CARRERA DE INGENIERÍA DE SISTEMAS

Análisis de Regresión Lineal Múltiple

ESTADÍSTICA APLICADA
PROF. MSC. VILMA S. ROMERO ROMERO
5TO CICLO | 21-06-2022

UNIVERSIDAD DE LIMA

Modelo de Regresión Lineal Múltiple

Extensión de la Regresión Lineal Simple, donde la diferencia radica en que se puede considerar dos o más variables regresoras o independientes en el modelo.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- k es la cantidad de variables predictoras.
- X_j representa la `j-ésima` variable predictora.
- β_j son los parámetros desconocidos a estimar.
  - $\boldsymbol{\beta_j}$: Cambio promedio o esperado en Y debido al incremento en una unidad de X_j manteniendo todas las demás variables predictoras constantes.
- ϵ es el error aleatorio.

UNIVERSIDAD DE LIMA

Generalmente, cuando se desarrolla una regresión lineal múltiple, uno está interesado en abarcar las siguientes preguntas:

- ¿Al menos uno de las variables predictoras X_1, X_2,. . ., X_k resulta significativa en la predicción de la respuesta?
- ¿Todos las variables predictoras ayudan a explicar Y, o solamente un subconjunto de las variables predictores son útiles?
- ¿Qué tan bueno ajusta el modelo a los datos?
- Dado un conjunto de valores de las variables predictoras, qué valor de respuesta se debería predecir, y qué tan precisa es la predicción?

UNIVERSIDAD DE LIMA

Estimación de Coeficientes

La estimación de los coeficientes $\hat{\beta}_0$, $\hat{\beta}_1 \ldots \hat{\beta}_k$ se realiza también mediante el método de Mínimos Cuadrados Ordinarios, que busca minimizar la suma de cuadrados de los residuales.

Tabla de Análisis de Varianza (ANOVA)

La variación total de la variable respuesta con respecto a su media se puede descomponer en:

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

Variabilidad alrededor de la media	=	Variabilidad debido a la regresión	+	Variabilidad debido al error
Suma de Cuadrados Total (SCT)	=	Suma de Cuadrados de la Regresión (SCR)	+	Suma de Cuadrados del Error (SCE)

Tabla de Análisis de Varianza (ANOVA)

Fuente de variación	Grados de libertad	Suma de Cuadrados	Cuadrado Medio	F_0	P-value
Regresión	k	SCR	$CMR = \frac{SCR}{k}$	$F_0 = \frac{CMR}{CME}$	$P(F > F_0)$
Error	$n - k - 1$	SCE	$CME = \frac{SCE}{n-k-1}$		
Total	$n - 1$	SCT			

UNIVERSIDAD DE LIMA

Prueba de Significación del modelo - Prueba F

Evaluar si la el modelo de regresión lineal múltiple, con las variables independientes utilizadas, es apropiado o no.

1) Hipótesis:
   $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$
   (Ninguna de las variables predictoras ayuda a explicar la variación en Y)
   H_1 : Al menos un $\beta_i \neq 0 \quad i = 1, \ldots, k$
   (Al menos una de las variables predictoras ayuda a explicar la variación en Y)
2) Especificar el nivel de significación α.
3) Calcular el valor del estadístico de prueba: $F_0 = \dfrac{CMR}{CME} \sim F_{(k,n-k-1)}$
4) Región crítica y regla de decisión:
   $$RC = \left\langle F_{(k,n-k-1,1-\alpha)}; \infty \right\rangle \Rightarrow \textbf{Rechazar } H_0 \text{ si } F_0 > F_{(k,n-k-1,1-\alpha)}$$

- Se puede calcular y utilizar el P-value:
  $$\text{P-value} = P(F_{(k,n-k-1)} > F_0) \Rightarrow \textbf{Rechazar } H_0 \text{ si P-value} \leq \alpha$$

UNIVERSIDAD DE LIMA

Prueba de Significación del modelo - Prueba F

- Al determinar que uno de los regresores es significativo, el siguiente paso es determinar cuál es.
- Se puede evaluar el *efecto parcial* de cada variable cuando se agrega al modelo mediante el estadístico t como se explicó en la regresión lineal simple.

UNIVERSIDAD DE LIMA

Prueba individual de las variables - Prueba T

1) Hipótesis:

   $H_0 : \beta_i = 0$ (la variable X_i no influye en el modelo)

   $H_1 : \beta_i \neq 0$ (la variable X_i influye en el modelo)

2) Especificar el nivel de significación α.

3) Calcular el valor del estadístico de prueba: $t_0 = \dfrac{\hat{\beta}_i}{es(\hat{\beta}_i)} \sim t_{(n-k-1)}$

4) Región crítica y regla de decisión:

$$RC = \left\langle -\infty; t_{(n-k-1,\alpha/2)} \right\rangle \cup \left\langle t_{(n-k-1,1-\alpha/2)}; \infty \right\rangle$$

$$\Rightarrow \textbf{Rechazar } H_0 \text{ si } t_0 < t_{(n-k-1,\alpha/2)} \text{ o } t_0 > t_{(n-k-1,1-\alpha/2)}$$

- Se puede calcular y utilizar el P-value:

$$\text{P-value} = 2 \times P(t_{(n-k-1)} > t_0) \Rightarrow \textbf{Rechazar } H_0 \text{ si P-value} \leq \alpha$$

Coeficiente de Determinación

Representa la proporción de la variabilidad en Y que puede explicarse por el conjunto de variables $X_1, X_2, X_3, \ldots, X_k$.

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- $0 \leq R^2 \leq 1$.
- Cuanto menor sea el valor de R^2 (valores cercanos a 0), peor será el ajuste del plano de regresión a los datos.
- Cuanto mayor sea el valor de R^2 (valores cercanos a 1), mejor será el ajuste del plano de regresión a los datos.

UNIVERSIDAD DE LIMA

Coeficiente de Determinación Ajustado

- Al añadir más variables al modelo, el R^2 siempre va aumentar. Por eso, es recomendable utilizar el R^2 ajustado en su defecto.

$$\boxed{R^2_{Aj} = 1 - (1 - R^2) \times \left(\frac{n-1}{n-k-1}\right)}$$

Adecuación del Modelo

Para determinar si el modelo es correcto y no inestable, se debe considerar:

- La relación entre la variable respuesta y las variables explicativas es lineal, al menos de manera aproximada.
- El término del error ϵ tiene media cero y varianza σ^2 constante.
- los errores no están correlacionados.
- Los errores tienen distribución normal.

Recordar que un residual está definido como:

$$e_j = y_j - \hat{y}_j \quad j = 1, 2, \ldots, n$$

Para comprobar las premisas anteriores, el análisis gráfico de los residuales resulta una forma muy efectiva.

Evaluación de los Supuestos

- **Normalidad de los errores**
  i) Hipótesis:
     H_0 : Los errores siguen una distribución normal
     H_1 : Los errores no siguen una distribución normal
  ii) $\alpha = 0.05$.
  iii) p-value de la prueba de Anderson-Darling.
  iv) Si p-value$> \alpha \Rightarrow$ Los errores siguen una distribución normal.

- **Supuesto de no multicolinealidad**
  - Si los valores de los factores de inflación de varianza son menores a 5, $VIF < 5 \Rightarrow$ No existe multicolinealidad entre las variables regresoras.

UNIVERSIDAD DE LIMA

Recursos Adicionales I

Devore, J. (2019). *Introducción a la probabilidad y estadística para ingeniería y ciencias*. Cengage, 1 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_CENGAGE.

Johnson, R. A. (2012). *Probabilidad y estadística para ingenieros*. Pearson Educación, 8 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_PEARSON.

Kokoska, S. (2015). *Introductory Statistics*. W. H. Freeman and Company, 2 edition.

Mendenhall, W., Beaver, R. J., and Beaver, B. M. (2015). *Introducción a la probabilidad y estadística*. Cengage, 14 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_CENGAGE.

Recursos Adicionales II

Millones, R., Barreno, E., Vásquez, F., and Castillo, C. (2017). *Estadística Descriptiva y Probabilidades: Aplicaciones en la ingeniería y los negocios*. Lima: Fondo Editorial de la Universidad de Lima, 1 edition. Código Biblioteca U.Lima: 519.53 E.

Triola, M. (2018). *Estadística*. Pearson Educación, 12 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_PEARSON.

UNIVERSIDAD DE LIMA