Clase 15 Diagnostico de regresión

Residuos

$$e_i = Y_i - (\hat{\beta}_0 + \hat{B}_1 X_{i1} + \dots + \hat{B}_{p-1} X_{i(p-1)}) = Y_i - X_i^T \hat{B}$$

$$e = Y - X\hat{B}$$

Propiedades

- $E(e_i) = 0$
- $Var(e_i) = \sigma^2 [1 - X_i^T (X^T X)^{-1} X_i] = \sigma^2 (1 - h_{ii})$
- $Cov(\varepsilon_i, \varepsilon_j) = -\sigma^2 X_i^T (X^T X)^{-1} X_j = -\sigma^2 h_{ij}$
- Residuos NO son una muestra aleatoria de $N(0, \sigma^2)$

  No son independientes

Residuos Studentizados

internamente

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}} \sim N(0,1) \quad \text{para } i = 1, \dots, n$$

externamente

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}^2 (1 - h_{ii})}} \sim t_{n-p-1} \quad i = 1, \dots, n$$

$$\hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n-p-1} (n - p - r_i^2)$$

C15 Linealidad: Análisis gráfico

Test de Hipótesis

Test F de Fisher-Snedecor

- $H_0: \mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} X_{i(p-1)}$
- Bajo H_0, $F = \frac{MCFA}{MCEP} \sim F_{(k-p,\, n-k)}$
- a nivel de significancia α, rechazamos H_0 si $F > F_{((k-p),\, n-k;\, \alpha)}$

Clase 14 Anova / Inferencia

FV	SC	GL	MC	Test F
Regresión	$SCR = \sum_1^n (\hat{Y}_i - \bar{Y})^2$	$p-1$	$MCR = \frac{SCR}{p-1}$	$F = \frac{MCR}{MCE}$
Error	$SCE = \sum_1^n (Y_i - \hat{Y}_i)^2$	$n-p$	$MCE = \frac{SCE}{n-p}$	
Total	$SCT = \sum_1^n (Y_i - \bar{Y})^2$	$n-1$		

Test F de Fisher - Snedecor

$H_0: B_1 = B_2 = \ldots = 0$ $\quad H_1$: algún $B \neq 0$

- Bajo H_0, $F = \frac{MCR}{MCE} \sim F_{(p-1, n-p)}$
- A nivel α, rechazamos H_0 si $F > F(p-1, n-p, \alpha)$

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

el valor p es menor que eso, por lo que rechazamos Ho. Si es signif.

<u>Estimación recta</u>

$Y_i \sim N(\mu_i, \sigma^2)$

$\mu_i = \beta_0 + B_1 X_{i1} + \ldots + B_{p-1} X_{i(p-1)} = X_i^T B \quad i = 1, \ldots, n$

estimar para X_h específico.

$E(Y | X = X_h) = X_h^T B$

$\hat{\mu}_h = X_h^T \hat{B} \sim N[E(\hat{\mu}_h), Var(\hat{\mu}_h)]$

donde $E(\hat{\mu}_h) = X_h^T B$ y

$$Var(\hat{\mu}_h) = Var(X_h^T \hat{B}) = X_h^T Var(\hat{B}) X_h$$

$$= \sigma^2 X_h^T (X^T X)^{-1} X_h$$

$$\Rightarrow \boxed{\mu_h \in \left(\hat{\mu}_h \mp t(n-p; \alpha/2)\, \hat{\sigma} \sqrt{X_h^T (X^T X)^{-1} X_h} \right.}$$

Clase 13 Regresión Lineal Múltiple

Estimación y propiedades

$$\boxed{Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + \ldots + B_{p-1} X_{(p-1)i} + \varepsilon_i}$$

Condiciones Gauss-Markov

1) $E(\varepsilon_i) = 0 \quad i = 1, \ldots, n$
2) $Var(\varepsilon_i) = \sigma^2 \quad i = 1, \ldots, n$ (homocedasticidad)
3) $E(\varepsilon_i \varepsilon_j) = 0 \quad \forall i \neq j$ (observaciones incorrelacionadas)
4) $\varepsilon_i \sim N(0, \sigma^2) \quad i = 1, \ldots, n$ (Normalidad)
5) $X^T X$ no singular, en caso contrario, multicolinealidad

Notación Matricial

$$Y = XB + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad B_{p \times 1} = \begin{pmatrix} B_0 \\ B_1 \\ \vdots \\ B_{p-1} \end{pmatrix} \quad \varepsilon_{n \times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad X_{n \times p} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1(p-1)} \\ 1 & X_{21} & \cdots & X_{2(p-2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{n(p-1)} \end{pmatrix}$$

donde rango $(X_{n \times p}) = p < n$

Estimadores

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\tilde{\sigma}^2 = \frac{RSS(\hat{\beta})}{n} \text{ [Sesgado]} \longrightarrow \hat{\sigma}^2 = \frac{RSS(\hat{\beta})}{n-p} \text{ [Insesgado]}$$

Distribución estimadores

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$$ Si $p=2 \Rightarrow$ simple

$\hat{\sigma}^2$ indep. de $\hat{\beta}$

Efecto interacción

- Los modelos de reg. con interacción (o efectos no aditivos) son casos especiales del modelo de reg. lineal.
- Efecto variables indep. en var resp. no son aditivos.
- En estos casos, efecto de una var. indep. depende de los niveles de otras var. indep.

Regresión polinomial

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{2i}^2 + \varepsilon_i$$

Transformación de variables

Var. indep. No se observa relación lineal
ej. $\log(x)$, x^2, $\sqrt{x}$

Var. Dependiente
- No se observa relación lineal
- Residuos no se distribuyen normalmente
- $\log(Y)$, Y^2, $\sqrt{Y}$
- Box-Cox, Atkinson, Johnson

Box-Cox: Encontrar valor apropiado para obtener normalidad de residuos.

$$Y_i^*(\lambda) = \begin{cases} \dfrac{Y_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(Y_i) & \lambda = 0 \end{cases}$$

1) Coef. correlación basado en Normal
2) Test normalidad Shapiro - Wilk
3) Máxima Verosimilitud (Normal)

Buscamos el λ que maximice Log-Likelihood. (Verosimilitud)

En los test, H_0 es parametro = 0, con $\alpha = 5\%$
si valor p es menor a 5%, => es significativo y rechazamos H_0.

Ej. práctico: buscar λ que optimiza BoxCox.
1. Ajusto modelo y verifico si residuos son normales.