Modeling Problems

Structural Problems

- Multicollinearity
- Influential pts

Violation of Model Assumptions

- Heteroskedasticity
- Non-Normal residuals
- False assumption of linearity

Multicollinearity

Problem two or more predictors are highly correlated

Design Matrix

$$X = \begin{pmatrix} | & | & | & & | \\ \mathbf{1}_n & x_1 & x_2 & \cdots & x_{p-1} \\ | & | & | & & | \end{pmatrix}$$

↑ one of these columns is linearly dependent (or very close) on the others

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

↑ multicollinearity can cause numerical difficulties when calculating LS Ests.

Ex: $y \sim x_1 + x_2$

but $x_2 = 4x_1$

true parameters

$\beta^* = \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}$

Ex Simulation:

$$y_i = 1 + 2x_{1i} + 4x_{2i} + \varepsilon_i$$

$$\Updownarrow$$

$$y_i = 1 + 2x_{1i} + 4(4x_{1i}) + \varepsilon_i$$

$$= 1 + 18x_{1i} + 0x_{2i} + \varepsilon_i \quad \leftarrow \tilde{\beta}^* = \begin{pmatrix} 1 \\ 18 \\ 0 \end{pmatrix}$$

$$\Updownarrow$$

$$y_i = 1 + 22x_{1i} - x_{2i} + \varepsilon_i \quad \leftarrow \tilde{\beta}^* = \begin{pmatrix} 1 \\ 22 \\ -1 \end{pmatrix}$$

↳ non-identifiability

Damage

① $\hat{\beta} = (X^TX)^{-1}X^TY$

$$X^TX = U\Lambda U^T = U \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \approx 0 \end{pmatrix} U^T$$

$$(X^TX)^{-1} = U\Lambda^{-1}U^T = U \begin{pmatrix} 1/\lambda_1 & & & \\ & 1/\lambda_2 & & \\ & & \ddots & \\ & & & 1/\lambda_p \end{pmatrix} U^T$$

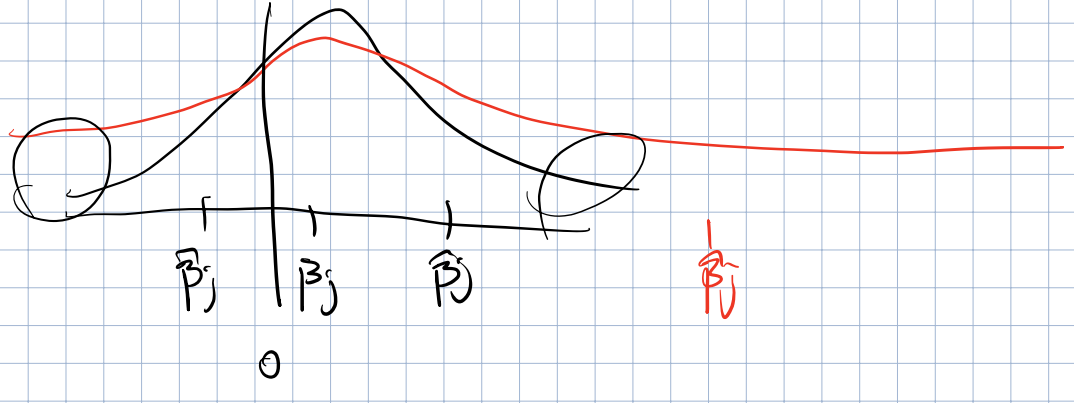
② $Var(\hat{\beta}) = \sigma^2 (X^TX)^{-1}$ ↑↑ when strong multicollinearity exists.

③ Impact on Inference:

$$t = \frac{\hat{\beta}_j - 0}{\hat{SE}(\hat{\beta}_j)}$$ ↑ ↓

losing statistical power harder for me to detect signal

no multicollinearity



Symptoms:

- When you add a predictor to the model, the estimates of your other $\hat{\beta}_j$'s vary a lot, some times even changing signs
- Global F-test rejects H_0 & all the indiv. t-tests fail to reject.

Ex: $corr(X_1, X_2) = 0$

No Multicollinearity

Model	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y \sim X_1$	-1	NA
$Y \sim X_1 + X_2$	-1	-5

$corr(X_1, X_2) = 0.9$

Multicollinearity

Model	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y \sim X_1$	-1	NA
	(*big change)	
$Y \sim X_1 + X_2$	10	-5

ANOVA Table

Typ 1 ANOVA Table

	SS	F	p	
X_1	↑	↑	↓	***
X_2	↓	↓	↑	NS

switch order

	SS	F	p	
X_2	↑	↑	↓	***
X_1	↓	↓	↑	NS

Notes:

- In reality, multicollinearity is always present
  ↳ our task is figuring out how much we can live w/
- unchecked multicollinearity makes it hard to understand the effect of each predictor on Y.

Detection:

① Correlation Matrix (Naive)

	X_1	X_2	X_3	X_4
X_1	1	.9	.2	.2
X_2	.9	1	.2	.2
X_3	.2	.2	1	.2
X_4	.2	.2	.2	1

② Variance Inflation Factor (VIF)

measures how much the variance of $\hat{\beta}$ is inflated by adding X_j to the model

$$VIF = \frac{1}{1-R_j^2} \quad \text{where}$$

R_j^2 is the coef. of det when you regress $X_j \sim X_1 + X_2 \cdots + X_{j-1} + X_{j+1} \cdots X_p$

If $VIF \approx 1 \iff X_j$ is nearly uncorrelated w/ other preds.

$1 \leq VIF \leq 4 \iff$ "light" multicollinearity

$4 \leq VIF \leq 10 \iff$ "moderate" "

$10 \leq VIF \iff$ "severe" "

Solutions

① Drop some of the suspicious predictors w/ high VIF

② Feature engineer a new feature/predictor which combines the info from the highly collinear preds

More Complicated Approaches:

① Regularized regression
- Ridge penalizes $\|\beta\|_2^2$
- LASSO penalizes $\|\beta\|_1$

② Dimension reduction on X

e.g. SVD/PCA

③ Partial Least Squares