma^2, \qquad i = 1, \ldots, n.$$

3. No correlación

$$Cov(Y_i, Y_j \,|\, x_i, x_j) = 0, \qquad i, j = 1, \ldots, n \text{ e } i \neq j.$$

4. * Variabilidad en X

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 > 0.$$

5. Normalidad (conjunta) en los errores

$$\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Función de verosimilitud

Partimos de un conjunto de observaciones independientes

$$(x_1, Y_1), \ldots, (x_n, Y_n)$$

tales que

$$Y_i \,|\, x_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right).$$

La verosimilitud del vector de parámetros $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)^T$ está dada por

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

Ecuaciones normales

La log-verosimilitud está dada por

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

Al maximizar la log-verosimilitud usando cálculo, nos encontramos con el siguiente sistema de ecuaciones

$$\begin{aligned} n\beta_0 + n\beta_1\bar{x}_n &= n\bar{y}_n \\ n\beta_0\bar{x}_n + \beta_1\sum_{i=1}^{n}x_i^2 &= \sum_{i=1}^{n}x_i y_i \\ n\sigma^2 - \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 &= 0 \end{aligned}$$

Estimación por Máxima Verosimilitud

Las primeras dos ecuaciones son las ecuaciones normales de MCO. De esto concluimos que los EMV son iguales a los EMCO

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

De la tercera ecuación despejamos el estimador para σ^2

$$\begin{aligned}\hat{\sigma}^2_{MV} &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2\end{aligned}$$

$\hat{\sigma}^2_{MV}$ se parece a $\hat{\sigma}^2_{MCO}$, excepto que el EMV tiene como denominador n mientras que el EMCO tiene como denominador $n-2$.

Propiedades de los EMV

Como los EMV de β_0 y β_1 coinciden con los EMCO, sabemos que son los MELI y que tienen varianzas

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

Además, dado que $\hat{\beta}_0$ y $\hat{\beta}_0$ son combinaciones lineales de v.a. normales, ambos estimadores siguen una distribución normal

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

Intervalos de confianza

La funciones

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

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

Resultados auxiliares

Proposición

En el modelo de RLS con errores normales se cumple que

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

$$\hat{\beta}_0, \hat{\beta}_1 \perp \hat{\sigma}^2$$

Recordatorio

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

$$T = \frac{X}{\sqrt{Y/n}} \sim t_{(n)}$$

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

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

Ahora las funciones T_i^* solamente dependen de β_i, por lo que son cantidades pivotales que podemos utilizar para construir intervalos de confianza para β_i, $i = 0, 1$.

Nota

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

Intervalos de confianza para β_0 y β_1

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

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

entonces

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

Por lo tanto un intervalo de confianza $100(1-\alpha)\,\%$ para β_i está dado por

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

Intervalos de confianza para β_0 y β_1

La expresiones de los intervalos de confianza para β_0 y β_1 son las siguientes.

- Para β_0:

$$\hat{\beta}_0 \quad \pm \quad t_{(n-2)}^{1-\alpha/2} \hat{\sigma}_{MCO} \sqrt{\left(\frac{1}{n} + \frac{\bar{x}_n^2}{S_{xx}} \right)}$$

- Para β_1:

$$\hat{\beta}_1 \quad \pm \quad t_{(n-2)}^{1-\alpha/2} \hat{\sigma}_{MCO} / \sqrt{S_{xx}}$$

Intervalo de confianza para σ^2

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

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

o de manera equivalente

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

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

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

Sobre el IC para σ^2

- En el caso de los intervalos de confianza para β_0 y β_1, al tomar tomar los cuantiles $t^{(\alpha/2)}_{(n-2)}$ y $t^{(1-\alpha/2)}_{(n-2)}$, se garantiza que los intervalos tienen longitud mínima, debido a que la distribución $t_{(n-2)}$ es simétrica.
- Sin embargo, en el caso de la varianza σ^2, la distribución $\chi^2_{(n-2)}$ 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.

Cuantiles de la distribución χ^2

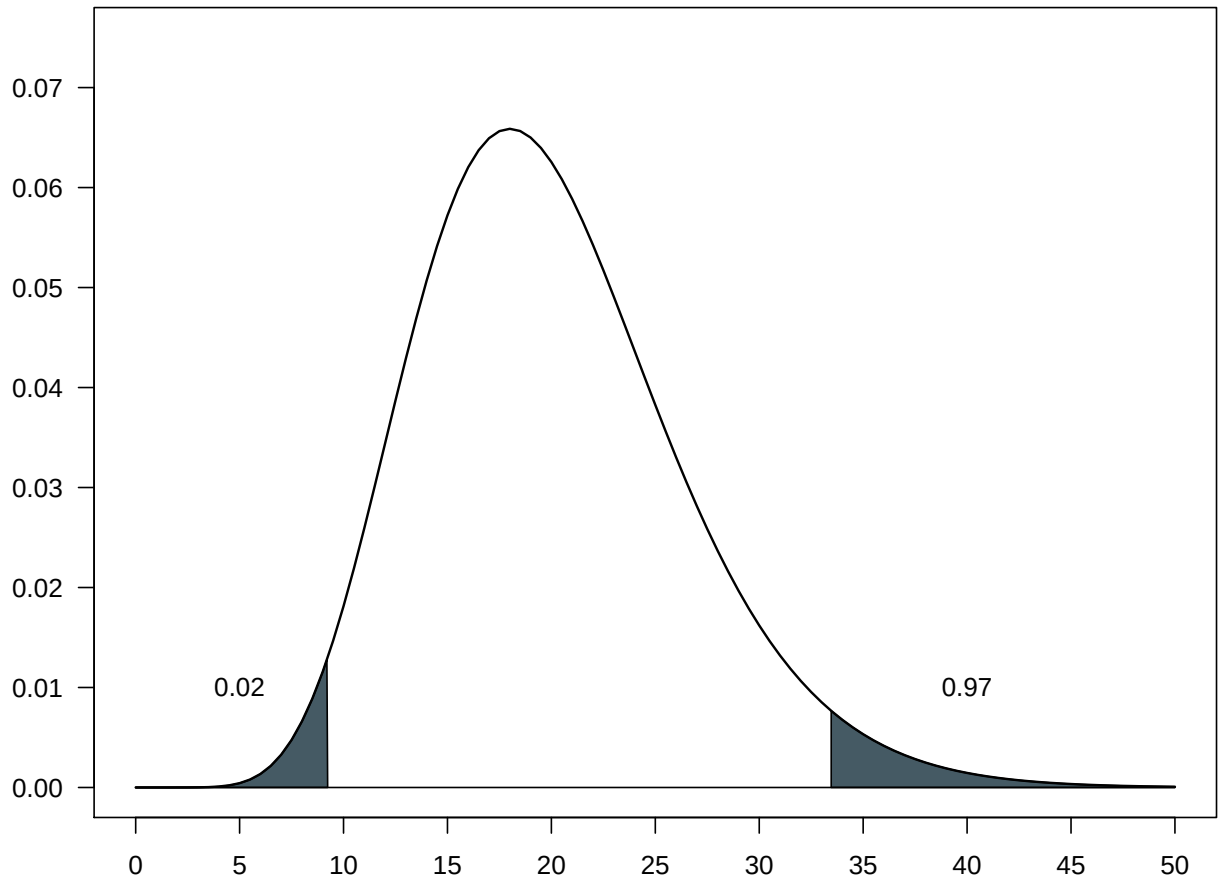


Figura: Densidad χ^2_{50}.

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

A partir de la cantidad pivotal para σ^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:

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

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

De igual manera se puede mostrar que

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

Para comparar los estimadores $\hat{\sigma}^2_{MCO}$ y $\hat{\sigma}^2_{MV}$ se debe utilizar el *Error Cuadrático Medio*, debido a que $\hat{\sigma}^2_{MV}$ es sesgado.

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, $\beta_0 + \beta_1 x$, 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

- De lo anterior se sigue que

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

- Nuevamente, si $\hat{\sigma}_x^2$ se obtiene de reemplazar $\hat{\sigma}_{MCO}^2$ por σ^2 en la expresión de σ_x^2, entonces

$$\frac{\hat{\mu}_x - \mu_x}{\hat{\sigma}_x} \sim t_{(n-2)}$$

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

$$\left(\hat{\mu}_x - t_{n-2}^{(1-\alpha/2)}\hat{\sigma}_x, \hat{\mu}_x - t_{n-2}^{(1-\alpha/2)}\hat{\sigma}_x\right)$$

De forma explícita

$$\hat{\beta}_0 + \hat{\beta}_1 x \quad \pm \quad t_{(n-2)}^{1-\alpha/2}\hat{\sigma}_{MCO}\sqrt{\frac{1}{n} + \frac{(x-\bar{x}_n)^2}{S_{xx}}}.$$

Intervalos simultáneos

- Los intervalos de confianza 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

Método de Bonferroni

- Se basa en la desigualdad de Bonferroni.
- Para construir k intervalos simultáneos propone usar $t_{(n-2)}^{(1-\alpha/2k)}$ en lugar de $t_{(n-2)}^{(1-\alpha/2)}$.
- Es recomendable para valores de k pequeños.

Método de Working-Hotelling-Scheffé

- Se basa en la distribución conjunta de $\hat{\beta}_0$ y $\hat{\beta}_1$.
- Es recomendable para construir muchos intervalos simultáneos, incluso infinitos intervalos como las bandas de confianza.
- Para cualquier número de intervalos propone utilizar $\sqrt{2F_{(2,n-2)}^{(1-\alpha)}}$ en lugar de $t_{(n-2)}^{(1-\alpha/2)}$.

Intervalos de predicción

- También es posible hacer inferencias sobre nuevas observaciones de Y. Estas nuevas observaciones son variables aleatorias. Se denorará por Y_x a una observación de Y asociada a $X = x$.
- Los supuestos del modelo RLS establecen que

$$Y_x \sim N\left(\beta_0 + \beta_1 x, \sigma^2\right)$$

- 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

$$\left(\beta_0 + \beta_1 x - z^{(1-\alpha/2)}\sigma, \beta_0 + \beta_1 x + z^{(1-\alpha/2)}\sigma\right)$$

donde $z^{(1-\alpha/2)}$ es el cuantil $1 - \alpha/2$ de la distribución $N(0, 1)$.

Intervalos de predicción

- Como los parámetros del modelo RLS son desconocidos, debemos estimarlos, esto hace que la varianza del la predicción crezca y tenga dos componentes: una debida a la estimación de los parámetros del modelo y otra debida a la variabilidad de Y_x.

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

- Si $\hat{V}(Y_x)$ se obtiene de $V(Y_x)$ al sustituir σ^2 por $\hat{\sigma}^2_{MCO}$, entonces

$$\frac{\hat{\mu}_x - Y_x}{\sqrt{\hat{V}(Y_x)}} \sim t_{(n-2)}$$

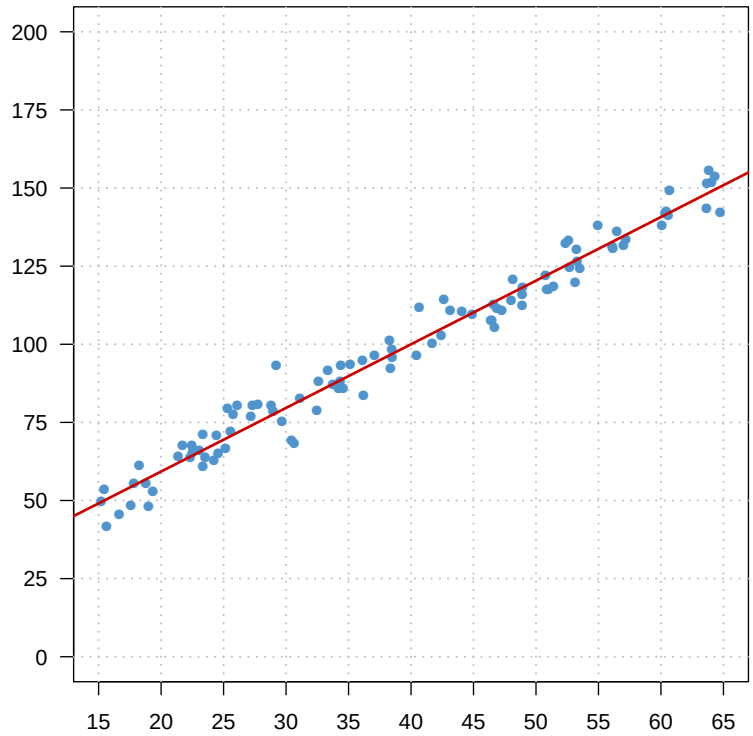
y de esta cantidad pivotal se obtiene un intervalo de predicción para Y_x es

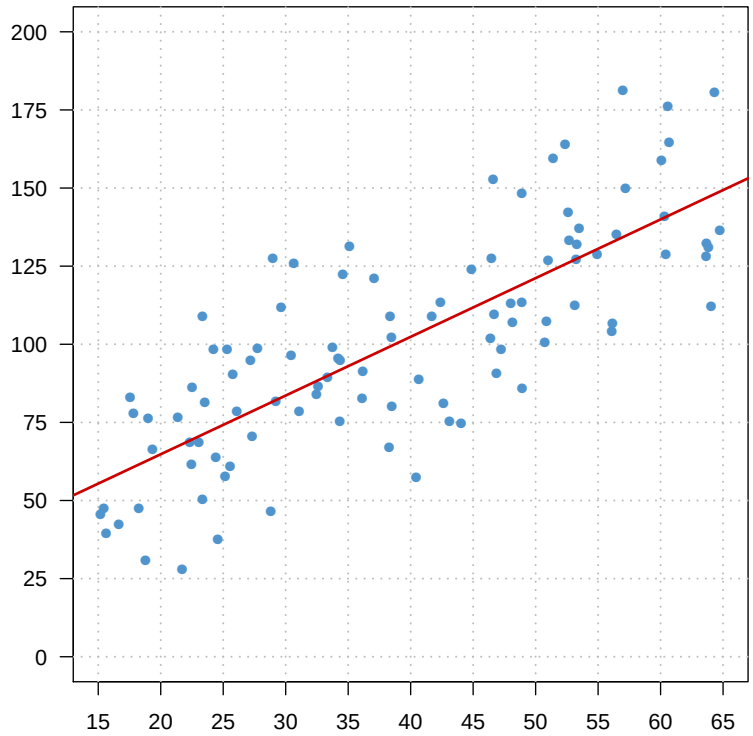
$$\left(\hat{\mu}_x - t^{(1-\alpha/2)}_{n-2} \sqrt{\hat{V}(Y_x)}, \hat{\mu}_x + t^{(1-\alpha/2)}_{n-2} \sqrt{\hat{V}(Y_x)} \right)$$

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

$$T^* = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{S_{xx}^{-1}\hat{\sigma}^2_{MCO}}} \sim t_{(n-2)}.$$

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.)

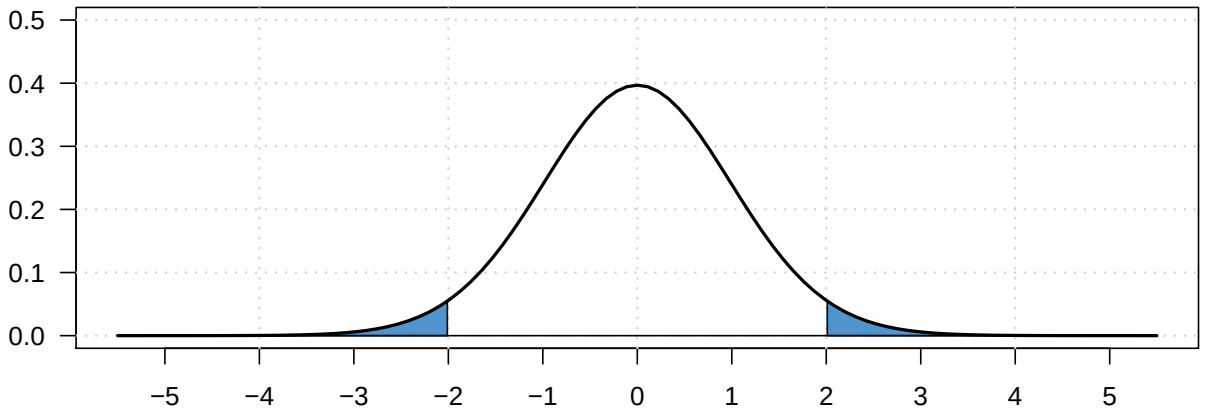
Para contrastar las hipótesis

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

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

$$| t^* | > t_{(n-2)}^{(1-\alpha/2)}$$

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



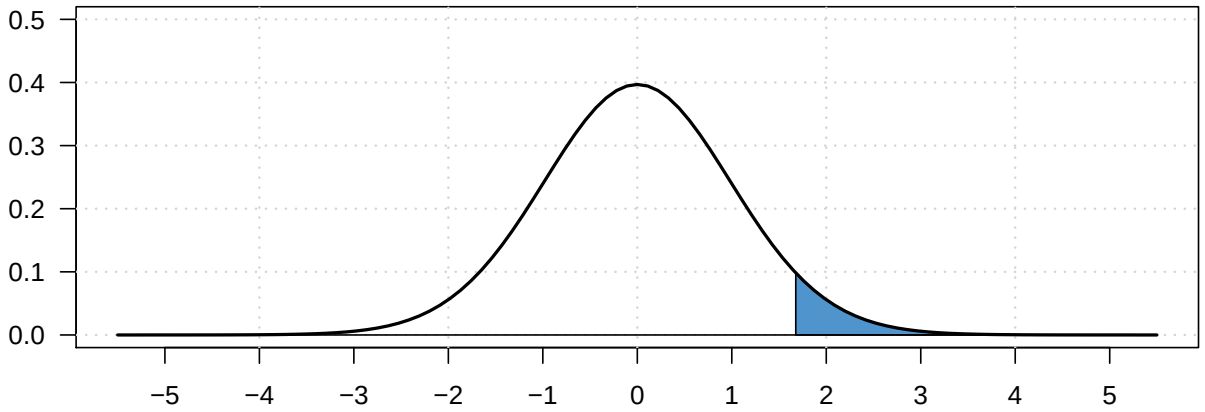
Pruebas de hipótesis para β_1 (cont.)

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

$$t > t_{(n-2)}^{(1-\alpha)}$$



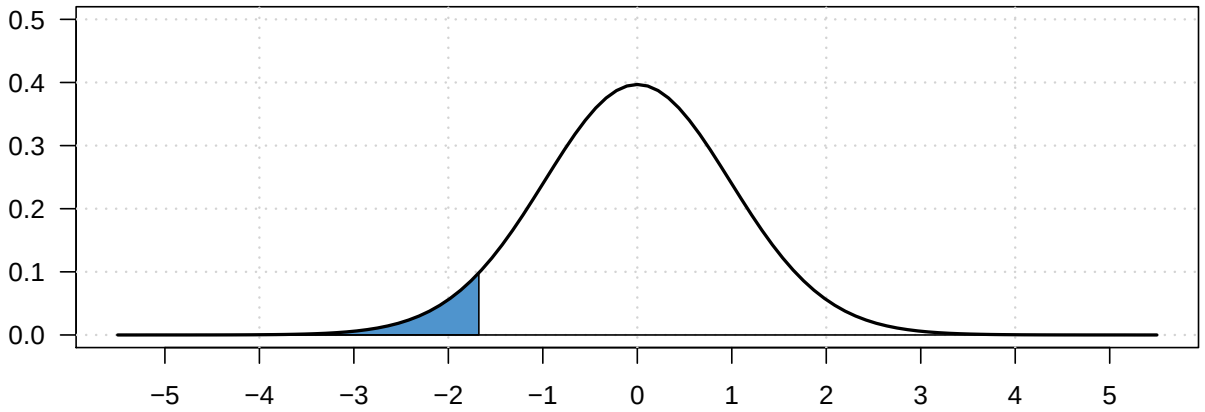
Pruebas de hipótesis para β_1 (cont.)

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

$$t < t^{(\alpha)}_{(n-2)}$$



Pruebas de hipótesis para β_0

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

donde:

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

Pruebas de hipótesis para σ^2

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

donde:

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

y $\chi^2_{(n)}(\alpha)$ denota el cuantil α de la distribución χ^2 con n grados de libertad.

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:

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

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:

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

- 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:

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

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:

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

- Al combinar los resultados anteriores obtenemos:

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

- Entonces

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

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

Recordatorio

Si $X \sim \chi^2_{(n)}$, $Y \sim \chi^2_{(m)}$ y $X \perp Y$ entonces

$$\frac{X/n_1}{Y/n_2} \sim F_{(n,m)}$$

donde $F_{(n,m)}$ denota la distribución F con n y m grados de libertad.

ANOVA (cont.)

- De lo anterior se sigue que, bajo H_0,

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

- A partir de este resultado es posible construir una prueba de hipótesis para contrastar $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. La regla de desición es, rechazar H_0 con un tamaño de prueba, o significancia, α si $F > F^{(1-\alpha)}_{(1,n-2}$. Esta es la prueba F que se realiza en el análisis de variaza.
- Lo usual es presentar los resultados del ANOVA en un arreglo que recibe el nombre de tabla ANOVA, como la que se muestra en la siguiente lámina.

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: estadístico F.

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

$$\frac{\left|\hat{\beta}_1\right|}{\sqrt{\hat{V}\left(\hat{\beta}_1\right)}} > t_{(n-2)}^{(1-\alpha/2)} \quad \Leftrightarrow \quad \frac{S_{xx}\hat{\beta}_1^2}{SC_{error}/(n-2)} > \left(t_{(n-2)}^{(1-\alpha/2)}\right)^2$$

- 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

$$\frac{SC_{reg}}{SC_{error}/(n-2)} > \left(t_{(n-2)}^{(1-\alpha/2)}\right)^2$$

- 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 $T \sim t_{(n-2)}$, 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 $x = a + by$, con $b > 0$. $r = -1$ indica una relación lineal inversa $x = a + by$, con $b < 0$. $r = 0$ 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