Announcements

- HW2 regrade
- HW3
- EXAMS Graded

85+ Awesome

(75-85) Solid

(65-75) Passing

(60-65) Push!

< 60 Reevaluate Study Methods

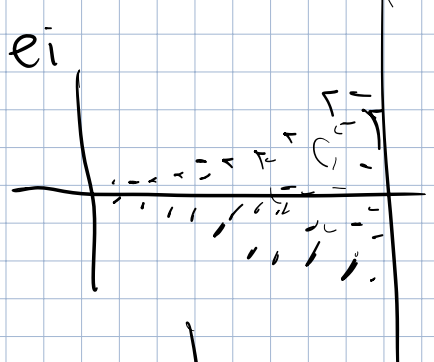
Modeling Problems in Regression

Sometimes we encounter issues which can harm the validity of our model fits or their inference.

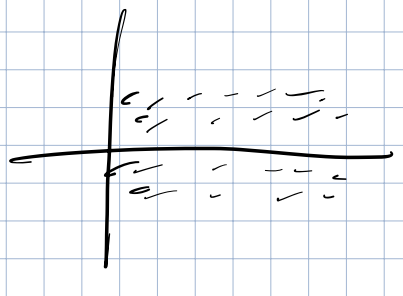
Summary:

- Structural Problems:
  - Multicollinearity
  - influential points
- Violations of Model Assumptions
  - Heteroskedasticity
  - Non-Normal residuals
  - False assumption of linearity

next wk



$$\sigma^2(x) = \sigma^2 x$$



For each problem, I'll try to give you:

A. Damage caused

B. Detection

C. Solution

Let's start w/ Multicollinearity

Problem: Two or more of the predictors are highly correlated.

X ← two (or more) of columns are (almost) linearly dependent

$\Rightarrow$ deficient rank $\text{rank}(X) < p$

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

may not exist, or is highly unstable

what do I mean by unstable?

$X^T X = U \Lambda U^T$

U = orthonormal matrices

Λ = diagonal matrix

$(X^T X)^{-1} = (U \Lambda U^T)^{-1}$ $\left(\begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \neq 0 \end{pmatrix}\right)$

$= (U^T)^{-1} \Lambda^{-1} U^{-1}$

$= U \Lambda^{-1} U^T$

$\Lambda^{-1} = \begin{pmatrix} 1/\lambda_1 & & & \\ & 1/\lambda_2 & & \\ & & \ddots & \\ & & & 1/\lambda_0 \end{pmatrix}$

$VAR(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

When multicollinearity is present:

- the coefficients can swing wildly based on which other predictors are in the model
- the coefficient is highly sensitive to small changes in the model (in the X matrix)

Symptom

Adding or dropping predictors causes massive change in coefs & maybe even sign flips.

EX:

$corr(X_1, X_2) = 0$

A dataset w/ no multicollinearity

Model	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y \sim X_1$	-1	NA
$Y \sim X_1 + X_2$	-1	-2

↖ no major change

$corr(Z_1, Z_2) = -.9$

A dataset w/ severe multicollinearity

Model	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y \sim Z_1$	-1	NA
$Y \sim Z_1 + Z_2$	5	-3

↖ major change

$\Rightarrow$ $se(\hat{\beta}_j)$ are getting inflated

Effect of this:

① when we do a t-test:

$$t = \frac{\hat{\beta}_j - 0}{\hat{SE}(\hat{\beta}_j) \uparrow} \quad \downarrow$$

the inflation in $SE(\hat{\beta}_j)$ makes it harder to reject H_0 $\Rightarrow$ to detect a signal when you have many related predictors in the model.

② If you look @ anova(typ=1) & change the order, the significance / F-stat / p-vals may change dramatically

Note:

- In reality, some degree of multicollinearity always exists. Our job is to judge how much we are willing to live with.

- It becomes hard to understand the effects of individual predictors if they're all highly correlated.

Detection Methods

① Naive → look @ correlation matrix of predictor

Watch out for very high r values in magnitude btw predictors

② VIF = Variance inflation factor

VIF measures how much the variance of the $\hat{\beta}_j$s is inflated by adding that specific pred to a pre-existing model:

For the j^{th} predictor:

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

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

- If X_j is uncorrelated w/ all other preds...

  $VIF = 1$

- If $1 \leq VIF \leq 4 \Rightarrow$ "light multicollinearity"

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

  $VIF > 10 \Rightarrow$ "severe"