Regresión Lineal Simple - Semana 02

Johnatan Cardona Jiménez
jcardonj@unal.edu.co
Profesor Asistente - Departamento de Estadística
Universidad Nacional de Colombia, Sede Medellín

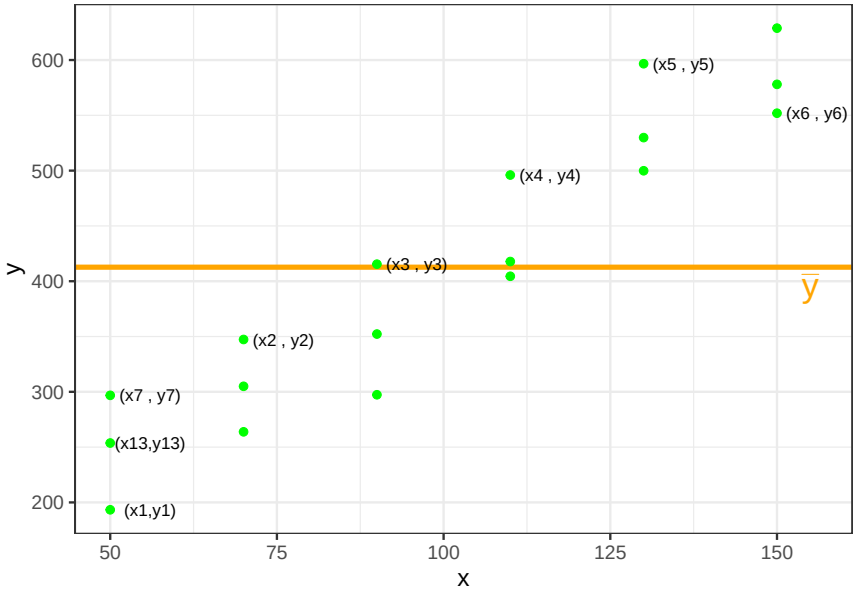
Semestre 02-2024

Análisis de varianza para probar la significancia de la regresión

Supongamos que sobre el siguiente conjunto de datos deseamos ajustar un modelo de regresión lineal:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y	x		y	x
193.3318	50	10	417.8053	110
347.3251	70	11	529.8588	130
415.4232	90	12	628.8226	150
496.0158	110	13	253.7020	50
596.6579	130	14	304.9893	70
551.9266	150	15	297.3414	90
296.8287	50	16	404.4873	110
263.8259	70	17	499.8732	130
352.2322	90	18	578.0051	150


600
500
400
300
200
y
50
75
100
125
150
x
(x5 , y5)
(x6 , y6)
(x4 , y4)
(x3 , y3)
ȳ
(x2 , y2)
(x7 , y7)
(x13,y13)
(x1,y1)

La significancia de la regresión se puede probar realizando el siguiente procedimiento de prueba de hipótesis sobre β_1:

$$H_0 : \ \beta_1 = 0$$
$$H_1 : \ \beta_1 \neq 0$$

En las clase anterior se definió un procedimiento para realizar esta prueba basado en un estadístico de prueba $T = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\widehat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$. Ahora se desarrollará una prueba alternativa, la cual arroja un resultado equivalente al obtenido con la estadística T. Esta prueba se basa en una descomposición de la variabilidad total observada en la variable respuesta:

$$\text{SST} = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

El objetivo es dividir esta suma en dos componentes: una asociada al modelo propuesto (SSR) y otra debida al error aleatorio (SSE).

$$\text{SST} = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2,$$

A esta última ecuación se le conoce como Identidad de Suma de Cuadrados, donde

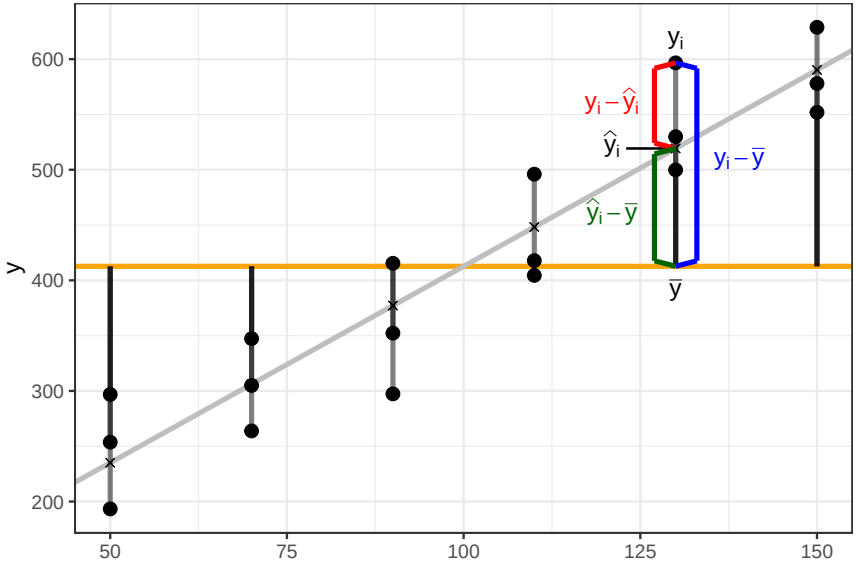
$$\text{SSR} = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \quad \text{y} \quad \text{SEE} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

Cuando se tiene un buen ajuste, se espera que SSR sea mucho mayor al SSE.

$$\underbrace{y_i}_{SST} = \underbrace{\beta_0 + \beta_1 x_i}_{SSR} + \underbrace{\epsilon_i}_{SSE} \tag{1}$$

- Veamos lo que representa gráficamente cada una de las sumas que componen a la SST:

Ilustración del enfoque de varianza



Interpretación de las sumas de cuadrados:

- $\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$, que está relacionada con las diferencias entre los valores ajustados por el modelo de regresión y el promedio de las observaciones de la respuesta, al cual se le conoce como Suma de Cuadrados de la Regresión, abreviado SSR.
  Se puede demostrar que:

$$\text{SSR} = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 = \widehat{\beta}_1 S_{xy} = \widehat{\beta}_1^2 S_{xx}$$

- $\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$, que está relacionada con las diferencias entre las observaciones de la respuesta y los valores ajustados por el modelo de regresión, esto es, los residuales del modelo (que son estimaciones de los errores del modelo), por lo que a esta componente se le conoce como Suma de Cuadrados del Error, abreviado SSE.
  Se puede demostrar que:

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = S_{yy} - \widehat{\beta}_1 S_{xy}$$

Nuevamente, **la identidad de suma de cuadrados** está dada por

$$\text{SST} = \text{SSR} + \text{SSE},$$

Cada una de estas sumas de cuadrados tiene asociados unos grados de libertad (g.l), que representan la cantidad de información libre en la suma de cuadrados.

Una forma de calcular los g.l es la diferencia entre el número de observaciones y el número de parámetros estimados en la suma de cuadrados.

- Se sabe que SST se construye con n observaciones y se estima la media de la respuesta con el promedio, de manera que SST tiene $n-1$ g.l.
- Analizando la expresión para SSE, se tienen las mismas n observaciones y se deben estimar β_0 y β_1, y así SSE tiene $n-2$ g.l.
- Finalmente, los grados de libertad de la SSR se calculan como la diferencia entre los grados del libertad de la SST y la SSE, así la SSR tiene sólo 1 g.l.

En virtud de lo anterior, los grados de libertad (g.l) de las sumas de cuadrados también forman una identidad, así:

$$\begin{array}{ccccc} \text{g.l(SST)} & = & \text{g.l(SSR)} & + & \text{g.l(SSE)} \\ (n-1) & = & (1) & + & (n-2) \end{array}$$

A continuación, se definen **los cuadrados medios** como la razón entre las sumas de cuadrados y sus respectivos grados de libertad. Esto es,

- $\text{MSR} = \text{SSR}/\text{g.l(SSR)} = \text{SSR}/1 = \text{SSR}$.
- $\text{MSE} = \text{SSE}/\text{g.l(SSE)} = \text{SSE}/(n-2)$.

Con el fin de establecer inferencias basadas en el enfoque del análisis de varianza se requiere conocer el valor esperado de cada una de los cuadrados medios, es decir, lo que se estima con cada suma de cuadrados.

Se puede demostrar que:

- $E\,[\text{MSE}] = \sigma^2$.
- $E\,[\text{MSR}] = \sigma^2 + \beta_1^2 S_{xx}$.

El primer resultado se conocía de la estimación insesgada de σ^2 obtenida en la clase anterior.

Bajo H_0 (esto es, si $\beta_1 = 0$) y asumiendo que todas las observaciones Y_i provienen de la misma distribución normal con media $\mu = \beta_0$ y varianza σ^2, se puede demostrar lo siguiente:

- SSR/σ^2 **se distribuye como una variable aleatoria Chi-cuadrado con 1 grado de libertad**.
- SSE/σ^2 **se distribuye como una variable aleatoria Chi-cuadrado con** $n-2$ **grados de libertad**.
- Los términos SSR/σ^2 y SSE/σ^2 **son estimaciones independientes de** σ^2.

De lo anterior, **se construye el siguiente estadístico:**

$$F_0 = \frac{(\text{SSR}/\sigma^2)/1}{(\text{SSE}/\sigma^2)/(n-2)} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{\text{MSR}}{\text{MSE}} \sim F_{1,n-2}.$$

Bajo la hipótesis nula $H_0 : \beta_1 = 0$, el estadístico F_0 sigue una distribución F con un grado de libertad en el numerador y $n-2$ grados de libertad en el denominador:

En el caso de la regresión lineal simple, la prueba sobre la significancia de la regresión (es decir, si la pendiente de la recta es significativamente diferente de cero) puede realizarse mediante el análisis de varianza usando un valor crítico $F_{\alpha;1,n-2}$ de la distribución F.

Esto es, si $F_0 > F_{\alpha;1,n-2}$ **a un nivel de significancia α, entonces se rechaza la hipótesis nula de que la variabilidad en la variable respuesta es debida sólo al error aleatorio** (en favor de la hipótesis de que la regresión en X es significativa).

Tabla de Análisis de Varianza para el modelo de RLS (ANOVA)

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Media	F Calculado
Regresión o Modelo	SSR	1	$MSR = \frac{SSR}{1}$	$F_0 = \frac{MSR}{MSE}$
Error o Residual	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

También se puede evaluar el valor p de la prueba que es igual a $P(F_{1,n-2} > F_0)$ y determinar si es menor que el nivel de significancia α, lo cual llevaría a rechazar la hipótesis nula: **"el modelo de regresión de Y en función X no es significativo para explicar la variabilidad observada en Y"**.

La conclusión obtenida por el análisis de varianza debe ser la misma que la obtenida cuando se prueba la significancia de la pendiente de la recta de regresión.

R^2 de una regresión: Coeficiente de determinación muestral

Es una medida del ajuste del modelo que provee un indicador de que tan bien la predictora X predice a la respuesta Y. Se calcula como:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

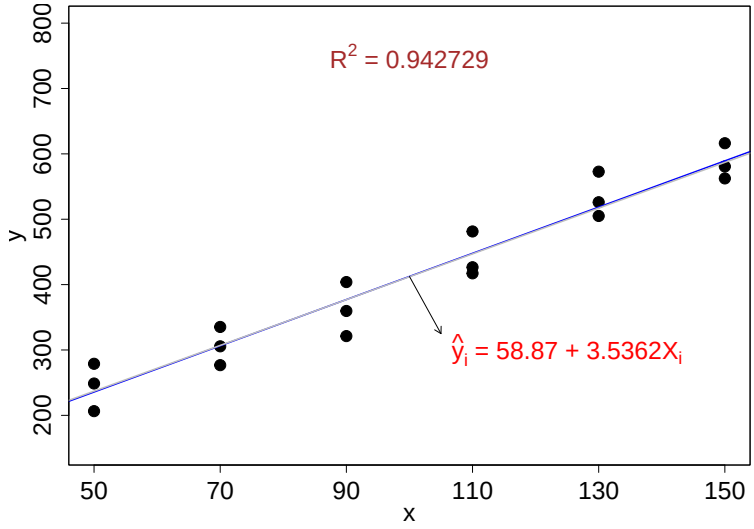
El R^2 se puede interpretar como la proporción de la variabilidad total observada en la variable respuesta que es explicada por la relación lineal con la variable predictora considerada.

Interpretaciones erróneas de R^2.

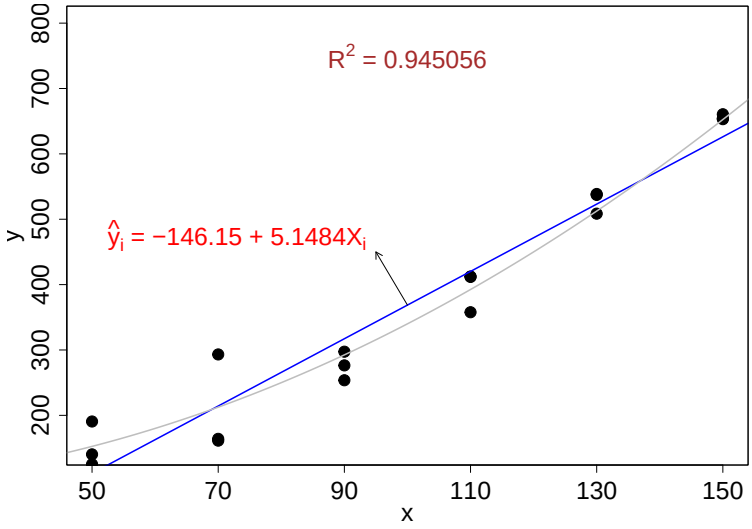
- Un R^2 alto indica que el modelo puede hacer predicciones útiles.
- Un R^2 alto indica que la recta de regresión tiene buen ajuste.
- Un R^2 cercano a cero indica que X y Y no están relacionados.

Las dos primeras indican que aunque un R^2 cercano a 1 indica una mayor asociación lineal, no necesariamente garantiza que los supuestos básicos del modelo lineal se estén cumpliendo y menos que el modelo lineal no pueda presentar falta de ajuste.

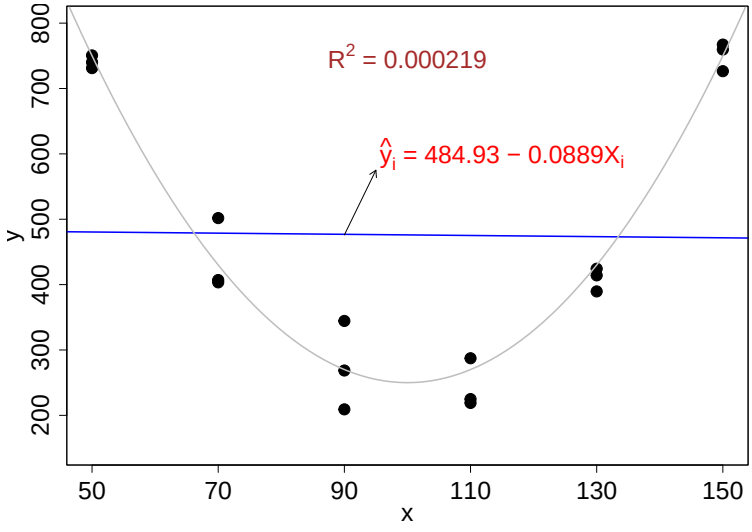
Ejemplo 1: Valor alto de R^2 - Modelo Lineal.



Ejemplo 2: Valor alto de R^2 - Modelo No Lineal



Ejemplo 3: Valor de R^2 Cercano a Cero - Modelo No Lineal



Inferencias con respecto a la Respuesta Media $E[Y|x_0]$ y Valores Futuros $Y|x_0 = y_0$

Interesa realizar inferencias sobre la respuesta, para un valor apropiado $X = x_0$, así:

- Estimación puntual y por intervalo de la respuesta media $E[Y|x_0]$.
- Predicción de valores futuros $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0 = E[Y|x_0] + \varepsilon_0$.

En ambos casos el único medio para producir tales inferencias es la ecuación de regresión ajustada.

Recuerde que la ecuación de regresión ajustada, en un valor dado $X = x_0$, es:

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

Note que $\widehat{Y}_0$ también es una combinación lineal de las variables aleatorias $Y_1, \ldots, Y_n$. en efecto,

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0 = \left(\sum_{i=1}^{n} m_i Y_i \right) + \left(\sum_{i=1}^{n} c_i Y_i \right) x_0 = \sum_{i=1}^{n} (m_i + x_0 \, c_i) Y_i,$$

con las constantes $m_i = \frac{1}{n} - \bar{x} \, c_i$ y $c_i = \frac{x_i - \bar{x}}{S_{xx}}$ como fueron especificadas previamente. Por lo tanto, bajo los supuestos del modelo $\widehat{Y}_0$ es una variable aleatoria normal.

con esperanza:

$$E\left[\widehat{Y}_0\right] = E\left[\widehat{\beta}_0 + \widehat{\beta}_1 x_0\right] = E\left[\widehat{\beta}_0\right] + E\left[\widehat{\beta}_1\right] x_0 = \beta_0 + \beta_1 x_0 = E[Y|x_0]$$

y varianza:

$$\begin{aligned} V\left[\widehat{Y}_0\right] &= V\left[\sum_{i=1}^{n}(m_i + x_0\, c_i)\, Y_i\right] = \sum_{i=1}^{n}(m_i + x_0\, c_i)^2\ V(Y_i) \\ &= \sigma^2 \sum_{i=1}^{n}\left[\left(\frac{1}{n} - \bar{x}c_i\right) + x_0 c_i\right]^2 = \sigma^2 \sum_{i=1}^{n}\left[\frac{1}{n} + (x_0 - \bar{x})c_i\right]^2 \\ V\left[\widehat{Y}_0\right] &= \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \end{aligned}$$

En resumen,

$$\widehat{Y}_0 \sim N\left(E[Y|x_0]\ ,\ \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

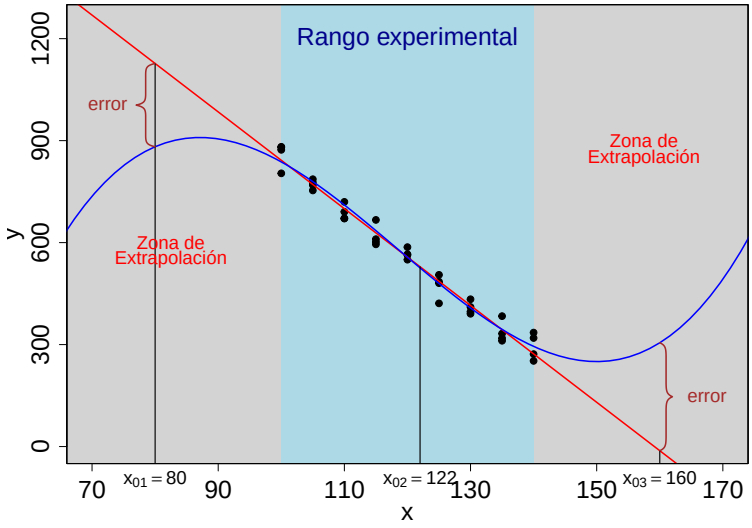
Esto es, $\widehat{Y}_0$ es un estimador insesgado de la respuesta media.

Note que, $\widehat{Y}_0$ también es un estimador para un valor futuro y_0, pero en este caso es un estimador sesgado. De ahí que la cantidad $y_0 - \widehat{Y}_0$ represente al error de predicción.

Tanto las estimaciones de valores de la respuesta media como las predicciones de valores futuro deben cumplir una condición sobre el valor fijo $X = x_0$ para que tal estimación/predicción sea válida.

- Sólo se podrán hacer inferencias sobre la respuesta cuando $X = x_0 \in [X_{\text{min}}, X_{\text{max}}]$, donde X_{min} y X_{max} son los valores mínimo y máximo de la variable predictora, que fueron fijados en la muestra. A esta intervalo también se le llama región de diseño o región de observación.
- Cumplir con lo anterior indica que x_0 es un punto de interpolación.
- Esto evita que x_0 sea un punto de extrapolación, esto es, un punto por fuera del rango experimental donde el modelo fue ajustado y que no garantiza que el modelo se mantenga.

Ilustración de puntos de interpolación y extrapolación



Intervalo de confianza para la Respuesta Media $\mu_{Y|x_0} = E[Y|x_0]$

Se puede demostrar que bajo los supuestos del modelo:

$$T = \frac{\widehat{y}_0 - E[Y|x_0]}{\sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2}$$

Por tanto un intervalo de confianza del $(1-\alpha)\%$ para $\mu_{Y|x_0}$ es:

$$\widehat{y}_0 \pm t_{\alpha/2,n-2} \times \sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

con $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ y $t_{\alpha/2,n-2}$ es el percentil $1-\alpha/2$ de la distribución t-Student con $n-2$ grados de libertad.

Intervalo de predicción para una observación futura de la respuesta Y_0

Dicho intervalo estima los posibles valores para un valor particular de la variable respuesta (no para su media) en un valor dado, $X = x_0$. Asumimos que este valor particular es un valor futuro de la variable aleatoria Y y por tanto, no fue utilizado en la regresión.

Si Y_0 es un valor futuro y $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ es su estimador, entonces estas dos variables aleatorias son estadísticamente independientes, dado que Y_0 no es utilizado para hallar a $\widehat{\beta}_0$ y $\widehat{\beta}_1$.

Por tanto, el estadístico:

$$T = \frac{\widehat{y}_0 - Y_0}{\sqrt{\widehat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2}$$

De ahí que, un intervalo de predicción del $(1-\alpha)\%$ para Y_0 está dado por:

$$\widehat{y}_0 \pm t_{\alpha/2,n-2} \times \sqrt{\widehat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

Donde, $t_{\alpha/2,n-2}$ es el percentil $1-\alpha/2$ de la distribución t-Student con $n-2$ grados de libertad.

Pruebas de hipótesis para la Respuesta Media

Para la respuesta media se pueden probar hipótesis a partir de la construcción y el análisis de los intervalos de confianza definidos anteriormente. Esto es, para probar a un nivel de significancia α, el siguiente juego de hipótesis:

$$H_0: \ E[Y|x_0] = c_0$$
$$H_1: \ E[Y|x_0] \neq c_0$$

Donde $c_0 \in \mathbb{R}$. **Criterio de decisión:** se calcula un intervalo de confianza del $(1-\alpha)100\%$ para $E[Y|x_0]$ y si el valor c_0 está incluido en el intervalo, entonces no se rechaza H_0, o si el valor c_0 no está incluido en el intervalo, entonces se rechaza H_0.