Estadística II - Taller 02 Semestre: 2024-01

Profesores: Johnatan Cardona Jimenez, Freddy Hernández Barajas, Raul Alberto Perez

Monitor: Ronald Palencia

Parte teorica

Dado el contenido que se ha visto hasta el momento, el taller será mayormente teórico y operativo.

1) ¿Qué representa el intercepto β_0 en un modelo de regresión lineal múltiple?

A) El cambio en la variable respuesta por cada unidad de cambio en las variables predictoras.

B) El valor promedio de la variable respuesta cuando todas las variables predictoras son cero.

C) La varianza de los errores del modelo.

D) El coeficiente de determinación del modelo.

Respuesta: B) El valor promedio de la variable respuesta cuando todas las variables predictoras son cero.

El intercepto β_0 representa la respuesta media de Y cuando todas las variables predictoras $X_1, X_2, \ldots, X_k$ son cero. Si esta combinación de valores cero para las variables predictoras no se observa en el rango experimental, entonces β_0 no es interpretable. Matemáticamente, β_0 es el valor de Y cuando $X_1 = X_2 = \cdots = X_k = 0$.

2) En la regresión lineal múltiple, ¿qué representa β_j (para $j > 0$)?

A) El cambio en la variable respuesta por cada unidad de incremento en la variable predictora X_j, manteniendo constantes las otras variables predictoras.

B) La varianza de la variable predictora X_j.

C) El error estándar asociado a la variable predictora X_j.

D) La correlación entre la variable respuesta y la variable predictora X_j.

Respuesta A

Los coeficientes β_j representan el cambio esperado en la variable de respuesta Y por cada unidad de cambio en la variable predictora X_j, manteniendo todas las otras variables predictoras constantes. Esto refleja el efecto parcial de X_j sobre Y, suponiendo que todas las demás variables en el modelo permanecen fijas.

3) ¿Qué implica la independencia de los errores ϵ_i en un modelo de regresión lineal múltiple?

A) Que los errores no tienen varianza.

B) Que el valor esperado de los errores es cero.

C) Que los errores de una observación no están correlacionados con los de otra.

D) Que los errores siguen una distribución uniforme.

Respuesta C

La independencia de los errores ϵ_i implica que los errores en las observaciones son independientes entre sí, lo que significa que el error en una observación no está correlacionado con el error en cualquier otra observación. Esta es una suposición crítica que ayuda a asegurar que el modelo sea bien especificado y que las inferencias estadísticas basadas en el modelo sean válidas.

4) ¿Qué problema se indica con la multicolinealidad en un modelo de regresión lineal múltiple?

A) Correlación cero entre las variables predictoras.

B) Fuerte asociación lineal entre dos o más variables predictoras.

C) Varianza infinita de los estimadores de los parámetros.

D) Coeficientes de regresión negativos.

Respuesta B La multicolinealidad se refiere a una fuerte asociación lineal entre dos o más variables predictoras en el modelo. Puede llevar a estimaciones de coeficientes inestables y significancias estadísticas cuestionables, ya que se vuelve difícil aislar el efecto individual de cada variable predictora sobre la variable respuesta. La multicolinealidad puede detectarse mediante el análisis de la matriz de correlación entre las variables predictoras o el cálculo de factores como el factor de inflación de la varianza (VIF).

5) Los estimadores de mínimos cuadrados son insesgados en un modelo de regresión lineal multiple.

A) Falso

B) Verdadero

Verdadero. Para un estimador $\hat{\beta}$ de mínimos cuadrados, se tiene que $E[\hat{\beta}] = \beta$. Esto se debe a que el estimador $\hat{\beta} = (X^T X)^{-1} X^T y$ es una función lineal de y, y bajo las suposiciones del modelo de regresión lineal, el valor esperado de y es $X\beta$. Por lo tanto, $E[\hat{\beta}] = (X^T X)^{-1} X^T E[y] = (X^T X)^{-1} X^T X\beta = \beta$, lo que muestra que el estimador es insesgado.

6) La homoscedasticidad es necesaria para que los estimadores de mínimos cuadrados sean eficientes.

A) Verdadero

B) Falso

Verdadero. La homoscedasticidad, que implica varianzas constantes de los errores ϵ_i para todas las observaciones, es crucial para la eficiencia de los estimadores de mínimos cuadrados en el sentido de mínima varianza en la clase de estimadores lineales insesgados. Sin homoscedasticidad (es decir, si hay heteroscedasticidad), los estimadores de mínimos cuadrados pueden no ser los más eficientes.

7) Todas las entradas de la matriz de correlaciones son menores a uno

A) Falso

B) Verdadero

Es falso porque las entradas de la diagonal principal son iguales a 1.

Parte practica

Un ingeniero realizó un experimento para determinar el rendimiento total del aceite por lote de cacahuate, para ello tuvo en cuenta variables como la presión, temperatura del CO2 aplicado, la humedad y el tamaño de partícula de los cacahuates. Los 16 datos recolectados aparecen a continuación

Aquí hay un breve resumen del conjunto de datos:

X1: Presión

X2: Temperatura

X3: Humedad

X4: Tamaño de particula

y : Rendimiento

Table 1: Primeros 5 registros

X1	X2	X3	X5	y
415	25	5	1.28	63
550	25	5	4.05	21
415	95	5	4.05	36
550	95	5	1.28	99
415	25	15	4.05	24
550	25	15	1.28	66

a) Calcule la matriz de varianzas-covarianzas.

b) Calcule la matriz de correlaciones.

c) Escriba un modelo con las covariables en forma escalar.

d) Añada una columna de unos al principio de los datos (excluyendo la covariable), de ahora en adelante dicha matriz será nombrada X

e) Calcule las matrices X^TX, $(X^TX)^{-1}$, $(X^TX)^{-1}(X^Ty)$, $X(X^TX)^{-1}(X^Ty)$, $y - X(X^TX)^{-1}(X^Ty)$.

a)

Calculando la matriz de varianzas-covarianzas

```
#con la funcion var se calcula la matriz de varianzas-covarianzas
var(datos)
```

```
     X1        X2        X3         X5          y
X1 4860    0.0000  0.000000   0.000000 270.000000
X2    0 1306.6667  0.000000   0.000000 368.666667
X3    0    0.0000 26.666667   0.000000   3.333333
X5    0    0.0000  0.000000   2.046107 -32.870667
y   270  368.6667  3.333333 -32.870667 690.866667
```

```
#tambien se puede hacer cov(datos)
```

b)

Se procede a calcular la matriz de correlaciones

```
cor(datos) #se hace con la funcion cor
```

```
           X1         X2          X3          X5           y
X1 1.0000000 0.0000000 0.00000000  0.0000000  0.14734945
X2 0.0000000 1.0000000 0.00000000  0.0000000  0.38802021
X3 0.0000000 0.0000000 1.00000000  0.0000000  0.02455824
X5 0.0000000 0.0000000 0.00000000  1.0000000 -0.87427338
y  0.1473494 0.3880202 0.02455824 -0.8742734  1.00000000
```

c)

Se escribe el modelo de regresión lineal múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_5 x_{5i} + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

d)

```
datosmat <- as.matrix(datos[, -5]) #excluyendo la respuesta
X <- cbind("intercepto" = 1, datosmat) #añadiendo columna de unos
y <- as.matrix(datos$y) #extrayendo la respuesta
```

e)

```
first <- t(X) %*%  X #multiplicando X^t y X
second <- solve(first) #invirtiedo lo de arriba
third <- second %*% t(X) %*% y #estimacion de los parametros
fourth <- X %*% third #estimacion de la respuesta

fifth <- y - fourth #residuales
third
```

```
                [,1]
intercepto  52.07904991
X1           0.05555556
```

```
X2           0.28214286
X3           0.12500000
X5         -16.06498195
```

usando lm

```r
mod <- lm(y ~ X1 + X2 + X3 + X5, data = datos) #ajustando el modelo
#se puede usar y ~ .

summary(mod)
```

```
Call:
lm(formula = y ~ X1 + X2 + X3 + X5, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max
-12.250  -4.438   0.125   5.250   9.500

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.07905   15.22756   3.420 0.005723 **
X1            0.05556    0.02848   1.951 0.077041 .
X2            0.28214    0.05493   5.137 0.000325 ***
X3            0.12500    0.38450   0.325 0.751207
X5          -16.06498    1.38809 -11.573 1.69e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.69 on 11 degrees of freedom
Multiple R-squared:  0.9372,    Adjusted R-squared:  0.9144
F-statistic: 41.06 on 4 and 11 DF,  p-value: 1.503e-06
```