X. Consequences of a Problematic Regression Space

Usual situation

DATA: $(Y_i, Z_i^T)^T$, $i = 1, \dots, n$, $Z_i = (Z_{i1}, \dots, Z_{ip})^T \in \mathcal{Z}$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} = (1_n, X^1, \dots, X^{k-1})$$

$$X_i = t_X(Z_i), \quad i = 1, \dots, n$$

Full-rank linear model with intercept assumed

$$Y | Z \sim (X\beta, \sigma^2 I_n), \quad \operatorname{rank}(X_{n \times k}) = k < n$$

From now on:

- $\mathbb{E}(Y|Z) \in \mathcal{M}(X)$ (= model is correct)
- $X = (1, \dots)$ (intercept in model)
- $\operatorname{rank}(X) = k$ (full-rank)

10.1 Multicollinearity

- Principal assumption of linear model $\mathbb{E}(Y|Z) \in \mathcal{U}(X)$

Many covariates available $\Rightarrow$ Choose X with many columns (= all available covariates)?

10.1.1 Singular value decomposition of a model matrix

$$X_{n\times k} = \sum_{j=0}^{k-1} \sqrt{\lambda_j}\, u_j v_j^T = U D V^T$$

- $U_{n\times k} = (u_0, \dots, u_{k-1})$
  - first k orthonormal eigenvectors of the $n\times n$ matrix XX^T
- $V_{k\times k} = (v_0, \dots, v_{k-1})$
  - (all) orthonormal eigenvectors of the $k\times k$ (invertible) matrix X^TX
- $D = \operatorname{diag}(d_0, \dots, d_{k-1})$, $d_j = \sqrt{\lambda_j}$, $j = 0, \dots, k-1$

$d_0^2 \quad\quad d_{k-1}^2$

$\lambda_0 \geq \dots \geq \lambda_{k-1} > 0$

- the first k eigenvalues of matrix XX^T
- (all) eigenvalues of matrix X^TX

$d_0 \geq d_1 \geq \dots \geq d_{k-1} > 0$: singular values of matrix X

$\sqrt{\lambda_0} \quad \sqrt{\lambda_1} \quad \sqrt{\lambda_{k-1}}$

- $X^TX = \sum_{j=0}^{k-1} \lambda_j v_j v_j^T = V \Lambda V^T$, $\Lambda = \text{diag}(\lambda_0, \dots, \lambda_{k-1})$

$$= \sum_{j=0}^{k-1} d_j^2 v_j v_j^T = V D^2 V^T$$

- $(X^TX)^{-1} = \sum_{j=0}^{k-1} \lambda_j^{-1} v_j v_j^T = V \Lambda^{-1} V^T$

$$= \sum_{j=0}^{k-1} d_j^{-2} v_j v_j^T$$

Is it inverse?

$V \Lambda \underbrace{V^T \cdot V}_{I} \Lambda^{-1} V^T = V \Lambda \Lambda^{-1} V^T = V V^T = I$ yes

- $\text{tr}(X^TX)^{-1} = \sum_{j=0}^{k-1} \lambda_j^{-1} = \sum_{j=0}^{k-1} \frac{1}{d_j^2}$

REMARK: If $\text{rank}(X) = r < k$, then

$$d_0 \geq \dots \geq d_{r-1} > d_r = \dots = d_{k-1} = 0$$

Moore-Penrose pseudoinverse of X^TX is

$$(X^TX)^+ = \sum_{j=0}^{r-1} \frac{1}{d_j^2} v_j v_j^T$$

What happens if $r = k$ (full-rank) but

$d_{k-1} \to 0$ then

- X^TX tends to a singular matrix
- columns of X tend to being linearly dependent
- $\text{tr}((X^TX)^{-1}) \to \infty$

$\equiv$ multicollinearity

10.1.2 Multicollinearity and its impact on precision of the LSE

(i) $\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_n)^T = HY$, $H = X(X^TX)^{-1}X^T = QQ^T$

orthonormal basis of $\mathcal{M}(X)$

$= \text{BLUE of } \mu = X\beta = E(Y|Z)$

$\text{var}(\hat{Y}|Z) = \sigma^2 H$

$\sum_{i=1}^{n} \text{var}(\hat{Y}_i|Z) = \text{tr}(\sigma^2 H) = \sigma^2 \text{tr}(H) = \sigma^2 \text{tr}(QQ^T) =$

$= \sigma^2 \text{tr}(Q^TQ) = \sigma^2 \text{tr}(I_k) = \sigma^2 k$

no problem

(ii) $\hat{\beta} = (\hat{\beta}_0, \dots, \hat{\beta}_{k-1})^T = (X^TX)^{-1}X^TY$

$= \text{BLUE of } \beta$

$\text{var}(\hat{\beta}|Z) = \sigma^2 (X^TX)^{-1}$

$\sum_{j=0}^{k-1} \text{var}(\hat{\beta}_j|Z) = \text{tr}(\sigma^2 (X^TX)^{-1}) = \sigma^2 \text{tr}((X^TX)^{-1}) =$

$= \sigma^2 \sum_{j=0}^{k-1} \frac{1}{d_j^2}$

$\downarrow \; d_{k-1} \to 0$

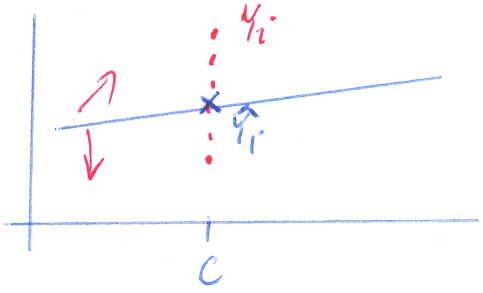
∞

Multicollinearity

- no impact on precision of LSE of $\mu = E(Y|Z)$
- possibly serious inflation of the standard errors of LSE of β

Extreme case $X = (1, c1)$, $\operatorname{rank}(X_{n\times 2}) = 1$

model $\mathbb{E}(Y/Z=z) = \beta_0 + \beta_1 c$



$\hat{y}_1 = \ldots = \hat{y}_n = \bar{y}$

"LSE" of $\beta = (\beta_0, \beta_1)^T$:

We want to write

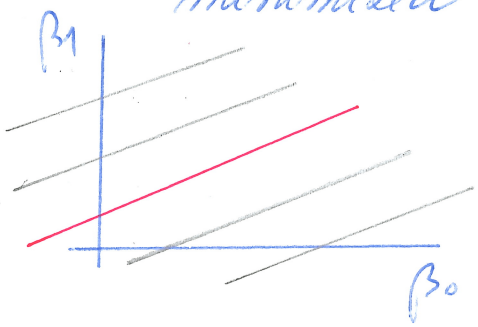
$$\bar{y} = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot c$$

x_i "x_i (regressor)

→ infinitely many pairs $(\hat{\beta}_0, \hat{\beta}_1)^T$ satisfy $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot c$

= "$\operatorname{var}(\hat{\beta}_0/Z) = \operatorname{var}(\hat{\beta}_1/Z) = \infty$"

Differently: $SS(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 c)^2$ is minimized for $(\beta_0, \beta_1)^T$ satisfying $\beta_0 + \beta_1 c = \bar{y}$



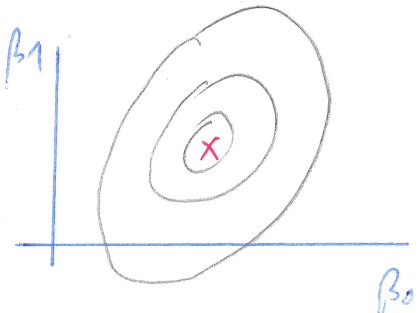
← contours of $SS(\beta)$

= "inverted roof"

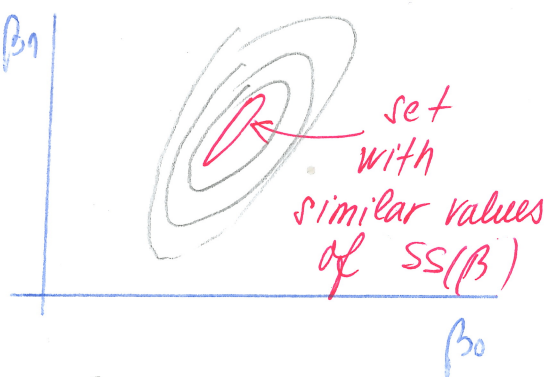
(roof of a house after hurricane)

$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, $\operatorname{rank}(X) = 2$ with singular values $d_0 \geq d_1 > 0$

$$SS(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



$d_1 \to 0$ →



also remember that $\frac{\partial^2 SS}{\partial \beta \partial \beta^2} = 2X^TX$

Lemma 10.1 Bias in estimation of the squared norms

MOTIVATION: suppose we want to estimate

$$\|\mathbb{E}(Y|Z)\|^2 = \|X\beta\|^2 \quad \text{or} \quad \|\beta\|^2$$

We know: $\mathbb{E}(\hat{Y}|Z) = X\beta \quad (= \mathbb{E}(Y|Z))$

$$\mathbb{E}(\hat{\beta}|Z) = \beta$$

$\rightarrow \|\hat{Y}\|^2$ estimator of $\|X\beta\|^2$

$\|\hat{\beta}\|^2$ estimator of $\|\beta\|^2$

- both estimators are biased, but how large is this bias?

$\rightarrow$ Let $Y|Z \sim (X\beta, \sigma^2 I_n)$, $\operatorname{rank}(X_{n\times k}) = k$. Then

$$\mathbb{E}\left(\|\hat{Y}\|^2 - \|X\beta\|^2 \,|\, Z\right) = \sigma^2 k \quad \leftarrow \text{no problem (multicollinear.)}$$

$$\mathbb{E}\left(\|\hat{\beta}\|^2 - \|\beta\|^2 \,|\, Z\right) = \sigma^2 \operatorname{tr}\{(X^TX)^{-1}\} \quad \downarrow \text{problem}$$

Proof: all expectations conditional, given Z:

- $\mathbb{E}\|\hat{Y} - X\beta\|^2 = \mathbb{E}\sum_{i=1}^{n} \underbrace{(\hat{Y}_{i\cdot} - X_{i\cdot}^T\beta)^2}_{\operatorname{var}(\hat{Y}_{i\cdot})} = \sum_{i=1}^{n} \operatorname{var}(\hat{Y}_{i\cdot}) = \operatorname{tr}(\sigma^2 H) = \sigma^2 k$

- $\mathbb{E}\|\hat{Y} - X\beta\|^2 = \mathbb{E}(\hat{Y} - X\beta)^T(\hat{Y} - X\beta) = \mathbb{E}\hat{Y}^T\hat{Y} + \mathbb{E}(\beta^T X^T X\beta) - 2\mathbb{E}(\beta^T X^T \hat{Y})$

$$= \mathbb{E}\|\hat{Y}\|^2 + \|X\beta\|^2 - 2\beta^T X^T \underbrace{\mathbb{E}\hat{Y}}_{X\beta} = \mathbb{E}\|\hat{Y}\|^2 - \|X\beta\|^2$$

$$\Rightarrow \mathbb{E}\|\hat{Y}\|^2 - \|X\beta\|^2 = \sigma^2 k$$

In the same way, since $\mathbb{E}\hat{\beta} = \beta$:

- $\mathbb{E}\|\hat{\beta} - \beta\|^2 = \sum_{j=0}^{k-1} \operatorname{var}\hat{\beta}_j = \operatorname{tr}(\operatorname{var}\hat{\beta}) = \sigma^2 \operatorname{tr}\{\sigma^2(X^TX)^{-1}\} = \sigma^2 \operatorname{tr}(X^TX)^{-1}$

- $\mathbb{E}\|\hat{\beta} - \beta\|^2 = \mathbb{E}\|\hat{\beta}\|^2 - \|\beta\|^2 \Rightarrow \mathbb{E}\|\hat{\beta}\|^2 - \|\beta\|^2 = \sigma^2 \operatorname{tr}(X^TX)^{-1}$

10.1.3 Variance inflation factor and tolerance

6

most important items from slides:

$$T_Y = \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = \sqrt{SS_T}$$

$$R^2 = 1 - \frac{SS_e}{SS_T} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\underbrace{\sum (Y_i - \bar{Y})^2}_{T_Y^2}}$$

7

Consider linear model with

X^j = response , $X^{(-j)} = (\mathbb{1}, X^1, \dots, X^{j-1}, X^{j+1}, \dots, X^{k-1})$ as model matrix

$\rightarrow \hat{X}^j = \begin{pmatrix} \hat{X}_{1j} \\ \vdots \\ \hat{X}_{nj} \end{pmatrix}$ = fitted values

, $\bar{X}^j$ = sample mean over X^j

$$T_j = \sqrt{\sum_i (X_{ij} - \bar{X}^j)^2} = \sqrt{SS_T \text{ (model } M_j : X^j \sim X^{(-j)})}$$

$$R_j^2 = 1 - \frac{\sum_i (X_{ij} - \hat{X}_{ij})^2}{\sum_i (X_{ij} - \bar{X}^j)^2} = 1 - \frac{SS_e (X^j \sim X^{(-j)})}{SS_T (X^j \sim X^{(-j)})}$$

8

slide + additional comments

R^2 = how closely can Y be expressed as a linear combination of $X^1, \dots, X^{k-1}$

R_j^2 = how closely can X^j be expressed as a linear combination of $X^1, \dots, X^{j-1}, X^{j+1}, \dots, X^{k-1}$

$R_j^2 = 1$ = X^j is a perfect linear combination of $X^1, \dots, X^{j-1}, X^{j+1}, \dots, X^{k-1}$

Theorem 10.2 Estimated variances of the LSE of the regression coefficients

For a given dataset for which a linear model $Y|Z \sim (X\beta, \sigma^2 I_n)$, $\text{rank}(X_{n\times k}) = k$, $X = (1, X^1, \ldots, X^{k-1})$ is applied, diagonal elements of the matrix $\widehat{\text{var}}(\beta|Z) = MSE(X^TX)^{-1}$, can also be calculated, for $j = 1, \ldots, k-1$ as

$$\underbrace{\widehat{\text{var}}(\hat{\beta}_j|Z)}_{\text{square root of this is standard error of } \hat{\beta}_j} = \underbrace{\left(\frac{T_y}{T_j}\right)^2 \cdot \frac{1-R^2}{n-k}}_{\text{not related to collinearity of regressors}} \cdot \underbrace{\frac{1}{1-R_j^2}}_{\text{quantifies strength of collinearity of regressor } j \text{ with remaining regressors.}}$$

Proof: see notes (if interested)

Def 10.1 Variance inflation factor and tolerance

For given $j = 1, \ldots, k-1$, the variance inflation factor and the tolerance of the jth regressor of the linear model $Y|Z \sim (X\beta, \sigma^2 I_n)$, $\text{rank}(X_{n\times k}) = k$ are values VIF_j and $Toler_j$, respectively, defined as

$$VIF_j = \frac{1}{1-R_j^2}, \quad Toler_j = 1 - R_j^2 = \frac{1}{VIF_j}$$

$$R_j^2 = 0 \Rightarrow VIF_j = 1, \quad Toler_j = 1$$

$$R_j^2 \to 1 \Rightarrow VIF_j \to \infty, \quad Toler_j \to 0.$$

Interpretation and use of VIF

Remember: $(1-\alpha)100\%$ confidence interval for β_j (under normality):

$$\hat{\beta}_j \pm t_{n-k}\left(1-\frac{\alpha}{2}\right)\sqrt{\widehat{var}(\hat{\beta}_j|Z)}$$

$$\hat{\beta}_j \pm t_{n-k}\left(1-\frac{\alpha}{2}\right)\boxed{\sqrt{\cdots}}\sqrt{VIF_j}$$

$\sqrt{\cdots}$: not related to collinearity

$\sqrt{VIF_j}$: related to collinearity

$$\rightarrow VIF_j = \left(\frac{VOL_j}{VOL_{0,j}}\right)^2$$

VOL_j = length (volume) of the confidence interval for β_j

$VOL_{0,j}$ = length (volume) ... if it was $R_j^2 = 0$

Note: In case of categorical covariates their collinearity is viewed through possible inflation of $\hat{\beta}^Z$ = LSE of regress. parameters based on (pseudo) contrast parameterization

→ confidence ellipsoid for β^Z:

$$\propto \beta^Z: (\beta^Z - \hat{\beta}^Z)^T (MSe\, \underline{V}^Z)^{-1} (\beta^Z - \hat{\beta}^Z) < (G-1) F_{G-1, n-k}(1-\alpha)\}$$

(G: # of levels; $\underline{V}^Z$: block of matrix $(X^TX)^{-1}$)

→ volume VOL_Z

→ $VOL_{0,Z}$ = volume of above ellipsoid if all columns in X corresponding to (pseudo) contrasts are orthogonal to remaining columns

→ generalized VIF

$$gVIF_z = \left(\frac{Vol_z}{Vol_{0,z}}\right)^2$$

- to take into account "dimensionality"

$$\left(\frac{Vol_z}{Vol_{0,z}}\right)^{\frac{1}{m}}, \quad m = G-1 \; (= \#\text{ of } \beta^z)$$

is to be calculated and compared,

i.e., calculate and compare values

of $gVIF_z^{\frac{1}{2m}}$, $m = G-1$

→ R function vif

(package car)

10.1.4 Basic treatment of multicollinearity

12

remember: $\sum \text{var}(\hat{y}_i | z) = \sigma^2 k$

$$\sum \text{var}(\hat{\beta}_j | z) = \sigma^2 \underbrace{\sum_{j=0}^{k-1} \frac{1}{d_j^2}}$$

$\to \infty$, $d_{k-1} \to 0$

+ comments using slides

+13

Example: $Y = IQ$

14

$z = (\text{gender}, zn7, zn8)$

$zn7 \approx zn8$ (corr = 0,95)

$$E(Y | \ldots) = \beta_0 + \beta_1 \underset{(boy)}{\mathbb{I}} + \underbrace{\beta_2 zn7 + \beta_3 zn8}$$

$$\approx \beta_2 zn + \beta_3 zn$$

$$= \underbrace{(\beta_2 + \beta_3)} \cdot zn$$

this sum identifiable from data, how to factorize it if $zn7$ and $zn8$ both in model?

$\to \infty$ many ways

15-18

REMEMBER: If multicollinearity not treated it is very dangerous to claim that some factor is not associated with response because it's non-significant in the model!