Modeling Problems

Structural Problems

- Multicollinearity
- Influential Pts

Violations of Model Assumptions

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

Multicollinearity

$\boxed{\text{problem}}$: two or more predictors are highly correlated

$\Downarrow$

could cause numerical difficulties when calculating LS Estimates

Design Matrix

$$X = \underbrace{\begin{pmatrix} \mathbb{1}_n & X_1 & X_2 & \cdots & X_{p-1} \end{pmatrix}}$$

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

$$\text{rank}(X) = p$$

For ex:

$$X_1 \qquad X_2 = \underline{4} X_1$$

$$y_i = 1 + \underline{2} X_{1i} + \underline{4} X_{2i} + \varepsilon_i \qquad \beta^* \begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix}$$

$$\Updownarrow$$

$$y_i = 1 + 2X_{1i} + 4(4X_{1i}) + \varepsilon_i \qquad \tilde{\beta}^* = \begin{pmatrix} 1 \\ 18 \\ 0 \end{pmatrix}$$

$$= 1 + \underline{18} X_{1i} + \varepsilon_i + \underline{0} X_{2i} +$$

$$\Updownarrow$$

$$y_i = 1 + 22X_{1i} + (-1) X_{2i} + \varepsilon_i$$

non-identifiability

Damage

If I don't address this issue

here's some negative consequences:

① $\hat{\beta} = (X^TX)^{-1}X^Ty$

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

$(X^TX)^{-1} = U\Lambda^{-1}U^T \quad \Lambda = \begin{pmatrix} 1/\lambda_1 & & & \\ & 1/\lambda_2 & & \\ & & \ddots & \\ & & & 1/0 \end{pmatrix}$

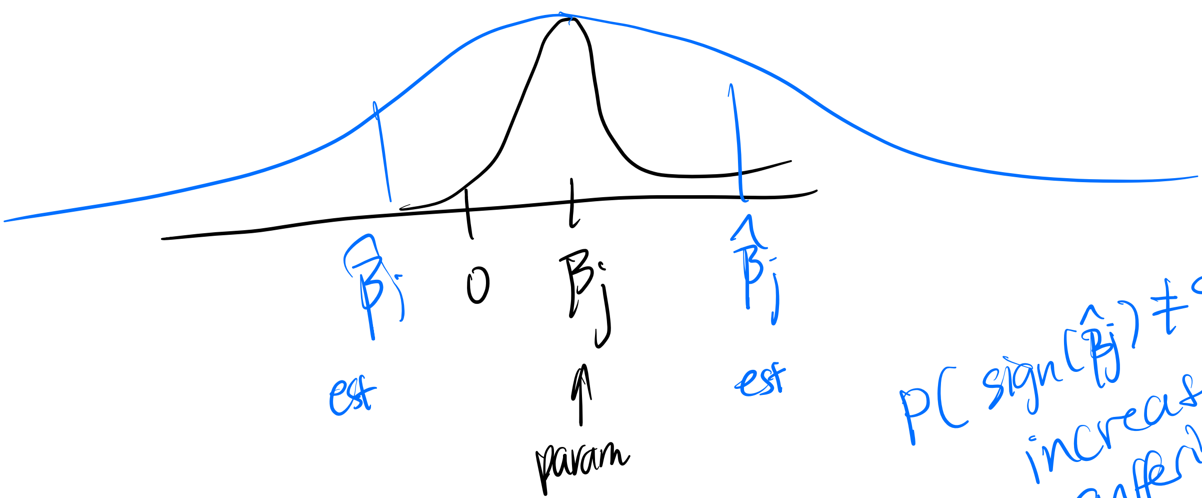
② $Var(\hat{\beta}) = \sigma^2 \underline{(X^TX)^{-1}}$

unaddressed multicol. leads to inflated standard errors.

Impact on Inference:

③ $t = \dfrac{\hat{\beta}_j - 0}{\hat{SE}(\hat{\beta}_j)} \uparrow$ $\downarrow$ losing statistical power

---



$P(\text{sign}(\hat{\beta}_j) \neq \text{sign}(\beta_j))$ increased when suffering from multicollinearity

Symptoms to look for:

- When you add a predictor to the model, the coefficients swing wildly, sometimes even changing signs.
- this is b/c the $\hat{\beta}$ is super sensitive to small changes in X.

EX)

No multicol

$Corr(X_1, X_2) = 0$

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

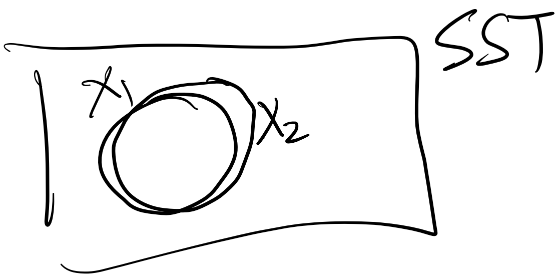
severe multicol...

$Corr(X_1, X_2) = .9$

Model	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y \sim X_1$	-1	NA
$Y \sim X_1 + X_2$	10	-5

Symptom

ANOVA(type = 1) $\rightarrow$ significance may change depending on order



	SS	F	p	
X_1	$\uparrow$	$\uparrow$	$\downarrow$	$\cancel{X}$
X_2	$\downarrow$	$\downarrow$	$\uparrow$	NS

Compare VS.

First var in model gets "credit" for explained SS

	SS	F	p	
X_2	↑	↑	↓	*
X_1	↓	↓	↑	NS

Notes:

- In reality multicollinearity is somewhat always present. Our job is mainly to decide how much we are ok with.

- Unchecked multicollinearity makes it very hard to understand the effect of each predictor on the response.

Detection

① Correlation Matrix (Naive)

② VIF – Variance Inflation Factor

VIF measures how much the variance of $\hat{\beta}$ are inflated by adding a specific predictor to the model

$$VIF_j = \frac{1}{1-R_j^2}$$ where

R_j^2 is the coef. of determination when we regress

$$X_j \sim X_1 + X_2 + \cdots + X_{j-1} + X_{j+1} + \cdots + X_{p-1}$$

If $VIF = 1 \iff$ no correlation b/w X_j & other preds

$1 \leq VIF \leq 4 \iff$ "light" multicol.

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

$VIF \geq 10 \iff$ "severe"

<u>Solutions</u>

1. Drop some of the suspicious looking predictors based on VIF

2. Feature engineer the highly correlated variables into a single new predictor which

carries/summarizes the info in the original preds.

More Complicated Approaches:

(1) Regularized Regression — Ridge ($\|\beta\|_2^2$ penalty), LASSO ($\|\beta\|_1$ penalty)

(2) Dimension Reduction on X
eg. PCA

(3) Partial Least Squares