Week 11

Multiple linear regression $[\hat{Y} = \hat\beta_0 + \hat\beta_1 x_1 + \cdots + \hat\beta_p x_p] + \varepsilon$


$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}_{E(Y \mid X_1 \ldots X_p)} + \varepsilon$$

Multicollinearity — Scatter plot

— VIF

Among X

| $X_1$ | $X_2$ ... | $\boxed{X_9}$ | $X_{10}$ |
|---|---|---|---|
| 1.2 | 2.3 ... | $\boxed{100}$ | 97 |

Model Selection

x ↔ Y

- Automatic selection
  (Forward / Backward / Stepwise)
  p-value
- Best subset approach
  (penalized goodness-of-fit)
  small p ← adj AIC → large p

NOTE

- Should we handle multicollinearity necessarily?

stepwise

should it be before the model selection?

Yes / no

- Highly correlated x's in the final model does NOT harm prediction

$\hat{Y}$

- Final model from [best subset approach]

$Y \sim X_1 + X_3 + X_5$

pval 0.07 0.03 —

0.07

p-value — $\hat{n}$

- Automatic selection & Best subset approach

- $\left(R^2 = \dfrac{SS_{model}}{SS_{Total}}\right)$ 79.2%

$\hat{Y}$ = 1.9.

$\hat{Y} = 0 + -$

Y: continuous  X: conti / categorical

(eg) professor data : Salary ~ publication + Gender

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * pub + \hat{\beta}_2 * Gender$$

1

Gender < Female 0 : reference group; Male 1 : comparison group

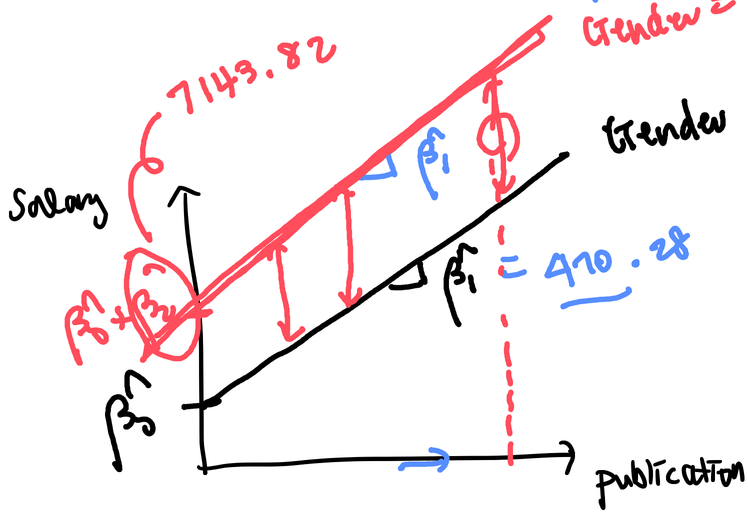
Gender = Female (Gender=0)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * pub$$

Gender = Male (Gender=1)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * pub + \hat{\beta}_2 \quad >0$$

$$= \underbrace{(\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 * pub}_{\text{Gender = Male}}$$



With interaction

Salary ~ pub + Gender + pub * Gender

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * pub + \hat{\beta}_2 * Gender + \hat{\beta}_{12}(pub \cdot Gender)$$

$\hat{\beta}_0 = 47680$ $\quad \hat{\beta}_1 = 102.1$ $\quad \hat{\beta}_2 = -1998.6$ $\quad \hat{\beta}_{12} = 535.9$

Gender = Female
Gender = 0

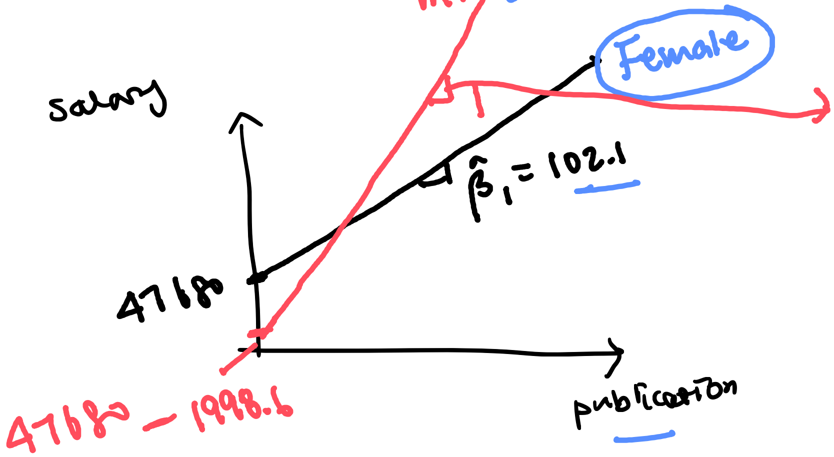
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * pub$$

Gender = Male
Gender = 1

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * pub + \hat{\beta}_2 + \hat{\beta}_{12} * pub$$

$$= (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_{12}) * pub$$

male




102.1 + 535.9
= 637

R output interpretation

- check significance of predictors

- $H_0: \beta_{Gender} = 0$   $H_a: \beta_{Gender} \neq 0$
- $H_0: \beta_{Gender \times pub} = 0$   $H_a: \beta_{Gender \times pub} \neq 0$

ANOVA   Toothlength ~ Dose

Dose: 0.5 → reference, 1, 2

| | |
|---|---|
| intercept | 10.60 |
| Dose1 | 9.13 |
| Dose2 | 15.49 |

ANCOVA

con

15.49   10.6 + 15.49   Dose=2

10.6 + 9.13   Dose=1

9.13

10.6   Dose=0.5

$\hat{y}$   $\hat{\beta}$   Dose

Dose=0.5   $\hat{y} = 10.6$

Dose=1   $\hat{y} = 10.6 + 9.13$

Dose=2   $\hat{y} = 10.6 + 15.49$

case1

| | pval |
|---|---|
| Dose1 | 0.3 |
| Dose 2 | 0.5 |

in Dose

case2

| | p-va |
|---|---|
| Dose1 | 0.005 |
| Dose 2 | 0.001 |

0.5 1

case3

| | pval |
|---|---|
| Dose1 | 0.1 |
| Dose 2 | 0.0005 |

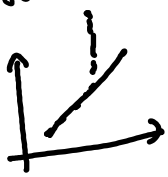
2

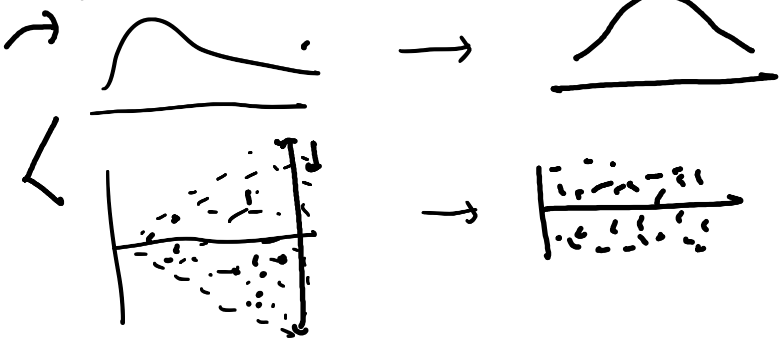
0.5 ↔ 1

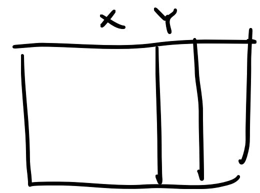
0.5 ↔ 2

Normality $\sqrt{Y}$ $\log Y$

Homoscedasticity





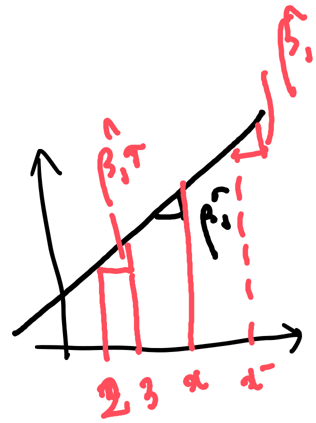



$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$\log Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$y^{4/3}$$

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X_1$$




$\hat{\beta}_1$

Additive change

$X=2 \xrightarrow{+1} X=3$

$X=6 \xrightarrow{+1} X=7$

$\hat{\beta}_1$

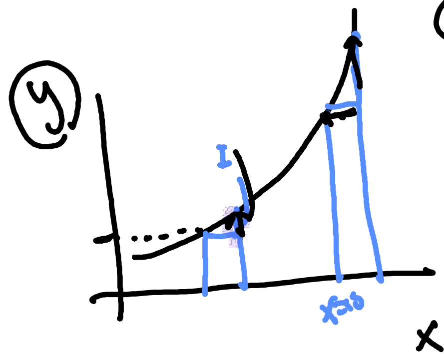
$y$

$X=2 \rightarrow X=3$

$X=6 \rightarrow X=7$

$$\log \hat{Y} = \hat{\beta_0} + \hat{\beta}_1 X_1$$

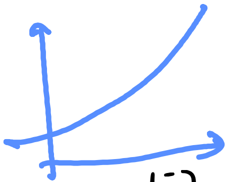
$Y$



$$\left[\log \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x\right]$$

One unit increase in $x$, what happens on $Y$?

$y = e^x$



(i) $X$ is continuous

$x$: $\hat{y}$, $\log(\hat{Y} \mid X = x) = \hat{\beta}_0 + \hat{\beta}_1 x \Rightarrow \hat{Y}(x) = e^{\hat{\beta}_0 + \hat{\beta}_1 x}$ (exp)

$x+1$: $\hat{y} \mid x+1$, $\log(\hat{Y} \mid X = x+1) = \hat{\beta}_0 + \hat{\beta}_1 (x+1) \Rightarrow \hat{Y}(x+1) = e^{\hat{\beta}_0 + \hat{\beta}_1 (x+1)}$ (exp)

$$\frac{\hat{Y}(x+1)}{\hat{Y}(x)} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 (x+1)}}{e^{(\hat{\beta}_0 + \hat{\beta}_1 x)}} = \frac{\cancel{e^{\hat{\beta}_0}} \cdot \cancel{e^{\hat{\beta}_1 x}} \cdot e^{\hat{\beta}_1}}{\cancel{e^{\hat{\beta}_0}} \cdot \cancel{e^{\hat{\beta}_1 x}}} = e^{\hat{\beta}_1}$$

$$\hat{Y}(x+1) = e^{\hat{\beta}_1} \cdot \hat{Y}(x)$$

1.34

$\beta$

$\hat{\beta}_1 = 0.3$: $\dfrac{\hat{Y}(x+1)}{\hat{Y}(x)} = e^{0.3} = 1.34$ ($= 1$) 34% ↑

$\hat{\beta}_1 = -0.3$: $\dfrac{\hat{Y}(x+1)}{\hat{Y}(x)} = e^{-0.3} = 0.74$ 26% ↓ $< 1$

2.34 $\hat{\beta}_1 = 0$

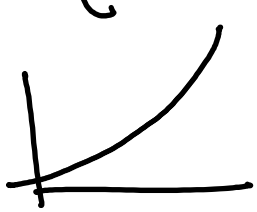
$\hat{\beta}_1 > 0$

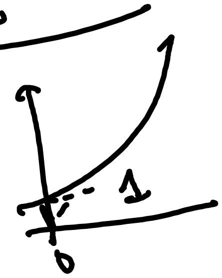
$\hat{\beta}_1 < 0$

$\hat{\beta}_1 = 0$ $e^0 = 1$

$e^{\hat{\beta}_1} < 1$

$e^{\hat{\beta}_1} > 1$






$e^{0.09} = 1.09$

$e^{-0.02} = 0.97 < 1$ 3% ↓

$\log Y$ $\log X$

X : categorical

$$X \begin{cases} \text{Female} & 0 \\ \text{Male} & 1 \end{cases}$$

$$\underbrace{\log \hat{Y} = \hat{\beta_0} + \hat{\beta_1} X}$$ Gender

Gender=0

exp $\qquad \log(\hat{Y} \mid \text{Female}) = (\hat{\beta_0}) \Rightarrow \hat{Y}(\text{Female}) = e^{\hat{\beta_0}}$ exp

exp $\qquad \log(\hat{Y} \mid \text{Male}) = (\hat{\beta_0} + \hat{\beta_1}) \Rightarrow \hat{Y}(\text{Male}) = e^{\hat{\beta_0}+\hat{\beta_1}}$

Gender=1 exp

$$\frac{\hat{Y}(\text{male})}{\hat{Y}(\text{Female})} = \frac{\cancel{e^{\hat{\beta_0}}} \cdot e^{\hat{\beta_1}}}{\cancel{e^{\hat{\beta_0}}}} = e^{\hat{\beta_1}} = e^{0.75} = 2.117$$

$= 111\% \uparrow$

BMI = 10

LBu = 5