Bioestadística con R

Roberto Bustillos, MVZ, M.Sc - Basado en un material de C. Armero

Universidad Central del Ecuador

Facultad de Medicina Veterinaria y Zootecnia

Maestría en Epidemiología y Salud Pública Veterinaria, 2016-2018

Contenido

Contraste de hipótesis para la pendiente del modelo.

- Contrastes de hipótesis:

$$
\begin{array}{lll}
H_0 : \beta_1 = 0, & H_1 : \beta_1 \neq 0 & \text{contraste de dos colas} \\
 & H_1 : \beta_1 > 0 & \text{contraste de una cola} \\
 & H_1 : \beta_1 < 0 & \text{contraste de una cola}
\end{array}
$$

- P-valores:

Contraste de hipótesis		P-valor
$H_0 : \beta_1 = 0,$	$H_1 : \beta_1 \neq 0$	$2\,P(t(n-2) \geq \mid t_{\hat{\beta}_1} \mid)$
	$H_1 : \beta_1 > 0$	$P(t(n-2) > t_{\hat{\beta}_1})$
	$H_1 : \beta_1 < 0$	$P(t(n-2) < t_{\hat{\beta}_1})$

- Si α es el nivel de significatividad en cualquiera de los contrastes planteados:

$$
\begin{array}{l}
\text{P-valor} \geq \alpha \rightsquigarrow \text{No rechazar } H_0 \\
\text{P-valor} < \alpha \rightsquigarrow \text{Rechazar } H_0
\end{array}
$$

- Importante: La hipótesis nula establece que la variable respuesta no depende linealmente de la predictora.

Descomposición de la suma de cuadrados I.

- Vamos a fijarnos en la siguiente expresión:

$$\underbrace{Y_i - \bar{Y}}_{(1)} = \underbrace{(Y_i - \hat{Y}_i)}_{(2)} + \underbrace{(\hat{Y}_i - \bar{Y})}_{(3)}, \quad i = 1, \ldots, n$$

(1) Desviación de Y_i con respecto a su media muestral $\bar{Y}$
(2) Desviación de Y_i con respecto a su valor ajustado $\hat{Y}_i$
(3) Desviación de $\hat{Y}_i$ con respecto a su media muestral $\bar{Y}$

- Además:

$$\underbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^{2} (\hat{Y}_i - \bar{Y})^2}_{SSR}$$

  - SST, Suma de cuadrados total; es una medida de la variabilidad de los datos de Y con respecto a su media muestral.
  - SSE, Suma de cuadrados residual: es una medida de la variabilidad de los datos de Y con respecto a los valores ajustados.
  - SSR, Suma de cuadrados explicada por el modelo; es una medida de la variabilidad de los valores ajustados $\hat{Y}_i$ con respecto a su media muestral.

Descomposición de la suma de cuadrados II.

- Recordamos que:

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{SST}=\underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}_{SSE}+\underbrace{\sum_{i=1}^{2}(\hat{Y}_i-\bar{Y})^2}_{SSR}$$

- Cada una de estas sumas de cuadrados tiene asociado un número (grados de libertad).

$$\underbrace{SST}_{n-1}=\underbrace{SSE}_{n-2}+\underbrace{SSR}_{1}$$

- Si $Y_i=\hat{Y}_i$, los residuos serán todos cero y, por lo tanto, su suma de cuadrados también, SSE=0. Esta es una situación ideal en la que todos los valores de Y estarían sobre la recta de regresión y $SST=SSR$.
- Si $\hat{Y}_i=\bar{Y}$, el modelo ajustado no explica nada de la variabilidad de las Y con respecto a su media, con lo que $SST=SSE$. Esta es la peor situación, el modelo de regresión no nos sirve porque la recta de regresión ajustada tendría pendiente cero e interceptación $\bar{y}$.

Tabla de ANOVA I.

Tabla ANOVA:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	Cociente F	P-valor
Regresión	SSR	1	MSR	MSR/MSE	$P(\mathrm{F}(1, n-2) > F)$
Error	SSE	$n-2$	MSE		
Total	SST	$n-1$			

Tabla de ANOVA II.

- Tabla ANOVA:

Fuente de variación	Suma de cuadrado	grados de libertad	Cuadrado medio	Cociente F	P-valor
Regresión	8896.334	1	8896.334	56.941	0.000
Error	1093.666	7	156.238		
Total	9990.000	8			

- El P-valor de la tabla ANOVA para el contraste de hipótesis $H_0 : \beta_1 = 0$, vs. $H_1 : \beta_1 \neq 0$ es 0.000 por lo que considerando α=0.05 concluiríamos rechazando H_0 y considerando, por tanto, el modelo de regresión como significativo.

Introducción.

- Hemos estudiado el modelo de regresión lineal simple sin mencionar en ningún momento el grado de asociación lineal entre las variables X e Y.
- Introduciremos dos medidas que se utilizan frecuentemente para describir el grado de asociación entre dos variables, el coeficiente de determinación y el coeficiente de correlación.

Coeficiente de determinación I.

- Recordamos que en el modelo de regresión lineal simple

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

- $STT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ es una medida de la variabilidad de los datos de Y cuando no se considera ninguna variable predictora X que pueda disminuir su incertidumbre.
- $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ es una medida de la variabilidad de Y explicada por el modelo (por X) y, por lo tanto, puede considerarse como una medida natural del efecto de X en la reducción de su variabilidad.
- $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ es una medida de la variabilidad de Y que no explica el modelo (X).

Coeficiente de determinación II.

- Coeficiente de determinación: Es la proporción de variabilidad de Y explicada por el modelo.

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

- El símbolo habitual para representar el coeficiente de determinación es R^2.

Coeficiente de determinación III.

Propiedades del coeficiente de determinación:

- $0 \leq R^2 \leq 1$
- R^2 es adimensional
- Un valor grande de R^2 indica que la variable predictora, X, explica mucha de la variabilidad de Y.
- Cuando $Y_i = \hat{Y}_i$ tendremos que $SSE = 0$ y, por lo tanto, $R^2 = 1$.
- Cuando la recta de regresión ajustada es paralela al eje de abcisas ($\hat{\beta}_1 = 0$) e $Y_i = \hat{Y}_i$ se cumplira que $R^2 = 0$.

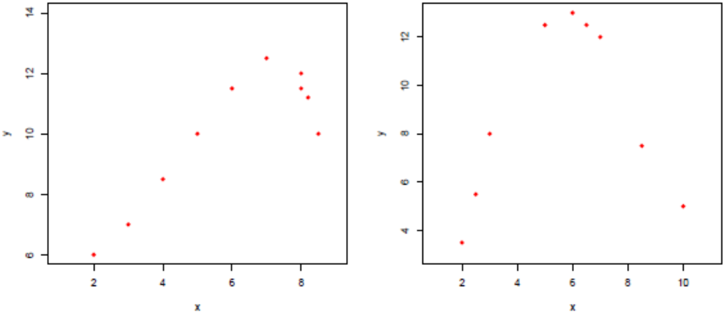
Coeficiente de determinación IV.

Limitaciones del coeficiente de determinación:

- Un coeficiente de determinación alto no significa, necesariamente, que el modelo tenga una capacidad predictiva alta. R^2 mide la reducción de SST explicada por el modelo de regresión, pero no proporciona información ni sobre la precisión de las estimaciones ni de las predicciones.
- No siempre un coeficiente de determinación alto indica que la recta de regresión ajustada es un buen modelo para los datos considerados.
- No siempre un coeficiente de determinación cercano a cero significa que no haya relación entre X e Y. R^2 sólo mide el grado de asociación lineal entre dos variables e ignora cualquier otro tipo de relación que no sea lienal.

Coeficiente de determinación V.

- $R^2 = 0.741$ y $R^2 = 0.086$



Coeficiente de correlación I.

- El coeficiente de correlación para una muestra de datos emparejados de la distribución (X,Y) es una medida del grado de asociación entre ambas variables.

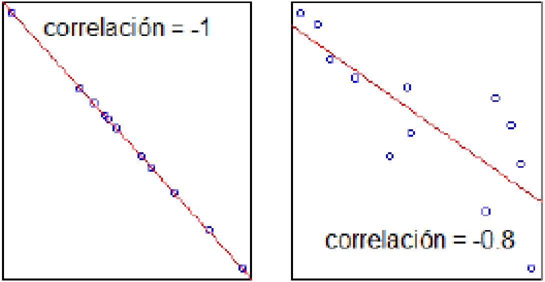
$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{s_{XY}}{s_X \, s_Y}$$

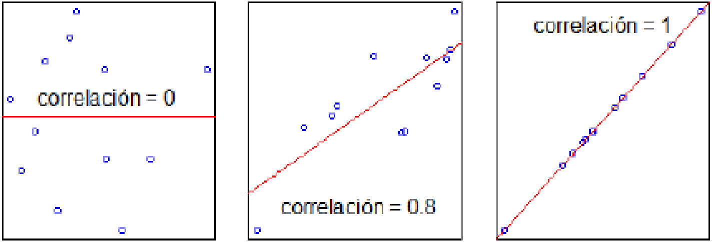
- El símbolo que se utiliza para representar la correlación es r.
- El valor absoluto del coeficiente de correlación muestral es la raíz cuadrada positiva del coeficiente de determinación: $|r| = +\sqrt{R^2}$

Coeficiente de correlación II.

Propiedades del coeficiente de correlación:

- $-1 \leq r \leq 1$
- Es adimensional.
- $r = 1$ si y sólo si todas las observaciones están sobre la recta de regresión ajustada con pendiente positiva.
- $r = -1$ si y sólo si todas las observaciones están sobre la recta de regresión ajustada con pendiente negativa.
- $r = 0$ cuando no existe relación lineal entre las observaciones de X e Y.


correlación = -1
correlación = -0.8


correlación = 0
correlación = 0.8
correlación = 1

Ejemplos en R.

Examples

```
# Tabla ANOVA
anova(model)
# Correlación
cor(x,y)
cor.test(x, y)
```