Regresión lineal 2

Wednesday, March 23, 2022 2:11 PM

Regresión lineal 2

Descomposición de suma de cuadrados.

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

Suma de cuad. totales | Suma de cuadr. de la regre° | Suma de cuadr. residuales

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

R^2 me dice q' tanto la variabilidad de una var me explica a la otra.
Entre más significativa la var. escogida, más va a mejorar mi R^2.

$R^2 = 1$ sii Error $= 0 = \hat{\varepsilon}_j$

$R^2 = 0$ cuando mi predic° es exactam. $\bar{y}$. (no sé re nada) $\beta_i = 0$.

Inferencia en el Modelo de regresión

$E[Y] = \beta_0 + \beta_1 z_1 + \cdots + \beta_n z_n$

Estudiamos la ~ muestral de $\hat{\beta}$ y de $\hat{\varepsilon}'\hat{\varepsilon}$ suponiendo ε normal.

$Y = Z\beta + \varepsilon$, Z rango completo

$\varepsilon \sim N_n(0, \sigma^2 I)$: el estimador es max

Homocedasticidad: varianzas iguales.
errores normales, incorr y con homocedasticidad.

- El estimador de máx verosimilitud de β es $\hat{\beta}$

$$\hat{\beta} = (Z'Z)^{-1} Z Y \sim N_{r+1}(\beta, \sigma^2 (Z'Z)^{-1})$$

- Región confianza $(1-\alpha)$ para β

n: # obs
r: # vars.

$$(\beta - \hat{\beta})' Z'Z (\beta - \hat{\beta}) \leq (r+1) s^2 F_{r+1, n-r-1}(\alpha)$$

$$s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{(n-r-1)}$$

IC simultáneo para los β_i

$\widehat{var}(\hat{\beta}_i)$ = i-ésimo elem de la diagonal de $s^2 (Z'Z)^{-1}$

$$\hat{\beta}_i \pm \sqrt{\widehat{var}(\hat{\beta}_i)} \sqrt{(r+1) F_{r+1, n-r-1}(\alpha)}$$

I.C 1 a 1 con confianza $(1-\alpha)$

$$\hat{\beta}_i \pm t_{n-r-1}\left(\frac{\alpha}{2}\right) \sqrt{\widehat{var}(\hat{\beta}_i)}$$

> si el 0 está incluído en el intervalo, β_i puede ser 0 y por tanto, no es tan significativa y podría sacarla de mi modelo.

Región de confianza

↳ Está centrada en $\hat{\beta}$ y la orientac° y la distancia dependen de los vecs. y vals. propios de $Z'Z$.

Ejemplo 2 vars.

$\hat{\beta} = \begin{bmatrix} 30.96 \\ 6.63 \\ 0.045 \end{bmatrix}$ β_i es la pendiente de la var 1, mientras la var 2 está constante.

Modelo ANOVA (lm)

H_0: $\beta_0 = \beta_1 = \cdots = \beta_r = 0$.

Hago pruebas de hip. para c/ $\beta_i = 0$

c/:	T para Ho	Prob > |T|
int.	3.93	0.001
z_1	3.53	0.004
z_2	0.16	0.87

β_2 no me ayuda a explicar el comportam. β_2 no significativa. CHAO! z_2.
↳ Si hiciera el IC, el 0 va a estar ahí.

Inferencia sobre parámetros del modelo.

Parte del análisis de reg. incluye inferencias sobre algunos params. del mod.

Ej: β_i's = 0 (no influyen esas vars.)

Denotamos esos predictores $z_{q+1}, \ldots, z_r$. $(\beta_{(2)})_{(q+1)\times 1}$

P.H: $H_0 \Rightarrow \beta_{q+1} = \dots = \beta_r = 0 \Leftrightarrow \beta(2) = 0$. $\quad \beta = \begin{pmatrix} \beta(1) \\ \beta(2) \end{pmatrix}_{(r-q)\times 1}$

$Z = (Z(1) \; ; \; Z(2))$ $\quad n\times(q+1)$ $\quad n\times(r-q)$

$q+1$ pues incluye a β_0

El modelo lineal se vuelve: $Y = Z\beta + \varepsilon = (Z(1) \,|\, Z(2)) \begin{bmatrix} \beta(1) \\ \beta(2) \end{bmatrix} + \varepsilon$

Evaluar H_0 con extra sum of squares (ESS)

$0 \le ESS = SSres(Z(1)) - SSres(Z)$ $\quad$ ▲ ESS = grande → este modelo restringido es peor, hay muchas cosas q' no se están explicando → Es mejor el mod 06.

suma de cuad. residuales del mod solo con las vars de $Z(1)$

$\hookrightarrow \sum \hat{\varepsilon}_i^2(1) - \sum \hat{\varepsilon}_i^2$

$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4 + \beta_r z_r + \varepsilon$

$Y(1) = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_r z_r + \varepsilon$

$\hookrightarrow SSres(Z(1))$

$ESS = SSres(Z(1)) - SSres(Z)$

$= (Y - Z(1)\hat{\beta}(1))'(Y - Z(1)\hat{\beta}(1)) - (Y - Z\hat{\beta})'(Y - Z\hat{\beta})$

donde $\hat{\beta}(1) = (Z(1)'Z(1))^{-1} Z(1)'Y$

Sea Z con rango completo, $\varepsilon \sim N_n(0, \sigma^2 I)$

La prueba $H_0: \beta(2) = 0$ equivale a una prueba q' usa ESS y $S^2 = \dfrac{(Y - Z\hat{\beta})'(Y - Z\hat{\beta})}{n-r-1}$ → grande cuando n es pequeño

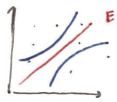
Concretam., se rechaza H_0 si

$$\frac{\frac{ESS}{r-q}}{S^2} > F_{r-q,\,n-r-1}(\alpha)$$

Para realizar la P.H, se construye el mod. con o sin estas vars (o términos)

Las sumas de cuad. residuales se comparan en ambos casos.

Inferencia de las predicciones



El $E[Y]$ lo queremos predecir y este también tiene su I.C.

Una vez establecido el mod. se pueden hacer predicciones.

Sea $z_0 = \begin{bmatrix} 1 \\ z_{01} \\ \vdots \\ z_{0r} \end{bmatrix}$ vals de vars. predictoras.

Dados estos vals z_0 y $\hat{\beta}$ se puede estimar el val de $E[Y_0]$ o de Y_0.

Para $E[Y]$: Sea Y_0 la v.a respuesta correspondiente a z_0

$E[Y_0 | z_0] = \beta_0 + \beta_1 z_{01} + \dots + \beta_r z_{0r} = z_0'\beta$

* El MLE (max likely.. estimator) de $z_0\beta$ es el $z_0'\hat{\beta}$
* $z_0'\hat{\beta}$ es el MVUE (min varianza unbiased estimator) de $E[Y_0|z_0]$
* Su varianza $var(z_0'\hat{\beta}) = z_0'(Z'Z)^{-1} z_0 \sigma^2$

Si ε normales, el IC $(1-\alpha)100\%$ para $E[Y_0|z_0] = z_0'\beta$ es:

* $\boxed{z_0'\hat{\beta} \pm t_{n-r-1}(\alpha/2)\sqrt{(z_0'(Z'Z)^{-1}z_0)\,S^2}}$ → plano/recta de reg. entre q' vals de tú.

Para Y_0: Tenemos $Y_0 = z_0'\hat{\beta} + \varepsilon_0$ $\quad \varepsilon_0 \sim N(0, \sigma^2)$

$var(\varepsilon_0) = \sigma^2(1 + z_0'(Z'Z)^{-1}z_0)$

El IC con conf. $(1-\alpha)$

* $\boxed{z_0'\hat{\beta} \pm t_{n-r-1}(\frac{\alpha}{2})\sqrt{S^2(1 + z_0'(Z'Z)^{-1}z_0)}}$ → (siempre es mayor q' el de $E[Y]$)

Donde está estaval, varía más q' donde está mi recta. ← Dado un val, entre q' y q' espero q' esté

Verifica° de los mod.

Podemos verificar si el modelo es adecuado antes de usarlo para realizar predicciones, inferencias, etc.

Si el modelo es válido, $\hat{\varepsilon}$ es un estimador de

$\varepsilon_n \sim N_n(0, \sigma^2 I)$ $\quad$ en este caso:

$$\hat{\varepsilon} \sim N(0, \sigma^2(I - \underbrace{Z(Z'Z)^{-1}Z'}_{H}))$$

Pueden variar las varianzas de $\hat{\varepsilon}$ si los elems de la diag. de H, h_{ii} son muy difs.

si al graficar los errores veo un patrón, no son indeps. de -
si hay zonas con mayores errores → ¡!!.

- Se puede usar:
  1. $\widehat{var}(\hat{\varepsilon}_i) = s^2(1 - h_{ii})$
  2. $\hat{\varepsilon}_i^* = \frac{\hat{\varepsilon}_i}{\sqrt{s^2(1-h_{ii})}}$

Para verificar la validez de la regre°.

- Métodos de verifica°.
  1. Graficar residuales $\hat{\varepsilon}_i$ vs $\hat{y}_i$ (no debe haber dependencia)
  1. Graficar $\hat{\varepsilon}_i$ vs z_i (si hay dep. funcional, indica q' hacen falta términos).
  3. QQ plots e histogramas residuales de $\hat{\varepsilon}_i$ (evaluar normalidad)