Exam 1: mean: 40.2 / 50 sd ≈ 9.6

median 43. 1st quartile (25%): 37. 3rd quartile (75%): 46.5

Diagnosis:

§1. Collinearity Issue

(§ 3.3.3.6 Intro to Stats learning)

- Assumption: $(X^T X)^{-1}$ exists
- Collinearity refers to the situation when 2 or more predictors are closely correlated with each other.
- The presence of collinearity can pose problems in Regression Analysis as it can be difficult to seperate out the individual effects of the collinear variables on Y.

large s.e.$(\hat{\beta}_j)$

In particular, collinearity reduces the accuracy of the estimates of $\hat{\beta}_j$ (if x_j is highly correlated with some other predictors).

To see this, consider a simple linear reg.
$$\underset{n\times 2}{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

suppose two columns are highly correlated.

$$\widehat{Var}\begin{pmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{pmatrix} = \hat{\sigma}^2 (X^T X)^{-1}$$

lecture 12:
$$(X^T X)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}_{2\times 2}$$

$$\det(X^T X) = n\sum_i x_i^2 - \left(\sum_i x_i\right)^2$$

"relatively small" (comparing with n)

$$\Rightarrow \text{s.e.}(\hat{\beta}_0) = \sqrt{\widehat{Var}(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2 \cdot \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}}$$

$$\text{s.e.}(\hat{\beta}_1) = \cdots = \sqrt{\hat{\sigma}^2 \cdot \frac{n}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2}}$$

large s.e.

- How to detect ?

① use the correlation matrix of predictors (large correlation close to 1 $\Rightarrow$ collinearity)

But this method cannot detect multi-collinarity

(one predictor is ~~highly~~ highly correlated with a linear combination of multiple predictors)

E.g $\quad X_1 \approx X_2 + X_3 + X_4$

$$\left.\begin{array}{l} Corr(X_1, X_2) \\ Corr(X_1, X_3) \\ Corr(X_1, X_4) \end{array}\right\} \text{ may not large.}$$

② A recommended method.

- We usually use the R^2 from the regression of

$X_j \sim$ all other predictors (including intercept)

denote them by X_{-j}

for $j = 1, \dots, p$.

- We denote $R^2_{X_j | X_{-j}}$ to be the R^2 from the above regression

- In practice, if $R^2_{X_j | X_{-j}} > 0.9$ or 0.8, this indicates a problematic amount of collinearity.

☆ Equivalently, people also use the " Variance Inflation Factor " (VIF)

For $j = 1, \dots, p$.

$$\mathrm{VIF}_j = \frac{\widehat{\mathrm{Var}}(\hat\beta_j) \text{ from the full regression } Y \sim X_1 + X_2 + \cdots + X_p}{\widehat{\mathrm{Var}}(\hat\beta_j) \text{ from the simple regression } Y \sim X_j}$$

$$= \frac{1}{1 - R^2_{x_j | x_{-j}}}$$

- $\mathrm{VIF}_j > 10$ or 5

---

§2. ☆ Gauss - Markov Theorem.

Under the linear regression model assumptions (n, p), the OLS estimator $\hat\beta_{(p+1)\times 1}$ is the Best Linear Unbiased Estimator of β (true)

(BLUE)

That is, if $\tilde\beta_{(p+1)\times 1}$ is any other unbiased linear estimator of β,

unbiased : $E(\tilde\beta \mid x) = \beta$

linear estimators means that we can write $\tilde\beta$ as a linear combination of $Y_{n\times 1}$, i.e.

$$\tilde\beta = A_{(p+1)\times n} Y_{n\times 1}$$

then

$$\underbrace{\mathrm{Var}(\hat\beta \mid x)}_{(p+1)\times(p+1)} \le \underbrace{\mathrm{Var}(\tilde\beta \mid x)}_{(p+1)\times(p+1)}$$

OLS

in the sense that for any $a \in R^{p+1}$

$$\underbrace{a^T \mathrm{Var}(\hat\beta \mid x)\, a}_{1\times 1} \le \underbrace{a^T \mathrm{Var}(\tilde\beta \mid x)\, a}_{1\times 1}$$

$$\| \qquad\qquad\qquad \|$$

$$\mathrm{Var}(a^T\hat\beta \mid x) \qquad \mathrm{Var}(a^T\tilde\beta \mid x)$$

Proof: Let $\tilde{\beta} = AY$. Then

$$\tilde{\beta} - \hat{\beta}_{OLS} = AY - (X^TX)^{-1}X^TY$$

$$= \underbrace{\left[A - (X^TX)^{-1}X^T\right]}_{C_{(p+1)\times n}} Y$$

Then $\tilde{\beta} = \hat{\beta} + CY = \hat{\beta} + C(X\beta + e)$

This implies

$$E(\tilde{\beta}|X) = E[\hat{\beta}_{OLS}|X] + E[C(X\beta+e)|X]$$

$$= \beta + CX\beta.$$

Because $\tilde{\beta}$ is unbiased, $E(\tilde{\beta}|X) = \beta$.

$$\Rightarrow \quad CX\beta = 0 \quad \text{for any } \beta.$$

$$\Rightarrow \quad CX = 0$$

This gives

$$Var(\tilde{\beta}|X) = Var(\tilde{\beta} - \beta|X)$$

$$= Var([C + (X^TX)^{-1}X^T]e|X)$$

This is because $\tilde{\beta} = \hat{\beta} + \underbrace{CX\beta}_{0} + Ce = \hat{\beta} + Ce$

and $\tilde{\beta} - \beta = Ce + \hat{\beta} - \beta$

$$= Ce + (X^TX)^{-1}X^Te = [C + (X^TX)^{-1}X^T]e$$

For matrix B only depending on X,

$Var(Be|X) = B\,Var(e|X)B^T$

$$= [C + (X^TX)^{-1}X^T]\underbrace{Var(e|X)}_{\sigma^2 I_n}(C + (X^TX)^{-1}X^T)^T$$

$$= \sigma^2 \left[C + (X^T X)^{-1} X^T \right] \left[C^T + X (X^T X)^{-1} \right]$$

$$= \sigma^2 \Big[CC^T + \underbrace{CX}_{=0} (X^T X)^{-1} + (X^T X)^{-1} \underbrace{X^T C^T}_{=0} + (X^T X)^{-1} \underbrace{X^T X}\, \underbrace{(X^T X)^{-1}} \Big]$$

$$= \sigma^2 \left[CC^T + (X^T X)^{-1} \right].$$

$$\Rightarrow \quad Var(\tilde{\beta} \mid X) = \sigma^2 CC^T + \underbrace{\sigma^2 (X^T X)^{-1}}_{Var(\hat{\beta}_{OLS} \mid X)}$$

Since $\underbrace{CC^T}_{(p+1)\times(p+1)} \geq 0$ (that is, for any $a \in \mathbb{R}^{p+1}$, $aCC^Ta^T \geq 0$),

we have

$$Var(\tilde{\beta} \mid X) \geq Var(\hat{\beta}_{OLS} \mid X). \qquad \#.$$

Next lecture: Regression with categorical predictors. (§5. Applied linear Reg).

$\quad$ (factors)