Análisis de Regresión Lineal Múltiple

Juan Sosa, PhD

Universidad Externado de Colombia

I - 2018

Objetivo

- Caracterizar o explicar el valor medio de una variable [dependiente] bajo condiciones específicas de otra(s) variable(s) [independientes].

Observaciones

- Establecer **asociación**, mas **no causalidad**.
- Aunque predecir [el valor promedio] no es el centro de atención, pero es posible hacerlo.
- Modelo [de carácter probabilístico] con muchas posibilidades!

Estructura del conjunto de datos

- Matriz de datos $\mathbf{X}$ de tamaño $n \times p$: variables independientes/explicativas.
- Vector de datos $\boldsymbol{y}$ de tamaño $n \times 1$: variable dependiente/explicada/respuesta.

Formulación

$$
\begin{aligned}
y_1 &= \beta_1 x_{1,1} + \beta_2 x_{1,2} + \ldots + \beta_p x_{1,p} + \epsilon_1 \\
y_2 &= \beta_1 x_{2,1} + \beta_2 x_{2,2} + \ldots + \beta_p x_{2,p} + \epsilon_2 \\
&\vdots \\
y_n &= \beta_1 x_{n,1} + \beta_2 x_{n,2} + \ldots + \beta_p x_{n,p} + \epsilon_n
\end{aligned}
$$

$$
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}
$$

$$
\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}
$$

Modelo (cont.)

Terminología

- x : variables independientes/explicativas.
- y : variable dependiente/explicada/respuesta.
- β : coeficientes de regresión.
- ϵ : errores/perturbaciones aleatorias.

Supuestos

$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ donde $\mathbf{I}_n$ es la matriz identidad de tamaño n.

Observaciones

- Los parámetros del modelo son $\beta_1, \beta_2, \ldots, \beta_p$ y σ^2.
- Las variables explicativas no tienen carácter aleatorio y pueden ser continuas o categóricas.
- El supuesto de normalidad y varianza constante no es una "camisa de fuerza".
- El carácter aleatorio del modelo recae en la variable dependiente a través del error.
- Los errores se asumen como independientes.

Implicaciones del modelo

Como $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ se obtiene que $\boldsymbol{y} \mid \mathbf{X} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ y además

$$\mathbb{E}\left[y_i \mid \boldsymbol{x}_i\right] = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} = \boldsymbol{\beta}^T \boldsymbol{x}_i$$
$$\mathbb{V}\mathsf{ar}\left[y_i \mid \boldsymbol{x}_i\right] = \sigma^2$$

Interpretación de β_i

Cambio en y correspondiente a una unidad de cambio en x_i cuando el resto de las variables explicativas se mantienen constantes.

Consecuencias

- El modelo caracteriza el valor medio de y para valores dados de x.
- La linealidad se encuentra en los coeficientes de regresión.
- La interpretación de los coeficientes no está dada en términos de cargas/pesos.

Estimación

Encontrar los valores de $\boldsymbol{\beta}$ y σ^2 para reproducir $\boldsymbol{y}$ tan precisamente como sea posible.

Método de máxima verosimilitud

Encontrar los valores de $\boldsymbol{\beta}$ y σ^2 que maximicen (optimicen) la función de verosimilitud:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2} \left(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i \right)^2 \right\}$$

Método de mínimos cuadrados ordinarios

Encontrar los valores de $\boldsymbol{\beta}$ que minimicen (optimicen) la función cuadrática:

$$\sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}^T \boldsymbol{x}_i \right)^2$$

Estimador de $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$$

$$\hat{\boldsymbol{\beta}} \sim \mathsf{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$$

Estimación (cont.)

Valores ajustados/predichos

$$\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = \mathbf{H}\boldsymbol{y}$$

Estimador del valor medio de $\boldsymbol{y}$.

Residuales

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \mathbf{I}_n\boldsymbol{y} - \mathbf{H}\boldsymbol{y} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{y}$$

Permiten evaluar la **bondad de ajuste** y los **supuestos** del modelo.

Sumas de cuadrados

La variabilidad total de la variable respuesta se descompone en la variabilidad debida a los valores ajustados y la variabilidad debida a los residuales:

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

$$SCT = SCR + SCE$$

Observación: SCE es la cantidad que se debe minimizar para ajustar el modelo.

Estimación (cont.)

Coeficiente de determinación

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

Proporción de la variabilidad total que explica el modelo ajustado.

Coeficiente de determinación ajustado

$$R_a^2 = 1 - \frac{n-1}{n-p}(1 - R^2)$$

Estimador insesgado de σ^2

$$\hat{\sigma}^2 = \frac{SCE}{n-p} = CME$$

CME : cuadrado medio del error.

Observación

Para apreciar el valor agradado de una regresión se recomienda comparar s_y con $\hat{\sigma}$.

Significancia global

Existe una relación significativa entre la variable respuesta y las variables explicativas en general?

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{frente a} \quad H_1 : \beta_j \neq 0 \text{ para algún } j$$

Estadístico de prueba: $F = \frac{SCR/(p-1)}{SCE/(n-p)} = \frac{CMR}{CME} \sim F_{p-1,n-p}$

Significancia particular

Existe una relación significativa entre la variable respuesta y una variable explicativa en particular?

$$H_0 : \beta_i = 0 \quad \text{frente a} \quad H_1 : \beta_i \neq 0$$

Estadístico de prueba: $t = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \sim t_{n-p}$

Intervalo de confianza: $\hat{\beta} \pm t_{n-p}\, s_{\hat{\beta}_i}$

Observación promedio para $\boldsymbol{x}_0$

$$\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_0 \pm t_{n-p}\, \hat{\sigma} \sqrt{\boldsymbol{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \boldsymbol{x}_0}$$

Observación particular para $\boldsymbol{x}_0$

$$\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_0 \pm t_{n-p}\, \hat{\sigma} \sqrt{1 + \boldsymbol{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \boldsymbol{x}_0}$$

Observación

La incertidumbre es mayor cuando se predice una observación particular que al predecir la respuesta media.

Multicolinealidad

Correlación entre las variables explicativas, i.e, información redundante en el modelo.

Observaciones

- Sobre carga de información.
- Una correlación absoluta mayor a 0.7 sugiere problemas de este estilo.
- Hay un aumento del error estándar de los coeficientes de regresión.
- Se hace difícil determinar la significancia de las variables explicativas en particular.

Validación del modelo

Supuestos

- $\mathbb{E}[\epsilon_i] = 0$.
- $\mathbb{V}\text{ar}[\epsilon_i] = \sigma^2$.
- La distribución de los errores es normal.
- Los errores son mutuamente independientes.

Residuales estandarizados

$$r_i^* = \frac{r_i}{s_{r_i}} = \frac{r_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

Recomendaciones

- Histograma y gráfico cuantil-cuantil (*qqplot*) de los residuales (normalidad).
- Graficar residuales (estandarizados) frente a valores predichos (varianza constante).
- Graficar residuales (estandarizados) frente a cada variable explicada (tendencias).

Outliers

Definición

Observación que diverge la tendencia general de la data. Hacen crecer el CME y por lo tanto los margenes de error de los intervalos de confianza.

Residuales estudentizados

$$\tilde{r}_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}$$

Detección

Una observación se considerada (potencialmente) como outlier si:

$$|r_i^*| > 2 \quad \text{o} \quad |\tilde{r}_i| > 3$$

Observación influyentes

Definición

Observación que cambia sustancialmente el análisis cuando hace parte de la data.

DFFITS (*difference in fits*)

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$$

Distancia de Cook

$$D_i = \frac{(y_i - \hat{y}_i)^2}{p\,\hat{\sigma}^2}\,\frac{h_{ii}}{(1-h_{ii})^2}$$

Detección

Una observación se considerada (potencialmente) como influyente si:

$$DFFITS_i > 2\sqrt{\tfrac{p}{n-p}} \quad \text{o} \quad D_i > 1$$

Observaciones

Selección de variables

- *Forward regression*.
- *Backward regression*.
- *Stepwise regression*.

Comparación de modelos

- Criterio de información de Akaike (AIC).
- Criterio de información Bayesiano (BIC)
- Pruebas de hipótesis para modelos anidados.

Todo tipo de variables explicativas

- Variables dummy.
- Variables polinómicas.

Alternativas

- Transformar la variable respuesta (logaritmo, raíz cuadrada, Box-Cox).
- Regresión lineal general.
- Regresión rígida.
- Regresión Lasso.
- Regresión ElasticNet.
- Regresión no lineal.
- Regresión no paramétrica.
- Regresión cuantílica.
- Regresión robusta.
- Modelos lineales generalizados.