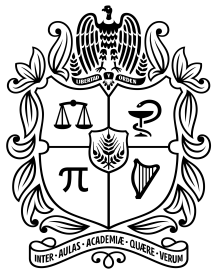
Estadística II - 3006915
Regresión Lineal Simple

Mateo Ochoa Medina

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín

Periodo académico 2023-2S



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Contenido

1. Análisis de varianza para probar la significancia de la regresión
2. Coeficiente de determinación R^2
3. Referencias

Contenido

1. Análisis de varianza para probar la significancia de la regresión
2. Coeficiente de determinación R^2
3. Referencias

Análisis de varianza para probar la significancia de la regresión

El análisis de varianza o ANOVA, consiste en la descomposición de la variabilidad total observada en la variable respuesta (SST o S_{yy}) en la suma de componentes o fuentes de variabilidad, de acuerdo al modelo propuesto (para el caso recuérdese que es el modelo de regresión lineal simple). Se espera que la recta ajustada explique en forma significativa la variabilidad observada en Y. Bajo los supuestos del modelo, la variabilidad total muestral de la respuesta satisface la siguiente descomposición:

$$\underbrace{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}_{\text{Variabilidad total}}=\underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}_{\text{Variabilidad explicada}}+\underbrace{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}_{\text{Variabilidad no explicada}} \tag{1}$$

La ecuación (1) se conoce como *identidad de suma de cuadrados*.

Consideraciones de la identidad de suma de cuadrados

Notas sobre $\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2+\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$:

- Se acostumbra llamar a

$$SSR=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2=\hat{\beta}_1^2S_{xx}=\hat{\beta}_1S_{xy} \tag{2}$$

  la *suma de cuadrados de regresión*.

- De esta manera, la identidad se puede reescribir como

$$SST=SSR+SSE=\hat{\beta}_1S_{xy}+SSE=\hat{\beta}_1^2S_{xx}. \tag{3}$$

- SSR tiene que ver con la medición de la cantidad de variabilidad en las observaciones y_i explicada por la recta de regresión ajustada.
- SSE tiene que ver con la variación residual que queda sin explicar por la recta de regresión ajustada.
- En virtud de la ecuación (3), se establece también la siguiente identidad para los grados de libertad ($g.l$) de las sumas de cuadrados:

$$\underbrace{g.l\left(SST\right)}_{n-1}=\underbrace{g.l\left(SSR\right)}_{1}+\underbrace{g.l\left(SSE\right)}_{n-2}. \tag{4}$$

Análisis de varianza para probar la significancia de la regresión

Sujetos a los supuestos del modelo de regresión, se cumple que:

1. Cuando $\beta_1 = 0$, $SSR/\sigma^2 \sim \chi^2_1$.
2. $SSE/\sigma^2 \sim \chi^2_{n-2}$.
3. SSR/σ^2 y SSE/σ^2 son estadísticamente independientes.
4. De las anteriores propiedades podemos decir que bajo $H_0 : \beta_1 = 0$, el estadístico,

$$F_0 = \frac{SSR/g.l(SSR)}{SSE/g.l(SSE)} = \frac{SSR/1}{SSE/(n-2)} = \frac{SSR}{MSE} \sim f_{1,n-2}. \quad (5)$$

5. F_0 de la ecuación (5) es igual al cuadrado del estadístico T del test de significancia de la pendiente β_1, es decir,

$$F_0 = \frac{SSR}{MSE} = \left(\frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}\right)^2 = \left(\frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}}\right)^2 \sim f_{1,n-2}. \quad (6)$$

Consideraciones del estadístico F_0

Notas sobre $F_0 = \frac{SSR}{MSE} = \left(\frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}}\right)^2 \sim f_{1,n-2}$:

- En el caso de la regresión lineal simple, la prueba sobre la significancia de la regresión es equivalente a probar si la pendiente de la recta es significativamente diferente de cero, es decir, el test puede realizarse de dos maneras: test t para β_1 o mediante el análisis de varianza. Por tanto, la conclusión obtenida por el análisis de varianza debe ser la misma que la obtenida cuando se prueba la significancia individual de β_1.
- En el test ANOVA el criterio de decisión a un nivel de significancia α es usando un valor crítico $f_{\alpha,1,n-2}$, con el cual rechazamos la hipótesis nula de que la variabilidad en la variable respuesta es debida sólo al error aleatorio (para aceptar la hipótesis de que la regresión en X es significativa), si $F_0 > f_{\alpha,1,n-2}$, donde $f_{\alpha,1,n-2}$ es tal que $P\left(f_{1,n-2} > f_{\alpha,1,n-2}\right) = \alpha$.

Consideraciones del estadístico F_0

Notas sobre $F_0 = \frac{SSR}{MSE} = \left(\frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}}\right)^2 \sim f_{1,\,n-2}$:

- En el test ANOVA también podemos evaluar el valor P de la prueba (significancia más pequeña, según los datos, que conduce al rechazo de H_0) el cual es igual a $P\left(f_{1,\,n-2} > F_0\right)$, y si es "pequeño", entonces se rechaza la hipótesis:

$$H_0 : \text{"El modelo lineal de } Y \text{ en } X \text{ no es significativo para explicar la variabilidad de } Y\text{"}$$
$$\Leftrightarrow$$
$$H_0 : \beta_1 = 0.$$

vs.

$$H_1 : \text{"El modelo lineal de } Y \text{ en } X \text{ es significativo para explicar la variabilidad de } Y\text{"}$$
$$\Leftrightarrow$$
$$H_1 : \beta_1 \neq 0. \tag{7}$$

Tabla ANOVA

El análisis de varianza suele presentarse en forma de tabla (tabla 1), conocida como la *tabla ANOVA*, donde los cuadrados medios corresponden a las sumas de cuadrados que allí se discriminan, divididas por sus respectivos grados de libertad.

Tabla 1: Tabla de análisis de varianza del modelo de regresión lineal simple.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico F_0
Regresión	SSR	1	$MSR = SSR/1$	MSR/MSE
Error	SSE	$n-2$	$MSE = SSE/(n-2)$	
Total	SST	$n-1$	$MST = SST/(n-1)$	

Nota sobre *MSR* y *MSE*:

Los cuadrados medios esperados están dados por $E(MSR) = \sigma^2 + \beta_1^2 S_{xx}$ y $E(MSE) = \sigma^2$.

Contenido

1. Análisis de varianza para probar la significancia de la regresión
2. Coeficiente de determinación R^2
3. Referencias

Coeficiente de determinación R^2

El coeficiente de determinación muestral, denotado por R^2, se calcula como

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \tag{7}$$

por tanto, podemos interpretarlo como la proporción de la variabilidad total observada en la variable respuesta, que es explicada por la relación lineal con la variable predictora considerada.

La ecuación (7) ha sido utilizada erróneamente como medida para evaluar la bondad del ajuste lineal, pues si bien valores cercanos a 1 indican una mayor asociación lineal, no necesariamente garantiza que los supuestos básicos del modelo lineal se estén cumpliendo y menos que no haya carencia de ajuste lineal.

Consideraciones del coeficiente de determinación

Notas sobre R^2:

- Creer que un R^2 alto indica que el modelo puede hacer predicciones útiles. Hay casos donde se tiene un R^2 alto y sin embargo, los intervalos de predicción son muy amplios indicando poca precisión del pronóstico.
- Creer que un R^2 alto indica que la recta de regresión ajustada tiene buen ajuste. Hay casos en los cuales se ajusta una recta obteniendo un R^2 cuando la verdadera relación no es lineal.
- Creer que un R^2 cercano a cero indica que X e Y no están relacionados. Cuando existe una relación no lineal entre X e Y, puede ocurrir que al ajustar considerando linealidad, el R^2 dé cercano a cero.

Contenido

1. Análisis de varianza para probar la significancia de la regresión
2. Coeficiente de determinación R^2
3. Referencias

Referencias

Kutner, M. H., Nachtsheim, C. J., Neter, J., y Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill Irwin, New York, quinta edición.

Montgomery, D. C., Peck, E. A., y Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley, New Jersey, quinta edición.

Álvarez, N. G. (2022). Notas de Clase Análisis de Regresión - 3006918, Capítulo 2: Regresión Lineal Simple. Notas no publicadas.

Álvarez, N. G. y Gómez, C. M. L. (2018). Notas de Clase - Estadística II (3006918): Análisis de Regresión Lineal e Introducción al Muestreo.