CI for Y_h : $\hat{Y}_h \pm B \cdot sd(pred)$

$B = t(1-\frac{\alpha}{2g}; n-2)$.

g: # of observations.

Scheffe procedure : $\hat{Y}_h \pm S \cdot sd(pred)$

$S = g \cdot F(1-\alpha; g; n-2)$. without derivation.

similar to W-H.

Em for Midterm #1.

★ (Chap 5) Matrix approach to LR.

$\mathbb{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$ $n \times n$ (↓ raw, ↓ col.)

Vectors : $(n \times 1)$ matrices.

$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

$\mathbb{A}^T$ → transpose : swap raws & cols.

Matrix $\underline{C}$, $\mathbb{E}(\underline{C}) = [\mathbb{E}(C_{ij})]$, $\underline{y}$, $\mathbb{E}\underline{y} = \begin{bmatrix} \mathbb{E}y_1 \\ \vdots \\ \mathbb{E}y_n \end{bmatrix}$

') dot product $\underline{y} \cdot \underline{y}$ $(\langle y, y \rangle)$ (scalar)

$\underline{y} \cdot \underline{y} = y_1^2 + y_2^2 + \cdots + y_n^2 = \|\underline{y}\|^2$

$\|\underline{y}\|$: norm / length / magnitude of $\underline{y}$.

$\underline{u}, \underline{v} \in \mathbb{R}^n$, $\underline{u} \cdot \underline{v} = u_1v_1 + u_2v_2 + \cdots + u_nv_n$.

•) matrix vector product.

$$\underset{m\times n}{A}\ \underset{n\times 1}{u} = \begin{bmatrix} \text{row}_1(A)\cdot u \\ \vdots \\ \text{row}_m(A)\cdot u \end{bmatrix}$$

e.g. $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} u_1+2u_2 \\ 3u_1+4u_2 \end{bmatrix}$.

Matrix–matrix product:

$$\underset{m\times n}{A}\cdot\underset{n\times p}{B} = \begin{bmatrix} | & & | \\ A\cdot B_1 & \cdots & A\cdot B_p \\ | & & | \end{bmatrix}$$

B_i = the ith col of B.

Write dot product in matrix form:

$$u\cdot v = u^T v = \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix}\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$$

$$= v^T u = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \Rightarrow \text{scalar.}$$

$$y\cdot y = \|y\|^2 = y^T y = \Sigma_i y_i^2$$

$yy^T \Rightarrow (n\times n)$ matrix. outer product. $\neq$ dot product.

$$\underset{n\times 2}{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \underset{2\times n\ \ n\times 2}{X^T X} = \begin{bmatrix} n & \Sigma_i x_i \\ \Sigma_i x_i & \Sigma_i x_i^2 \end{bmatrix}$$

$(X^TX)^T = X^TX. \rightarrow A^T = A \Rightarrow A$ is symmetric.

Fact: For any A, A^TA is always a symmetric matrix.

I_n ↑ $n \times n$ $= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & 0 & & 1 \end{bmatrix}$

- Inverse of a matrix: A is invertible if $\exists\, A^{-1}$ s.t. $A \cdot A^{-1} = A^{-1} \cdot A = I$.

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad A^{-1} = \frac{1}{\det(A)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

- Properties

$$(AB)^T = B^T A^T$$

$$(A^T)^T = A.$$

$$(A^T A)^T = A^T A.$$

$$(AB)^{-1} = B^{-1} A^{-1}$$

$$(ABC)^{-1} = C^{-1} B^{-1} A^{-1}$$

$$(A^T)^{-1} = (A^{-1})^T.$$

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \det(X^T X) = n \sum x_i^2$$

$$- (\sum x_i)^2 = n \sum (x_i - \bar{x})^2$$

$$= n S_{xx}.$$

- Back to Stats.

$$\mathrm{Var}(y) = \mathbb{E}[(y - \mathbb{E}(y))^2].$$

$$\mathbb{E}y = \begin{bmatrix} \mathbb{E}y_1 \\ \vdots \\ \mathbb{E}y_n \end{bmatrix}$$

$$\mathrm{Var}\, y = \mathbb{E}\big(\underbrace{(y - \mathbb{E}y)}_{n \times 1} \underbrace{(y - \mathbb{E}y)^T}_{1 \times n} \big).$$

Variance-covariance matrix.

$$\text{Var}(\underline{y}) = \begin{bmatrix} \text{Var}(y_1) & \text{Cov}(y_1,y_2) & \cdots & \text{Cov}(y_1,y_n) \\ \text{Cov}(y_2,y_1) & \ddots & & \vdots \\ \vdots & & \ddots & \\ \text{Cov}(y_n,y_1) & \cdots & & \text{Var}(y_n) \end{bmatrix}$$

Covariance matrix.

Recall $W = aY$ $\quad E(W) = aE(Y)$, $\quad \text{Var}(W) = a^2 \text{Var}(Y)$.

$\underline{W} = \underline{A}\underline{Y}$ $\quad E(\underline{W}) = \underline{A}E(\underline{Y}) \neq E(\underline{Y})\underline{A}$

$m \times 1 \quad m \times n \quad n \times 1$

$\text{Var}(\underline{W}) = \underline{A}\,\text{Var}(\underline{Y})\,\underline{A}^T$.

$m \times n \quad n \times n \quad n \times m$

1) SLR model matrix form.

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \ldots, n. \rightarrow n$ eqns.

$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$

$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$

$\vdots$

$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n.$

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \underline{x} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\underline{y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$$

$n \times 1 \quad n \times 2 \quad 2 \times 1 \quad n \times 1$.

$E(\varepsilon_i) = 0.$

$\text{Var}(\varepsilon_i) = \sigma^2$

$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0.$

$\rightarrow E\underline{\varepsilon} = \vec{0}.$

$$\text{Var}\,\underline{\varepsilon} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma^2 \end{bmatrix} = \sigma^2 I_n.$$

$\underline{y} \sim N(\mu, \Sigma)$

$p \times 1$; μ: mean vector ; Σ: covariance matrix.

$\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 \underline{I_n})$.

pdf: $f(y) = \frac{1}{(2\pi)^{\frac{p}{2}} (\det \Sigma)^{-\frac{1}{2}}} \exp\left[-\frac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right]$

☆ Section #7 2023春AP统计(B)

- Gaussian-Markov Theorem.

  Normality

  in-covariance (independent)

  Mean-zero.

- $e_i = Y_i - \hat{Y}$

  $e_i^* = \frac{Y_i - \hat{Y}}{\sqrt{MSE}}$ semi-studentized residual.

- Possible plots:

  e_i vs. X_i

  e_i^2 or $|e_i|$ vs. X_i

  e_i vs. $\hat{Y}_i$ Covariance should be zero.

  e_i vs. time. test dependent error term

  e_i vs. other predictors.

  Normal plot of e_i's (Q-Q plot)

- Issues with the model.

  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

  1. Non-linear function

     e.g. $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ → γX_i^2

  Heteroskedasticity test → 2. Non-constant variance

  runs test → 3. Not independent error term

  4. outliers.

  5. Error terms are not normally distributed.

  6. Missing predictor.

QQ plot.

Shapiro-Wilk normality test

☆ Lecture 10 2023年10月2日.

Matrix approach to LR
- SLR and inference
- multiple. LR

☆☆ SLR w/ normal error in matrix form

$\underset{n\times 1}{y} = \underset{n\times 2}{X}\cdot\underset{2\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$, $\varepsilon \sim N(0, \sigma^2 I_n)$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

☆☆ Least Squares Estimation of β.

$Q(\beta) = Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 = \|y - X\beta\|^2 = (y - X\beta)^T(y - X\beta)$

$$b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \hat\beta_{0LS} \\ \hat\beta_{1LS} \end{bmatrix} = \underset{\beta_0, \beta_1}{\operatorname{argmin}}\, Q(\beta_0, \beta_1) = \underset{\beta}{\operatorname{argmin}}\, Q(\beta).$$

fitted values: $\hat y = X\cdot b$ Soln.: $b = (X^TX)^{-1}X^Ty$

$\hat y = X(X^TX)^{-1}X^Ty$ Hat matrix, $H = X(X^TX)^{-1}X^T$

note: all params, fitted values, residuals, are linear combi of Y_i, $i = 1, 2, \dots, n$.

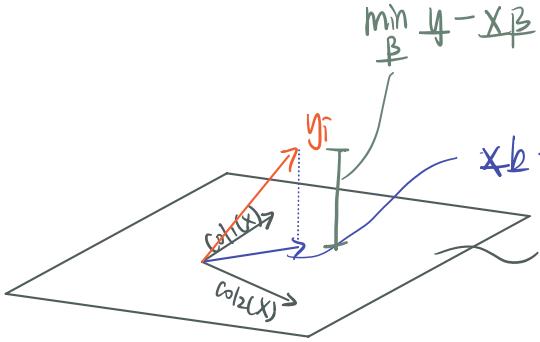
$b = A_1 y$
$\hat y = A_2 y$, $A_2 = H$) like before. $SSTO = SSE + SSR$.
$e = A_3 y$

☆☆ Normal equations:

$(X^TX)b = X^Ty$ $\quad AV = v_1A_1 + v_2A_2 + \cdots + v_nA_n$

$\underset{2\times 2}{X^TX}\ \underset{2\times 1}{b} = \underset{2\times n}{X^T}\ \underset{n\times 1}{y}$ $\quad \underset{n\times n}{A}\ \underset{n\times 1}{V}$; A_i = col. of matrix A.

$Xb = b_0\,\mathrm{col}_1(X) + b_1\,\mathrm{col}_2(X)$



projection of y onto col(X).

☆☆ Properties:

$H^T = H$ (symmetric)
$H^2 = HH = H$ (idempotent) pf: $X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T = H$.
$E\hat y = E[Xb + \varepsilon] = X\beta + E[\varepsilon] = X\beta$ ($E[\varepsilon] \to 0$)

$\mathrm{Var}\,\hat y = \mathrm{var}(Xb + \varepsilon) = 0 + \mathrm{Var}(\varepsilon) = \sigma^2 I_n$
distribution: $\hat y = \varepsilon + Xb \sim N(X\beta, \sigma^2 I_n)$

Vector of residuals: $e = y - \hat{y} = y - Hy = (I-H)y$

a) $E(e) = E[(I-H)y] = (I-H)E[y] = (I-H)(X\beta) = X\beta - H(X\beta) = 0.$

b) $\text{Var}(e) = \text{var}((I-H)y)$
$= (I-H)\,\text{var}(y)\,(I-H)^T$
$= \sigma^2(I-H)(I-H)^T$ ∵ $(I-H)$ symmetric.
$= \sigma^2(I-H)(I-H)$
$= \sigma^2(I - HI - IH + HH)$ ∵ $IH = H$, $HH = H$.
$= \sigma^2(I-H)$

Var-cov matrix of residual vector e.

$\text{Var}(e_1) = [\sigma^2(I-H)]_{11} = \sigma^2(1-h_{11})$

Recall $\text{Var}(e_i) = \sigma^2(1-h_{ii}) \neq \text{Var}\,\varepsilon_i = \sigma^2$

Inference in regression analysis.

$\text{Var}(b) = \text{var}((X^TX)^{-1}X^Ty)$
$= (X^TX)^{-1}X^T\,\text{var}(y)\,((X^TX)^{-1}X^T)^T$ ∵ $\text{Var}(y) = \sigma^2 I$
$= \sigma^2(X^TX)^{-1}X^T(X^T)^T((X^TX)^{-1})^T$
$= \sigma^2(X^TX)^{-1}(X^TX)((X^TX)^{-1})^T$
$= \sigma^2((X^TX)^{-1})^T$
$= \sigma^2((X^TX)^T)^{-1}$
$= \sigma^2(X^TX)^{-1}$
$= \text{Var}(b)$

$SSR = b_1^2 S_{XX} = \frac{S_{XY}^2}{S_{XX}}$ $\quad SSE = S_{YY} - \frac{1}{S_{XX}}S_{XY}^2$

$SSTO = S_{YY}$

a) $\text{Var}(b_0) = [\sigma^2(X^TX)^{-1}]_{11}$ $\quad s^2(b_0) = MSE((X^TX)^{-1})_{11}$ ①

b) $\text{Var}(b_1) = [\sigma^2(X^TX)^{-1}]_{22}$ $\quad s^2(b_1) = MSE((X^TX)^{-1})_{22}$ ②

c) $\text{Cov}(b_0, b_1) = [\sigma^2(X^TX)^{-1}]_{12}$

use in ① & ②

CI for β_0, β_1. Before: $s^2(b_0) = MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)$

$s^2(b_1) = MSE\left(\frac{1}{S_{XX}}\right)$

MSE using Linear Algebra:

$MSE = \frac{1}{n-2}SSE = \frac{1}{n-2}\|e\|^2 = \frac{1}{n-2}\|y - Hy\|^2 = \frac{1}{n-2}((I-H)y)^T((I-H)y)$
$= \frac{1}{n-2}(y^Ty - y^THy)$

Inference of mean response at X_h.

$\hat{y}_h = b_0 + b_1X_h = X_h^Tb$, $X_h = \begin{bmatrix}1\\X_h\end{bmatrix}$

$\text{Var}(\hat{y}_h) = \text{var}(X_h^Tb) = X_h^T\,\text{var}(b)\,X_h = \sigma^2 X_h(X^TX)^{-1}X_h = \sigma^2\left[\frac{1}{n} + \frac{1}{S_{XX}}(X_h - \bar{X})^2\right]$

$s^2(\hat{y}_h) = MSE\, X_h^T(X^TX)^{-1}X_h$

use this for hypothesis Testing / CI for mean response.

Prediction of resp. variable at X_h.

$s^2(\text{pred}) = MSE[1 + X_h^T(X^TX)^{-1}X_h]$

✡✡ Multiple Linear Regression (aka. honey prediction) Chapter 6.

we have 2 predictors, X_1, X_2.

obsn's: n triplets: (x_{11}, x_{12}, y_1)

(x_{n1}, x_{n2}, y_n)

model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d.

fitted plane: $\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2$

β_0: mean response when predictors are all 0.

β_1: increase in mean response when x_1 increases by 1 and x_2 is constant

β_2: the other way around (vice-versa).

What if p-predictors?

In general: $y_i = \underbrace{\beta_0 + \beta_2 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}}_{\text{hyperplane in } \mathbb{R}^p} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$

$\underset{n\times 1}{y} = \underset{n\times p}{X} \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$, $\varepsilon \sim N(0, \sigma^2 I_n)$

✡✡ General Linear Model (GLM)

Not to be confused with generalized linear models (GLiM)

$y_i = \underbrace{c_{i0}\beta_0 + c_{i1}\beta_1 + \cdots + c_{i,p-1}\beta_{p-1}}_{\text{functions of } x_i \text{ / constants.}} + \varepsilon_i$

Note: still LR model as still linear on params $\beta_0, \ldots, \beta_{p-1}$!

eg. polynomial regression.

$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \underbrace{\beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2}_{\text{2nd order term}} + \varepsilon_i$

✡✡ Qualitative predictions (categorical var.)

gender: M or F. $X_1 \begin{cases} 1, & \text{if F} \\ 0, & \text{M} \end{cases}$

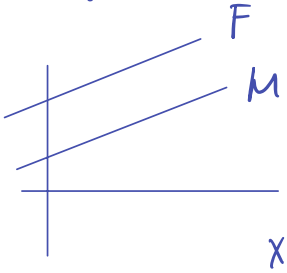
age: X_2

C categories → (C-1) variables

↳ "dummy var."

$\mathbb{E}y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ $\begin{cases} \text{F: } (\beta_0 + \beta_1) + \beta_2 X_2 \\ \text{M: } \beta_0 + \beta_2 X_2 \end{cases}$

different intercepts.



	X_3	X_4
disabled	1	0
Partially disabled	0	1
not disabled.	0	0

X_3	X_4	X_5
1	0	0
0	1	0
0	0	1

bad. X^TX stops being invertible.

✩ Lecture 11 2023年10月9日(一).

✩ Recall $\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$ ($n\times1$, $n\times p$, $p\times1$, $n\times1$) $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 \underline{I}) \sim N_n(\underline{0}_n, \sigma^2 \underline{I}_n)$

✩ LS: $\underline{b} = \arg\min_{\underline{\beta}} Q(\underline{\beta}) = \arg\min_{\underline{\beta}} \| \underline{Y} - \underline{X}\underline{\beta} \|^2$

↳ normal equations: $(\underline{X}^T\underline{X})\,\underline{\beta} = \underline{X}^T\underline{Y}$ ($p\times p$, ↳ $p\times1$, $p\times n$, $n\times1$) p equations. (full Rank.)

$\Rightarrow \underline{b} = (\underline{X}^T\underline{X})^{-1}\underline{X}^T\underline{Y}$

In order to calculate inverse, X must be full rank.

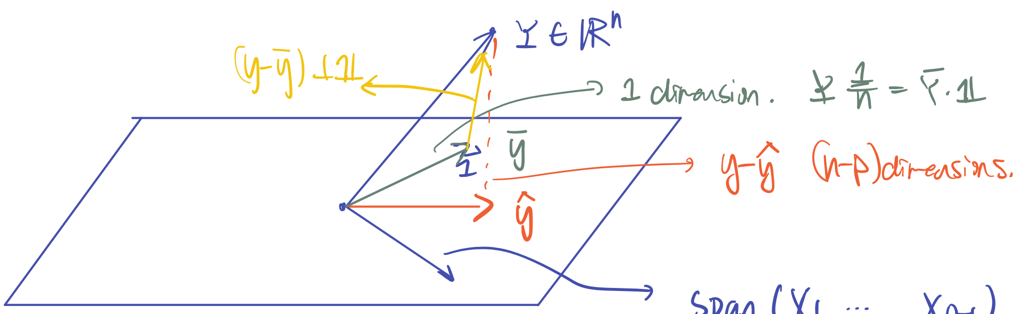
✩ MLE: $L(\underline{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\{-\frac{1}{2\sigma^2}(\sum_{i=1}^n y_i - \beta_0 - \beta_1 x_{i1} \cdots - \beta_{p-1} x_{i,p-1})^2\}$

maximizing wrt $\underline{\beta}, \sigma^2$, $\hat{\underline{\beta}}^{MLE} = \hat{\underline{\beta}}^{LS}$ ← minimum variance, unbiased, consistent. $\hat{\underline{\beta}}^{MLE} \xrightarrow{n\to\infty} \underline{\beta}$

results as before for p predictors.

Fitted values: $\hat{\underline{y}} = \underline{H}\underline{Y} = \underline{P}\underline{Y} = \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T\underline{Y}$ ($\underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T$: H.)

Residuals: $\underline{e} = (\underline{I} - \underline{H})\underline{Y}$



✩ ANOVA Table:

Source	SS	df = dim	MS.
error	SSE	(n-p)	$MSE = \frac{SSE}{n-p} = \|y-\hat{y}\|^2/(n-p)$
regression	SSR	(p-1)	$MSR = \frac{SSR}{p-1} = \|\hat{y}-\bar{y}\|^2/(p-1)$
total	SSTO	(n-1)	

(SSE + SSR = SSTO)

$\min_c \sum_i (y_i - c)^2$, $\underline{c} = c\cdot\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix} \Rightarrow \min_c \|\underline{y} - c\cdot\underline{\mathbb{1}}\|^2$. $\underline{c} = c\cdot\underline{\mathbb{1}}$, $c = \bar{y}$

Like before, $\mathbb{E}(MSE) = \sigma^2$ (unbiased)

$\mathbb{E}(MSR) = \sigma^2 + \underline{\beta}^T \underline{A}\,\underline{\beta}$

↳ symmetric and positive definite

all e-values are strictly positive.

If $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$, then $MSE \approx MSR$.
If at least one of $\beta_k \neq 0$, then $MSE < MSR$.

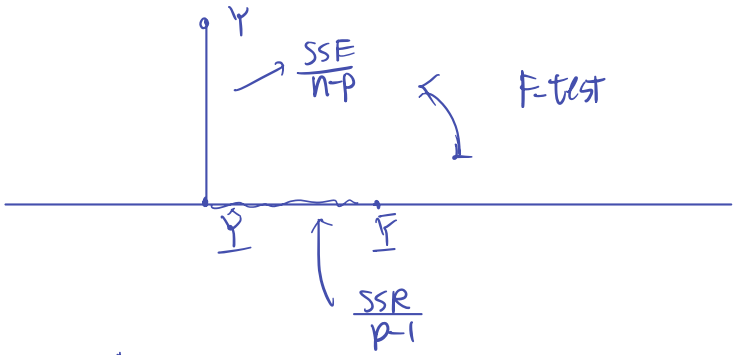
F-test (test of overall fit / model utility test).

$H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$,
H_a: at least one of them is non zero.

$$F^* = \frac{MSR}{MSE} = \frac{SSR/(p-1)}{SSE/(n-p)}$$

If $F^* \leq F(1-\alpha, p-1, n-p)$, conclude H_0.
If $F^* > F(1-\alpha, p-1, n-p)$, conclude H_a.

F test: is there a useful relationship bow Y and $X_1, \cdots, X_{p-1}$?



Now, Ftest is no longer equivalent to t-test (on β_1).

Coefficient of multiple determination (R^2)

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

relative reduction in the variation of y due to all predictors $X_1, \cdots, X_{p-1}$.

$R^2 = 0 \Rightarrow b_1 = b_2 = \cdots = b_{p-1} = 0$.
$y_i = b_0$, $SSR = \hat{y} - \bar{y}$ No linear relationship

$R^2 = 1 \Rightarrow SSE = 0$, $y - \hat{y} = 0$. Perfect fit.
regression pass through all datapoints.

As we add predictors, R^2 increasing "always".
$\underset{n\times 1}{\arg\min} \|Y - X\beta\|^2$ bic you search the minimum at a larger & larger dimension of space.

Adjusted R^2, penalize # of predictors.

$$R^2_{adjusted} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)} = 1 - \frac{MSE}{MSTO} = 1 - \frac{n-1}{n-p}\frac{SSE}{SSTO}$$

as $p \uparrow$, $\frac{n-1}{n-p} \uparrow$, $R^2_{adj} \downarrow$, (if SSE stays the same)

Note: R^2 = coeff. mult det. = coeff. simple det. when $Y_i \approx \hat{Y}_i$

Inference on β_k's

$E(b) = \beta$; $var(b) = \sigma^2(X^TX)^{-1}$
$S^2(b) = MSE(X^TX)^{-1}$
b $(n-p)$ dim.

$\hat{S}(b_k) = [\hat{S}^2(\underline{b})]_{k+1,k+1}$

$CI_{1-\alpha}$: $b_k \pm S(b_k)\, t(1-\frac{\alpha}{2}; n-p)$

H_0 : $\beta_k = 0$

H_a : $\beta_k \neq 0$

T.S : $t^* = \frac{b_k}{S(b_k)} \sim t_{n-p}$ If $|t^*| \leq t(1-\frac{\alpha}{2}; n-p)$, conclude H_0.

☆☆ If q coeffs estimated:

Bertfroni procedures

joint intervals $b_k \pm S(b_k)\, t(1-\frac{\alpha}{2q}; n-p)$

(recall joint interval for β_0, β_1)

Lecture 12 2023년 10월 11일 (수)

Inferences in MLR.

→ mean response var. $\underline{X_h} = \begin{bmatrix} 1 \\ X_{h,1} \\ \vdots \\ X_{h,p-1} \end{bmatrix} \in \mathbb{R}^p$ $\hat{Y}_h = \underline{X_h}^T \underline{b}$ = fitted value (estimated mean response).

$$s^2(\hat{Y}_h) = MSE(\underline{X_h}^T (\underline{X}^T\underline{X})^{-1} \underline{X_h}) = \underline{X_h}^T s^2(b) \underline{X_h}$$

$1\times p \quad p\times p \quad p\times 1$ → 1×1 (scalar)

$$CI_{1-\alpha} = \hat{Y}_h \pm t(1-\frac{\alpha}{2}, n-p)\, s(\hat{Y}_h)$$

CI for a single mean.

→ Confidence region for regression surface

$$\hat{Y}_h \pm W \cdot s(\hat{Y}_h) \qquad W = \sqrt{p F(1-\alpha; p, n-p)}$$

"Combined" CI for all possible means. (generalization of Working-Hotelling procedure).

→ simultaneous CI for g means

$$\hat{Y}_h \pm t(1-\frac{\alpha}{2g}, n-p) \cdot s(\hat{Y}_h)$$ (Bonferroni procedure).

→ prediction of a new observations

(PI) $\hat{Y}_h \pm t(1-\frac{\alpha}{2}, n-p)\, s(pred)$

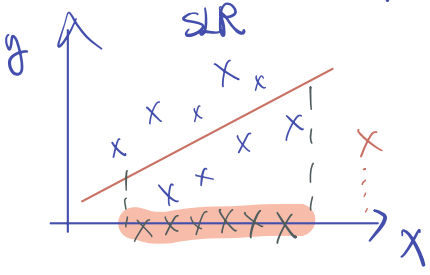
$s^2(pred) = MSE(1 + \underline{X_h}^T (\underline{X}^T\underline{X})^{-1} \underline{X_h})$

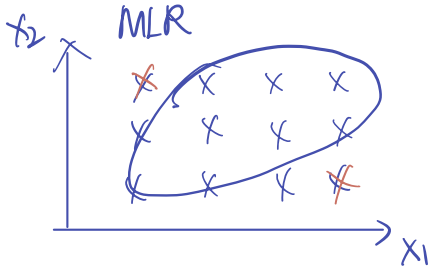
→ simultaneous prediction intervals:

Same as before in SLR: – Bonferroni
– Scheffe.

Always pick the Narrow one.

Issue: Hidden Extrapolations.





Cannot really detect the outlier by this graph.

2 predictors.

Diagnostics in MLR.

① Scatter plots for all predictors against all predictors.
Pairs (dataset).

② Correlation matrix. → niche: 3D Scatterplot (Y_i vs. $X_{i,k}$ & X_{ij})
Cor (dataset)

⑤ Fit the model. using lm(formula , df)
See the p-values.

✡✡ Diagnostics from SLR still apply
- e_i vs. $\hat{Y}_i$ → nonlinearity.
- Hist. & QQ plot for e_i's → Normality
- e_i vs. $X_{i,k}$ for all ks.
- e_i vs. omitted predictors. → do we need more predictors?
- $|e_i|$ or e_i^2 vs. $\hat{y}_i$ → outliers.
- $|e_i|$ or e_i^2 vs. $X_{i,k}$ → Nonconstancy of variance

Interaction terms: $X_{i,k_i} X_{i,j}$, plot e_i vs. interaction terms to see if needed.

✡✡ Remedial measures.
Shapiro-Wilks test for normality of error.
Brown-Forsythe test for nonconstancy of variance.

✡✡ If variance is NOT constance:
Transformation on y. Sometimes transform X also to fix the linearity.

✡✡ If nonlinear,
Transformation on X, or add some predictor f(x).

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_0$$
$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_0.$$

✡✡ Extra sum of squares (Ch. 7)
Model #1: $y \sim X_1$
Model #2: $y \sim X_1, X_2$
Q: worth to add X_2 to the model?
ANOVA: $SSTO_{\text{model}} = SSE(X_1) + SSR(X_1)$ ∈ M1
$= SSE(X_1 + X_2) + SSR(X_1 + X_2)$ ∈ M2

$$SSE(X_1) \geq SSE(X_1 + X_2)$$
$$SSR(X_1) \leq SSR(X_1 + X_2)$$

Extra sum of squares: $SSR(X_2 | X_1) = SSE(X_1) - SSE(X_1, X_2)$ (1)
$= SSR(X_1, X_2) - SSR(X_1)$ (2)

(1) How much the error decrease when we add X_2 to model $Y \sim X$?
(2) How much the explained var. in Y increased when we added X_2 to $Y \sim X$?
↳ need to test if $SSR(X_2|X_1)$ is large enough.

→ General Linear Test (Lecture 5).

(R) Reduced: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$. df: $n-2$

(F) Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. df: $n-3$.

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \Big/ \frac{SSE(F)}{df_F} = \frac{SSE(X_1) - SSE(X_1, X_2)}{(n-2)-(n-3)} \Big/ \frac{SSE(X_1, X_2)}{n-3}$$

$$= \frac{SSR(X_2 \mid X_1)}{1} \Big/ \frac{SSE(X_1, X_2)}{n-3} = MSR(X_2 \mid X_1) \Big/ MSE(X_1, X_2)$$

$$\sim F(1, n-3)$$

$$\Updownarrow$$

t test $(t^*)^2 = \dfrac{(b_2)^2}{s^2(b_2)}$

★★ Marginal Test.

Is X_2 worth adding to the model which already has X_1?

Relationship to model utility test (F test for overall fit)

$H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$

H_a: at least one $\beta_k \neq 0$.

R: $Y \sim 1$ df: $n-1$

F: $Y \sim 1, X_1, X_2, \cdots, X_{p-1}$ df: $n-p$

$$F^* = \frac{SSE(1) - SSE(1, X_1, \cdots, X_{p-1})}{(n-1)-(n-p)} \Big/ \frac{SSE(1, \cdots, X_{p-1})}{(n-p)}$$

$$= \frac{SSR(1, X_1, \cdots, X_{p-1})}{(p-1)} \Big/ \frac{SSE(1, X_1, \cdots, X_{p-1})}{(n-p)} = \frac{MSR}{MSE}$$

★★ General linear test

In general, we can test any subset of variables:

$H_0: \beta_3 = \beta_5 = 0$.

H_a: at least one is not zero.

↳ F test only

★★ Coefficient of partial determination.

$$R^2_{y2|1} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)} = \frac{SSR(X_2 \mid X_1)}{SSE(X_1)}$$

→ measures relative reduction in variation in y when X_2 is included in model $Y \sim X_1$.

$$SSTO = SSR(X_1) + SSR(X_2 \mid X_1) + SSE(X_1, X_2).$$

Lecture 13, Oct. 16th ASDA 413/613

Today: - general Linear test : alternative form
- coefficient of partial determ'n
  - added variable plot / partial regression plot
- standardized multiple regression model
- corr'n transform
- start multicollinearity

No section on Friday: Fall break!

General linear test (alternative form) (* not in KNNL)

(I) General linear test

various F-tests:

(A) Model utility test / overall fit

$$\begin{cases} H_0: \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0 \\ H_a: \text{at least one } \beta_k \neq 0 \end{cases}$$

$$F^* = \frac{MSR}{MSE} = \frac{\frac{SSR(X_1, \dots, X_{p-1})}{p-1}}{\frac{SSE(X_1, \dots, X_{p-1})}{n-p}} \sim F(p-1, n-p) \text{ under } H_0$$

(B) Individual β_k's : $H_0 : \beta_k = 0$ vs. $H_a : \beta_k \neq 0$

$$F^* = \frac{b_k^2}{s^2(b_k)} \sim F(1, n-p) \text{ under } H_0$$

(C) Test various subsets of β_k's, e.g.:

$H_0 : \beta_3 = \beta_5 = 0$ vs. $H_a : \beta_3 \neq 0, \beta_5 \neq 0$

We use G.L.T. : compare reduced models vs. full models

=> We can (and will) generalize this

$H_0 : \underline{C}\,\underline{\beta} = \underline{\gamma}$ vs. $H_a : \underline{C}\,\underline{\beta} \neq \underline{\gamma}$

$\underline{C}$: $m \times p$, $\underline{\beta}$: $p \times 1$, $\underline{\gamma}$: $m \times 1$

rank($\underline{C}$) : q (*)

(A) test of overall fit : $H_0 : \beta_1 = \dots = \beta_{p-1} = 0$

$$\underline{C} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \dots & \dots & 0 \; 1 \end{bmatrix}, \quad \underline{\gamma} = \underline{0} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \Big\} p-1$$

$(p-1) \times p$

$$\underline{C}\cdot\underline{\beta} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix} = \underline{\gamma} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

$\text{rank}(\underline{C}) = p-1$

(B) (individual) $\quad H_0 : \beta_k = 0$

$$\underline{C} = [0\ 0 \ldots 1\ 0 \ldots 0] \Rightarrow \underline{C}\cdot\underline{\beta} = \beta_k$$

(1 at the $k+1$ position)

$$\underline{\gamma} = 0$$

$$\underset{1\times p}{\underline{C}}\ \underset{p\times 1}{\underline{\beta}} = \underline{\gamma} = \underset{1\times 1}{\gamma} \qquad \text{rank}(C) = 1$$

Other interesting tests! E.g.: $H_0: \beta_1 = \beta_2 \iff \beta_1 - \beta_2 = 0$

$$\underline{C} = [0\ 1\ -1\ 0 \ldots 0] \qquad \underline{C}\cdot\underline{\beta} = \beta_1 - \beta_2 ; \ \underline{\gamma} = 0$$

Test statistic ?

$$\text{If } x \sim N(\mu, \sigma^2) \Rightarrow \frac{(x-\mu)^2}{\sigma^2} \sim \chi^2_1$$

Fact: $\underline{x} \sim N_q(\underline{\mu}, \underline{\Sigma}) \qquad (\underline{x}-\underline{\mu})^T \underline{\Sigma}^{-1} (\underline{x}-\underline{\mu}) \sim \chi^2_q$

So: $\underline{b} \sim N(\underline{\beta}, \sigma^2 (X^T X)^{-1}) \Rightarrow$

$$\underline{C}\cdot\underline{b} \sim N(\underline{C}\cdot\underline{\beta}, \underbrace{\sigma^2 \underline{C}(\underline{X}^T\underline{X})^{-1}\underline{C}^T}_{\text{var}(\underline{C}\underline{b})})$$

$\text{rank}(\underline{C}) = q$

$$(\underline{C}\underline{b} - \underline{C}\underline{\beta})^T (\text{var}(\underline{C}\underline{b}))^{-1} (\underline{C}\underline{b} - \underline{C}\underline{\beta}) \sim \chi^2_q$$

σ^2 has to be estimated as usual!

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-p}$$

$$\frac{(\underline{C}\cdot\underline{b} - \underline{\beta})^T [\text{var}(\underline{C}\underline{b})]^{-1} (\underline{C}\underline{b} - \underline{\beta})}{q} \Big/ \frac{SSE/\sigma^2}{n-p}$$

$$= \frac{n-p}{q} \frac{[(\underline{C}\underline{b} - \underline{\beta})^T (\underline{C}(\underline{X}^T\underline{X})^{-1}\underline{C}^T)^{-1} (\underline{C}\underline{b} - \underline{\beta})]}{SSE}$$

$$\sim F_{q, n-p} \qquad \frac{\chi^2_q}{q} \Big/ \frac{\chi^2_{n-p}}{n-p}$$

If $C = [0\ 1\ 0 \dots 0]$ => the above is $\frac{\hat{\beta}_1^2}{s^2(\beta_1)}$, i.e

$H_0 : \beta_1 = 0$

geometrically:

$$= \frac{\|\hat{Y}_F - \hat{Y}_R\|^2}{\|Y - \hat{Y}_F\|} \frac{n-p}{q} = \frac{\|Y - \hat{Y}_R\|^2 - \|Y - \hat{Y}_F\|^2}{\|Y - \hat{Y}_F\|^2} \cdot \frac{n-p}{q}$$

SSE(R) SSE(F) SSE(F)

where $\hat{Y}_R = X \cdot \beta_R$ where $\beta_R = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2$

s.t. $C\beta = \gamma$

very versatile F-test!

We can conduct tests: $H_0 : \beta_3 = \beta_5$

$H_0 : \beta_3 = \beta_5 = 2$

$H_0 : \beta_3 = 5,\ \beta_5 = 7$

$\vdots$

$H_0 : 2\beta_3 + 9\beta_5 = 12$

e.g.: $H_0: \beta_1 = \beta_2 = \beta_3$

↳ 2 eqns: $\beta_1 = \beta_2 \Rightarrow \beta_1 - \beta_2 = 0$

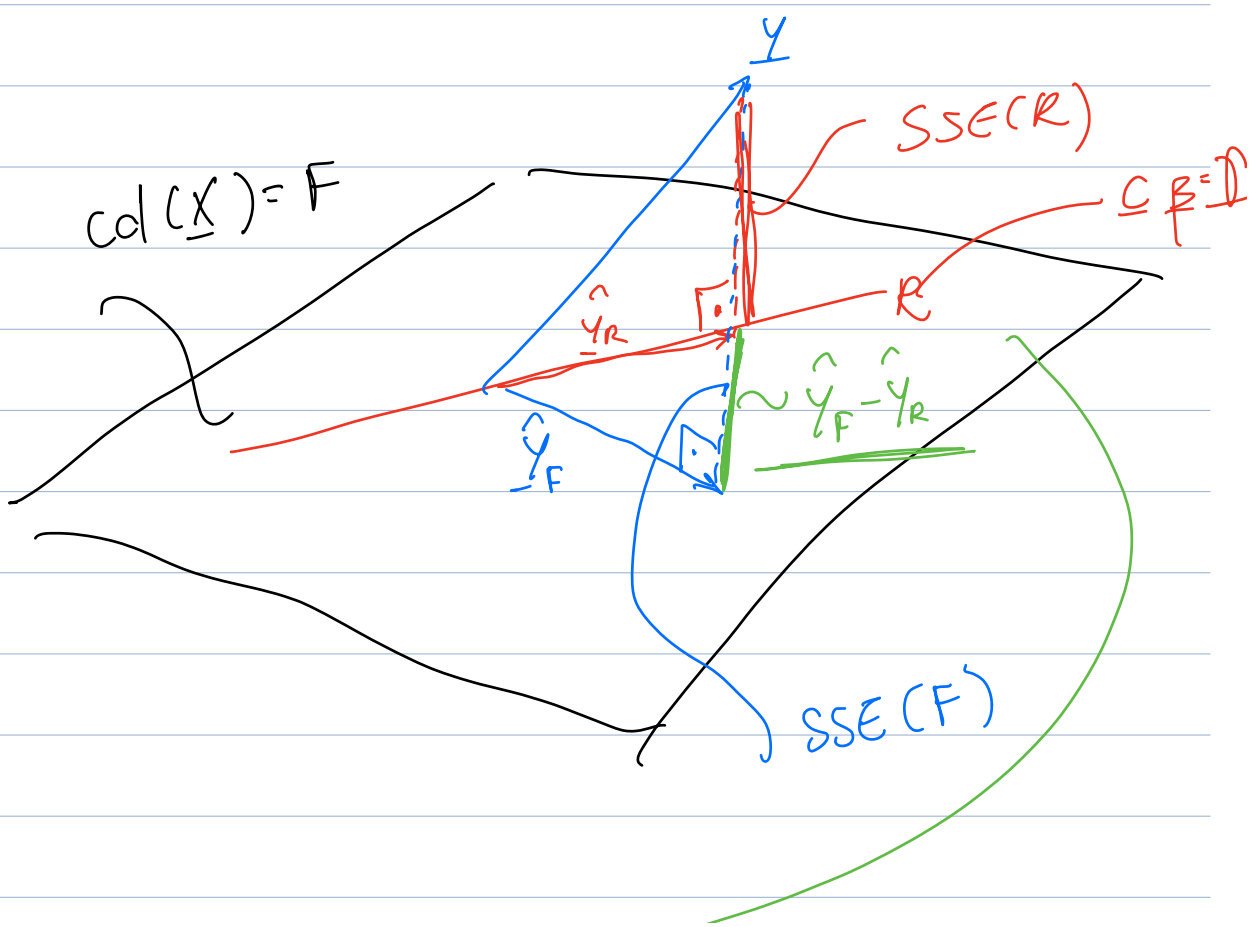
$\beta_1 = \beta_3 \Rightarrow \beta_1 - \beta_3 = 0$ ←

$\beta_2 = \beta_3 \Rightarrow \beta_2 - \beta_3 = 0$

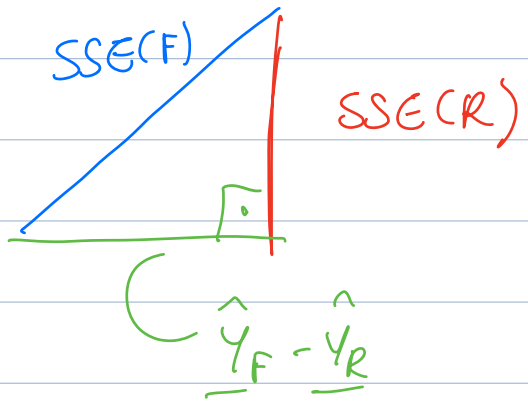
$$\underline{C} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$\underline{C}$ is not unique!

but outcome of the test will be the same!



Side view



- Coefficient of partial determination (Ch. 7.4)

→ coeff. simple det.: R^2 = SCR

→ coeff. multiple det.: R^2 = MLR

→ coeff. of partial det.: $R^2_{Y_{2|1}}$

$$R^2_{Y_{2|1}} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)}$$

= proportionate reduction in variation of Y when X_1 is included in model and we add X_2

meaning:

① $Y \sim X_1 \Rightarrow e_i(Y_i|X_1) = Y_i - \hat{Y}_i(X_1)$

② $X_2 \sim X_1 \Rightarrow e_i(X_2|X_1) = X_{i2} - \hat{X}_{i2}(X_1)$

Regress ① $e_i(Y|X_1) \sim$ ② $e_i(X_2|X_1) \Rightarrow$

R^2 = coeff. of simple defn:

=

coeff. of partial defn $R^2_{Y2|1}$

↳ measures the strength of linear relationship btw Y & X_2 when X_1 taken into account

↳ Interesting fact: $\hat{Y}_i = b_0 + b_1 X_1 + b_2 X_2$ — b_2 are the same!!

$\hat{e}(Y|X_1) = \hat{b}_0 + b_2 e_i(X_2|X_1)$

↳ check R-script!

▶ Additional diagnostic tool

/

↳ added variable plot / partial reg. plot

plot of $e_i(Y|X_1)$ vs $e_i(X_2|X_1)$

↳ check constancy of variance, nonlinear relationship

▶ Standardized multiple linear regression model
- to make β_k's comparable
- for numerical stability in the matrix $(X^TX)^{-1}$

standardization: $\frac{Y_i - \bar{Y}}{s_Y}$, $\frac{X_{ik} - \bar{X}_k}{s_k}$

↳ sample sd(y) ↳ sample sd(X_k)

▶ Correlation transform

↳ $Y_i^* = \frac{1}{\sqrt{n-1}} \frac{Y_i - \bar{Y}}{s_Y}$ $\quad X_{ik}^* = \frac{1}{\sqrt{n-1}} \frac{X_{ik} - \bar{X}_k}{s_k}$

In these standardized models, there is no intercept!

$$Y_i^* = \beta_1^* X_{1i}^* + \beta_2^* X_{2i}^* + \dots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^*$$

all sample means are 0! so model has to pass through origin → no β_0^* (intercept)

$$\underline{X} = \begin{bmatrix} X_{11}^* & \dots & X_{1,p-1}^* \\ \vdots & & \vdots \\ X_{n-1}^* & & X_{n,p-1}^* \end{bmatrix}$$

Fact: $\underline{X}^T \underline{X} = r_{XX}$

correlation matrix btw X_i & X_j!

$$r_{ij} = \frac{1}{n-1} \sum_k \frac{(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)}{S_i S_j}$$

$|r_{ij}| \leq 1$

Similarly, $\underline{X}^T \underline{Y} = \underline{r}_{YX} = \begin{bmatrix} r_{Y,1} \\ r_{Y,2} \\ \vdots \\ r_{Y,p-1} \end{bmatrix}$

$\rightsquigarrow r_{Y,j}$: sample corr'n btw Y & X_j

$$\underline{b}^* = \underline{r}_{xx}^{-1} \cdot \underline{r}_{xy} = \begin{bmatrix} b_1^* \\ \vdots \\ b_{p-1}^* \end{bmatrix}$$

can recover b_k's (unstandardized) as

$$b_k = \frac{s_y}{s_k} \cdot b_k^*$$

$$b_0 = \bar{y} - b_1^* \bar{x}_1 - b_2^* \bar{x}_2 - \dots - b_{p-1}^* \bar{x}_{p-1}$$

Multicollinearity (Ch 7.6)

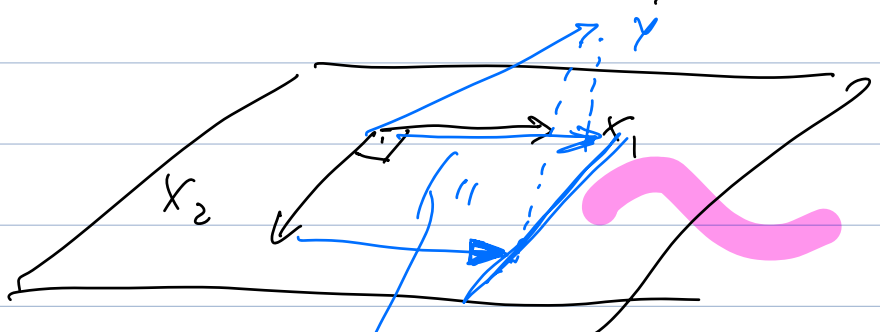
wrong def'n: predictors are correlated

||

multicollinearity / intercorrelation

a) when $corr(X_1, X_2) = 0$,

$$SSR(X_2 | X_1) = SSR(X_2)$$
$$SSR(X_1 | X_2) = SSR(X_1)$$



parallel

SSR is the same

If we fit 3 models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$Y = \tilde{\beta}_0 + \beta_1 X_1$$

$$Y = \hat{\beta}_0 + \beta_2 X_2$$

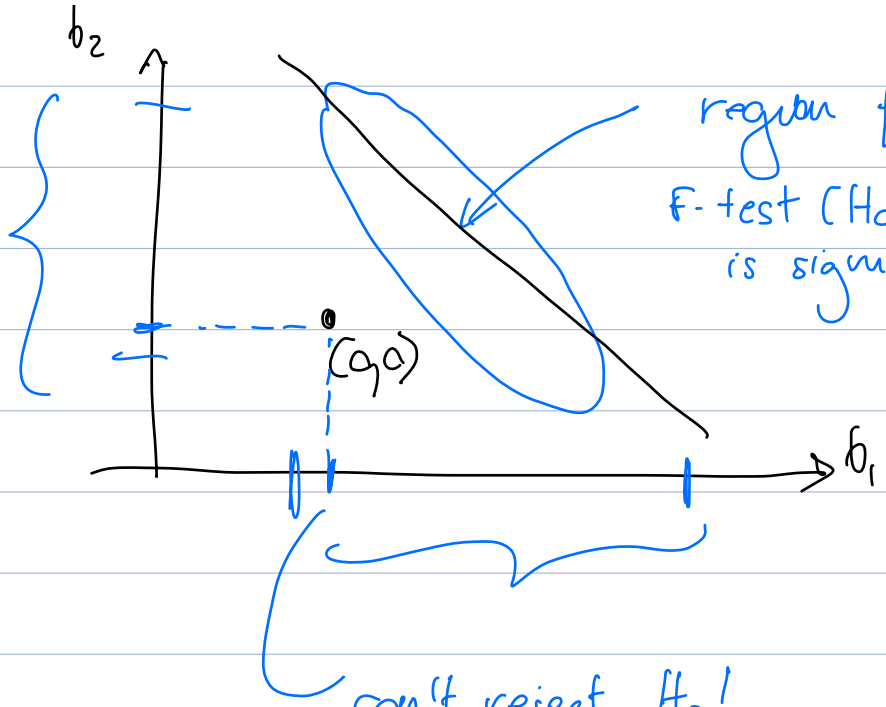
- when $corr(X_1, X_2) = 1$, eg. $X_1 = X_2$

$$b_0 = 2, \quad b_1 = 3, \quad b_2 = 5$$

$$\hat{Y} = 2 + 3X_1 + 5X_2$$
$$= 2 + 4X_1 + 4X_2$$
$$= 2 + (-10)X_1 + 8X_2$$

} $SD(b_k)$ are very large

c) $corr(b_2, b_1)$ is high!



region for F-test

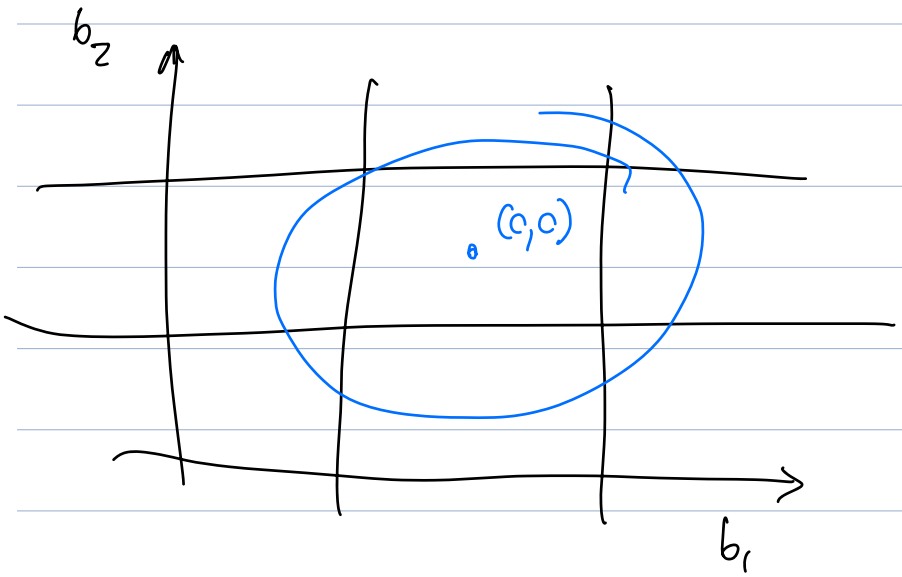
F-test (H_a: atleast one $\beta \neq 0$) is significant

b_2, b_1 are highly (negatively) correlated

can't reject H_a!

(H_a : $\beta_1 = 0$)

a) $corr(b_2, b_1) \approx 0$



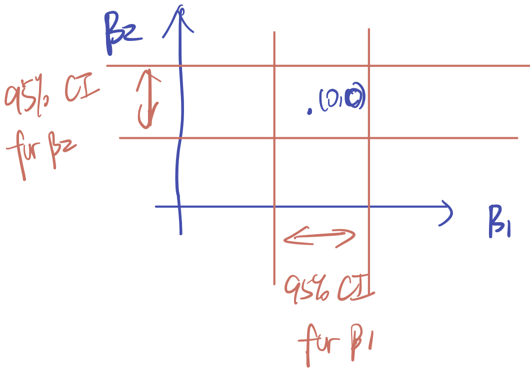
no ambiguity

Lecture 14 2023年10月18日(三). Chap 10.

Collinearity significant vs. non-significant.

β_k is significant $\Leftrightarrow$ H_a nonzero value of β_k (H.T.)
$\Leftrightarrow$ p value $< \alpha$.
$\Leftrightarrow$ $0 \notin CI_{1-\alpha}(\beta_k)$

Confidence region for marginal tests.



- $H_0: \beta_1 = 0$. Do NOT reject
- $H_0: \beta_2 = 0$. Do NOT reject.
- $H_0: \beta_1 = 0$ (do not reject)
  $\Leftrightarrow$ comparing 2 models $y \sim X_2$ vs. $y \sim X_1 + X_2$.

Confidence region for F-test.

$H_0: \beta_1 = \beta_2 = 0$.

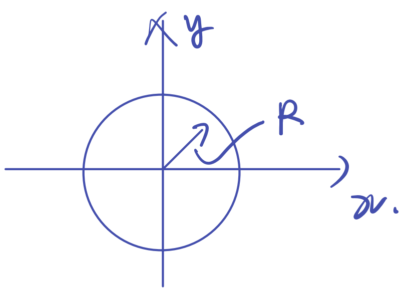
Recall: $F^* = \dfrac{(\underline{C}\,\underline{b})^T(\underline{C}(\underline{X}^T\underline{X})^{-1}\underline{C})^{-1}(\underline{C}\,\underline{b})}{SSE} \cdot \dfrac{n-p}{q}$ $\underline{\gamma} = \underline{0}$, $q = \text{rank}(\underline{C})$.

$F^* = \underline{b}^T \underline{\Omega}\, \underline{b}$
→ PSD matrix. all e-values of Ω are ≥ 0.
$\underline{\Omega}\,\underline{v} = \lambda \underline{v}$ $\forall\, v_i, \lambda_i \Rightarrow \lambda_i \geq 0$.

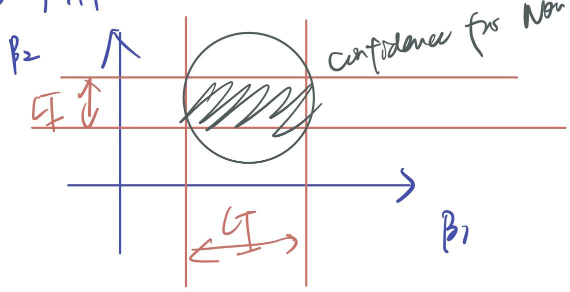
$F^* = \underline{b}^T \Omega\, \underline{b} \leq$ critical value → non-rejection region
quadratic form on $\underline{b}$

Assume. If $\underline{\Omega} = \underline{I}$, $\underline{b}^T \underline{I}\, \underline{b} = \underline{b}^T \underline{b} = \sum_{i=1}^{p-1} b_i^2 \leq$ C.V.

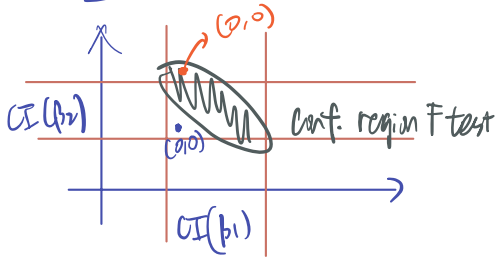
$x^2 + y^2 \leq R^2$



For β_1, β_2



→ If $\Omega \neq I \Rightarrow b^T \Omega b$ defines an ellipsoid.



$H_0: \beta_1 = \beta_2 = 0$, reject H_0.
(1) $H_0: \beta_1 = 0$ do not reject
(2) $H_0: \beta_2 = 0$ do not reject
(3) $H_0: \beta_1 = \beta_2 = 0$ reject.
} different scenarios.

H_0: (1) $\beta_1 = 0$ do not reject
(2) $\beta_2 = 0$ reject
(3) $\beta_1 = \beta_2 = 0$ do not reject.

In general, multicollinearity is when the corr(x_j, $\sum_{i \neq j} a_i x_i$) is high. $\Rightarrow \det(X^T X) \approx 0$.
$\Rightarrow$ large standard deviation for b_k's.
Non significant results.

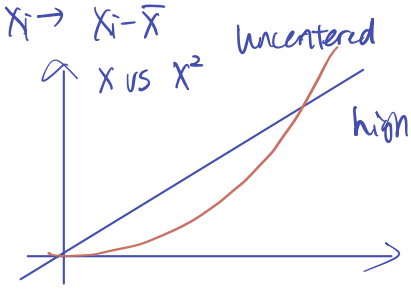
Note: correlation matrix (or pairwise correlations r_{ij}) do NOT measure this. → we need more diagnostics.

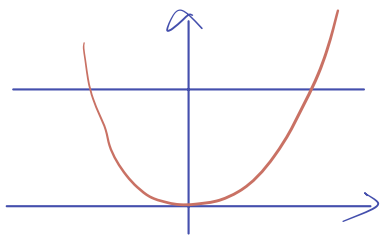
Regression Models with Qualitative and Quantitative Predictors (Chap. 8).

→ One pred, 2nd order model.
$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$

→ One pred, 3rd order model.
$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \beta_{111} X_i^3 + \varepsilon_i$

$X_i \to X_i - \bar{X}$





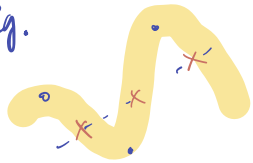
Caution: always center X_i's for polynomial regression.

→ Two predictors, 2nd order:
$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{11} X_{i1}^2 + \beta_{22} X_{i2}^2 + \beta_{12} X_{i1} X_{i2} + \varepsilon_i$
(β_{12}: interaction effect coeff.; $X_{i1} X_{i2}$: interaction term)

effect of X_1 on y depends on X_2

Caution for polynomial model: problem with interpolating & extrapolating.

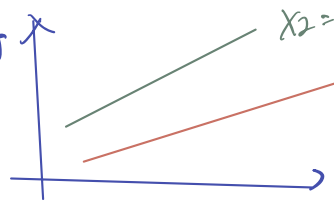
eg.



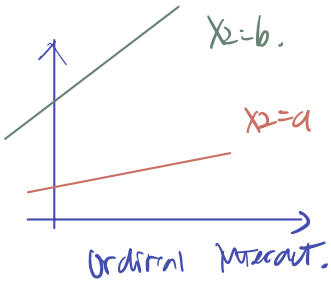
$X \to \pm\infty$ hyper order poly. oscillate & diverge.

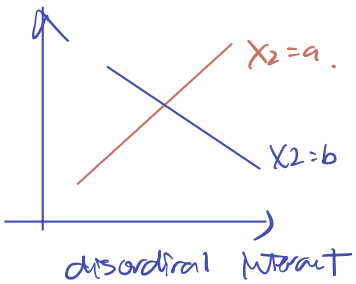
$E(y) = 10 + 2X_1 + 5X_2 + 0.5 X_1 X_2$
$X_2 = 1$, $E(y) = 10 + 2X_1 + 5 + 0.5X_1 = 15 + 2.5 X_1$
$X_2 = 3$, $E(y) = 25 + 3.5 X_1$

Interaction plot:



If parallel, β_{12} is NOT significant.





β_{12} is significant

quantitative predictors : numerical
qualitative " : categorical.
C categories → C-1 var. ("dummy")
Cat C: $x_1 = x_2 = \cdots = x_{c-1} = 0$.
↳ baseline category.

Why don't we use C variables, e.g.: $x_c = \begin{cases} 1 & \text{if baseline} \\ 0 & \text{o.w.} \end{cases}$

	x_1	x_2	⋯	x_c	"x_0"
Cat 1	1	0	⋯	0	1
Cat 2	0	1		0	1
Cat 3	0	0		0	1
⋮	⋮			⋮	⋮
Cat C	0	0	⋯	1	1

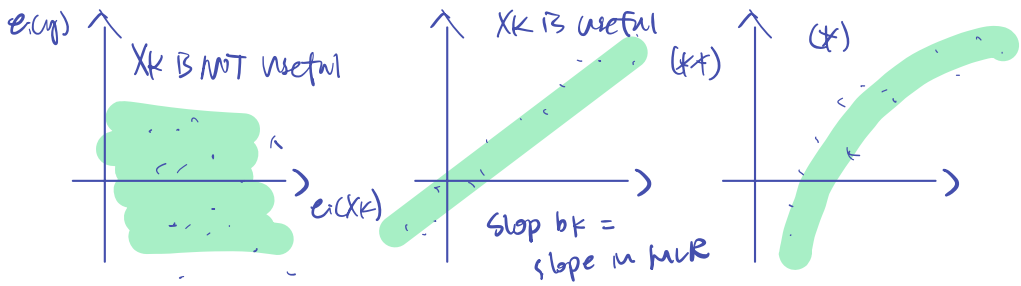
⇒ Can't invert X^TX.

$\sum_{1 \le i \le c} x_i$

·) If all var's are qualitative ⇒ ANOVA (analysis of var. chp. 16)
·) If var's are both qual. & quant. ⇒ ANCOVA (analysis of cov.)
↳ main (e.g. dosage of drug).

✱✱ Diagnostics
·) Added variable / partial reg. plot
y ~ all pred's but x_k → residual $e_i(y)$
x_k ~ all " " " → residual $e_i(x_k)$ } Contain unexplained variation in y & x_k.

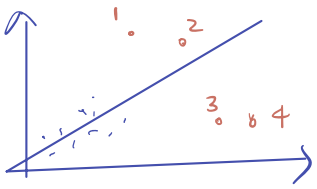
Plot $e_i(y)$ vs. $e_i(x_k)$



The plot should shows no patterns normally.
$y_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} + \cdots + b_{p-1} x_{i,p-1}$

Check: - improper fuction form (*)
- whether X* is useful (**).
- non constant variance. (like the cone shape)

) outlying observations



outliers?

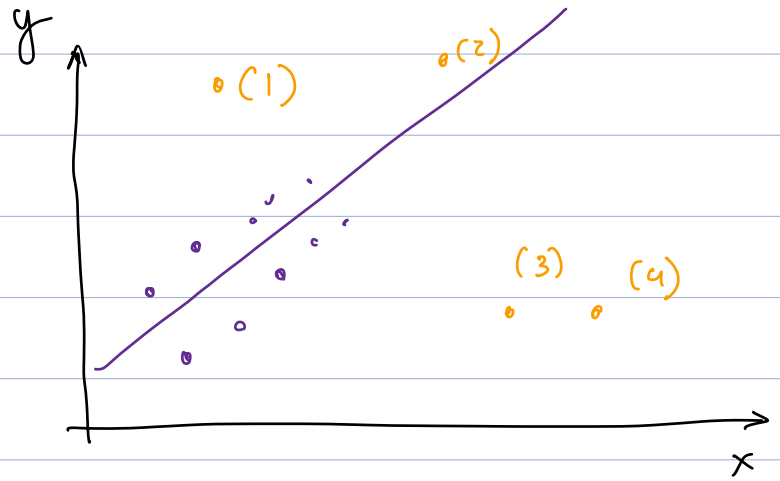
Outlying wrt X 2, 3, 4.
" Y 1, , 3, 4
" X & Y. 3, 4.

Compare with fitted line.

Lecture 15, ASDA 413/613

Today: → outlying obsn's / outliers (10.2)
→ hidden extrapolations (10.3)

0) outlying obsn's



→ outlying wrt x (2), (3), (4)
→ outlying wrt y (1), (3), (4) → compare w/ fitted values
→ outlying wrt x & y (3), (4)

$e_i = Y_i - \hat{Y}_i \rightarrow e_i^* = \frac{e_i}{\sqrt{MSE}}$ semi-studentized residuals

$\underline{e} = \underline{Y} - \underline{\hat{Y}} = (\underline{I} - \underline{H})\underline{Y}$

$$\text{var}(\underline{e}) = \text{var}((\underline{I}-\underline{H})\underline{Y}) = (\underline{I}-\underline{H})\,\text{var}(\underline{Y})\,(\underline{I}-\underline{H})^T$$

symm.

$$= (\underline{I}-\underline{H})\,\text{var}(\underline{Y})(\underline{I}-\underline{H}) = \sigma^2(\underline{I}-\underline{H})(\underline{I}-\underline{H}) =$$

$$\sigma^2 = \sigma^2(\underline{I}-\underline{H})$$

idempotent

$$\text{var}(e_i) = \sigma^2(1-h_{ii}) \qquad \text{cov}(e_i, e_j) = -\sigma^2 h_{ij}$$

h_{ii} = ith diagonal term of $\underline{H}$

$$h_{ii} = \left[\underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T\right]_{ii} = \underline{x}_i(X^TX)^{-1}\underline{x}_i^T$$

↳ Exercise: verify

↳ ith obs'n vector

def'n: h_{ii} = leverage obs'n i

$$h_{ij} = \underline{x}_i(X^TX)^{-1}\underline{x}_j^T \quad \rightarrow \text{verify!!}$$

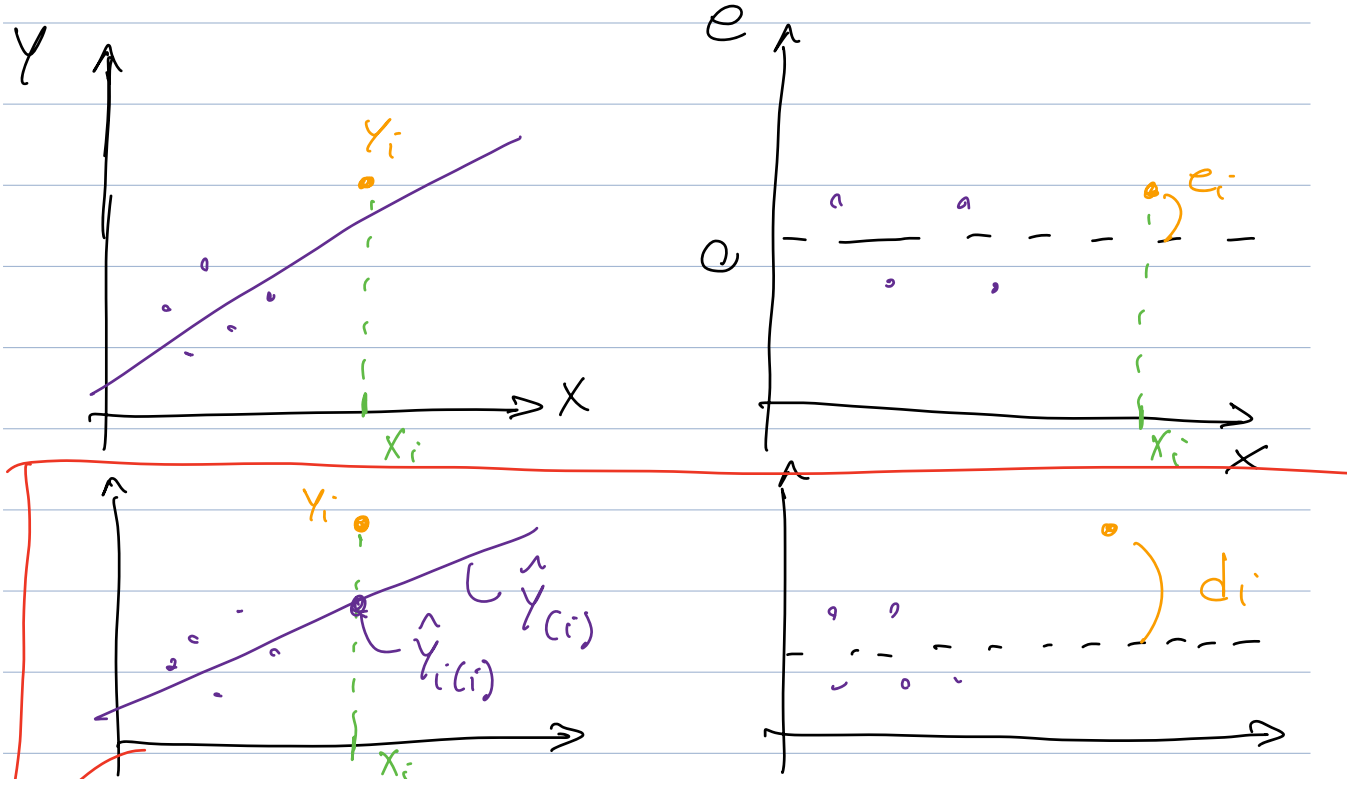
Note: x_i here is ith row of X, in row form! i.e. $x_i \in \mathbb{R}^{1\times p}$

- To correctly studentize residuals we compute:

$$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{\sqrt{MSE\,(1-h_{ii})}}$$

→ "internally" studentized residuals

WARNING: no consistent naming of standardized, studentized residuals, etc, especially in R! (always check the documentation)



↳ Idea: let's forget about Y_i!

d_i = deleted residual

$d_i = Y_i - \hat{Y}_{i(i)}$ ↳ removed obsn (i)

↳ pred'n at point x_i

Fact: $d_i = \frac{e_i}{1 - h_{ii}}$ (C.A. trick!)

! can compute this w/o fitting n different regressions

Proof sketch:

Fact: $(X_{(i)}^T X_{(i)})^{-1} = (X^T X)^{-1} +$ (p × p)

$$(X^T X)^{-1} \; x_i^T \; x_i \; (X^T X)^{-1}$$

p × p, p × 1, 1 × p, p × p

↳ design matrix w/ i^{th} row removed

From this we get:

$$b_{(i)} = b - (X^T X)^{-1} x_i^T \frac{e_i}{1 - h_{ii}}$$ (no refitting!)

$$d_i = Y_i - \hat{Y}_{i(i)} = Y_i - x_i b_{(i)} =$$

$$Y_i - x_i b + \frac{x_i (X^T X)^{-1} x_i^T e_i}{1 - h_{ii}} =$$

($x_i b = \hat{Y}_i$)

$$e_i + \frac{h_{ii} e_i}{1 - h_{ii}} = \frac{e_i (1 - h_{ii}) + h_{ii} e_i}{1 - h_{ii}} =$$

$$\frac{e_i}{1 - h_{ii}} \Rightarrow \boxed{d_i = \frac{e_i}{1 - h_{ii}}}$$ ∎

When is d_i large?

- e_i is large
- $h_{ii} \approx 1$

both

- e_i is large, $h_{ii} \approx 1$

$1 \times p$ $\quad p \times p \quad p \times 1$

$$S(d_i) = MSE_{(i)} \left[1 + \underline{X_i} \left(\underline{X}_{(i)}^T \underline{X}_{(i)} \right)^{-1} \underline{X_i}^T \right]$$

- $MSE_{(i)}$ → MSE when i^{th} case removed
- $\underline{X}_{(i)}$ → design matrix w/ i^{th} case removed
- $\underline{X_i}$ → i^{th} case

$$= \frac{MSE_{(i)}}{1 - h_{ii}}$$

Fact: $(n-p)MSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}}$

$$\Rightarrow SSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}} \Rightarrow$$

$$\frac{SSE\,(1-h_{ii}) - e_i^2}{n-p-1} = MSE_{(i)}(1-h_{ii})$$

studentize d_i:

$$t_i = \frac{d_i}{s(d_i)} = \frac{\frac{e_i}{1-h_{ii}}}{\sqrt{\frac{MSE_{(i)}}{1-h_{ii}}}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

↝ w/o refitting the regression line!

$$= e_i \sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}}$$

t_i = externally studentized residuals

If $MSE_{(i)} \approx MSE$ then $r_i \approx t_i$

ISR ESR

$$t_i \sim t(n-1-p)$$

$n-1$: # obsn's; p: # predictors

Test for outliers:

Outlier $|t_i| > t\left(1 - \frac{\alpha}{2n}, n-p-1\right)$

Bonferroni correction
= simultaneous test correction
(conservative!)

but it works!

Outlying obsn's wrt X:

e.g. (2)

related to Mahalanobis distance

H = distance measure in the space of pred's

Properties:

- $0 \leq h_{ii} \leq 1$ (w/o proof)
- $\sum_{j=1}^{n} h_{jj} = p$

$$\sum_{j=1} h_{jj} = \text{trace}(\underline{H}) = \text{tr}\left(\underline{X}\,(\underline{X}^T\underline{X})^{-1}\,\underline{X}^T\right) =$$

$\text{tr}(\underline{A}\,\underline{B}) = \text{tr}(\underline{B}\,\underline{A})$

$$= \text{tr}\left((\underline{X}^T\underline{X})(\underline{X}^T\underline{X})^{-1}\right) = \text{tr}(I_p) = p$$

$h_{ii} \approx$ distance $\underline{x}_i$ to the center of the data in predictor space

$$\hat{Y}_i = (HY)_i = h_{i1}Y_1 + h_{i2}Y_2 + \ldots + h_{ii}Y_i + \ldots\ldots + h_{in}Y_n$$

If $h_{ii} = 1 \Rightarrow \text{var}(e_i) = 0 \Rightarrow e_i = 0 \Rightarrow Y_i = \hat{Y}_i$

↳ highest possible lvg

h_{ii} = leverage, "influence" of i^{th} obsn'n on fitted values

Rule of thumb: (to detect outliers wrt X)

$$h_{ii} > \frac{2p}{n}$$

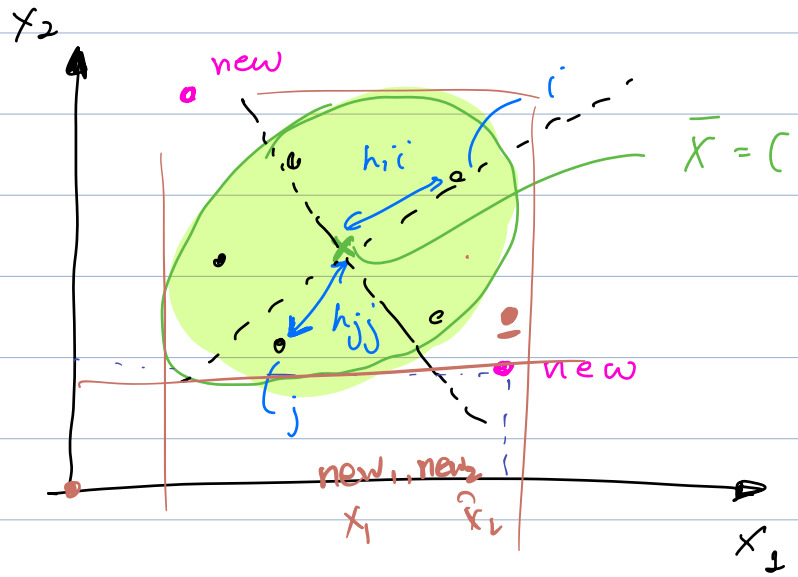
"2x the avg leverage"

$$\frac{\sum h_{ii}}{n} = \frac{p}{n}$$

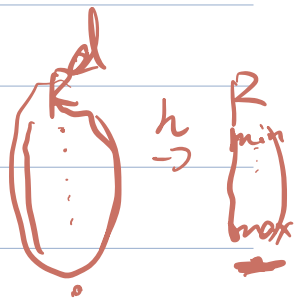
$h_{ii} > 0.5$ = "high leverage" NOT TESTED!

$0.2 < h_{ii} < 0.5$ = "moderate leverage"

→ Hidden extrapolations 只有1d的时候 $h_{ii} = \frac{1}{n} + \frac{1}{S_{xx}}(x_i - \bar{x})^2$ 可以有距离.

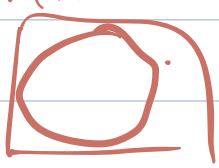


$\bar{x} = (\bar{x}_1, \bar{x}_2)$



(0,0)

$x_{(0,0)} = 0$



$h_{new,new} = x_{new}^T (X^T X)^{-1} x_{new}$

$\min(h_{ii}) \leq h_{new,new} \leq \max(h_{ii})$ => no hidden extrapolation!

MT2 stops here!

① Q3 (b)

② Q2 (ii).

③ hidden extrapolition vs. outliers.