Week 10.

$Y \sim X_1$

$Y \sim X_1 + X_2 + \cdots + X_p$

goal: ① Find linear relationship btw Y, X

② prediction $\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$

$Y \sim X_1$ (simple)

$Y \sim X_1 + X_2 + \cdots + X_p$ (multiple)

- Model significance

$H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ ← model is not useful

H_a: at least one $\beta \neq 0$

- Individual term significance $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

$H_0: \beta_M = 0$ vs. $H_a: \beta_M \neq 0$

$H_0: \beta_2 = 0$ vs. $H_a: \beta_2 \neq 0$

$\vdots$

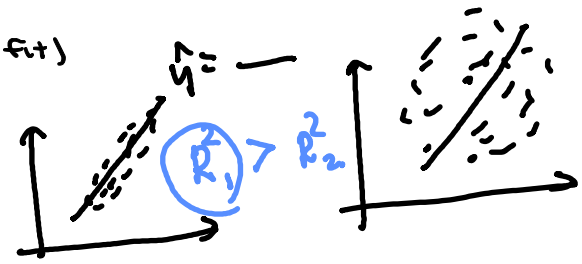
$H_0: \beta_p = 0$ vs. $H_a: \beta_p \neq 0$

- Estimated regression line

$\hat{y} = -b_0 + b_4 * M + 35.2 * S_0 + \cdots$

- R^2 (prediction power / goodness-of-fit)

$= \left(\frac{SS_{model}}{SS_{total}} \right)$



- Model diagnostics

model assumption

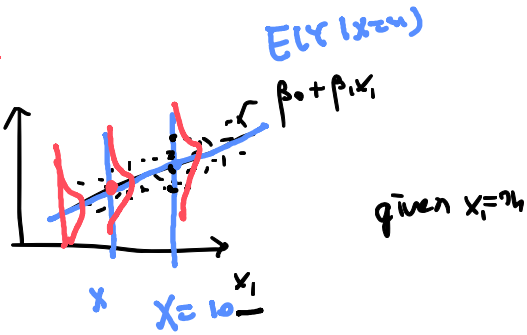
influential count D.

$Y = \beta_0 + \beta_1 X_1 + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$

mean, Error

$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$

mean, $N(0, \sigma^2)$



given $X_1 = x_1$, $X_2 = x_2$, $\vdots$, $X_p = x_p$

$Y \sim N($

Multicollinearity

⇒ Remove highly cor variables

Y $(X_1 \cdots X_{10})$ $n = 100$

X_1 ⋮ X_{10} $\hat{\beta}_1$ ⋮ $\hat{\beta}_{10}$

Model selection

"optimal" or "best"

Multicollinearity (eg) UScrime data

About predictors not model assumptions or Y

simulation studies

cirrhosis

alcohol

alcohol2 = 3*alcohol

alcohol3 = 3*alcohol + ε

① cirrhosis ~ alcohol

② cirrhosis ~ alcohol + alcohol2 cor(alcoh. alc2) = 1

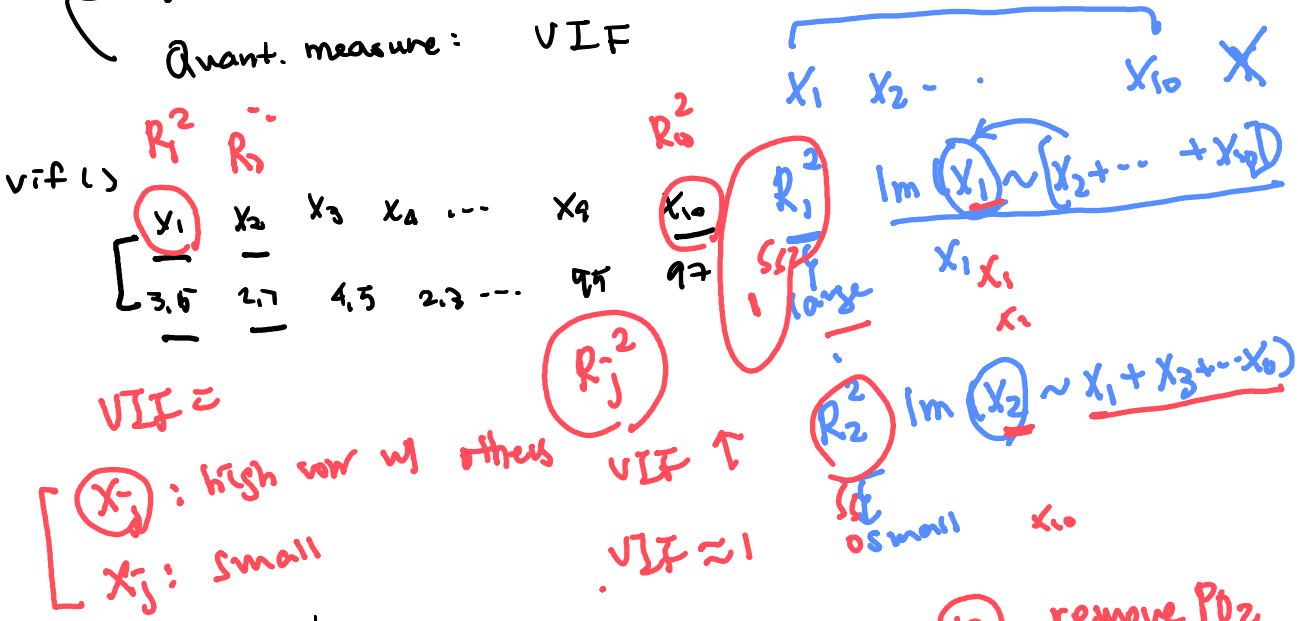
③ cirrhosis ~ alcohol + alcohol3 cor(alc. alc3) = 0.99

F-test : model is useful

t-test

How to detect? what is the remedy?

- Visualization : pairwise scatter plot
- Quant. measure : VIF

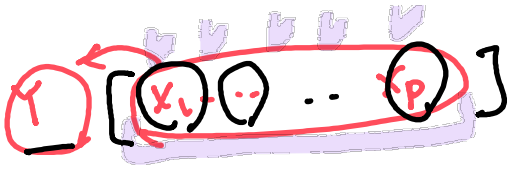


⇒ Remove one by one

① Remove predictor w/ highest VIF (w/ threshold) 10 remove PD2

② $Y \sim . - PD2$

③



Model Selection

10 [X_1 - X_{10}] 2^{10} $Y \sim x_1$ $Y \sim x_1 + x_2$

- Automatic Selection : backward / forward / stepwise ⇒ criterion ⇒ p-value ⇒ BIC
- Best subset approach : goodness-of-fit measure criteria {adjusted R^2 / AIC / BIC / Mallow's C_p}

90

$p = 100$ $R^2 = 88\%$ ↑ 3%

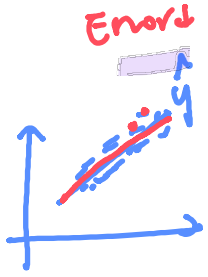
$p = 10$ $R^2 = 85\%$

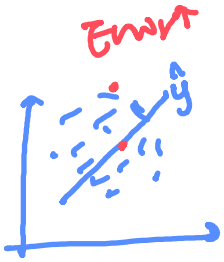
$p = 2$ $R^2 = 50\%$

good prediction ⟷ simplicity

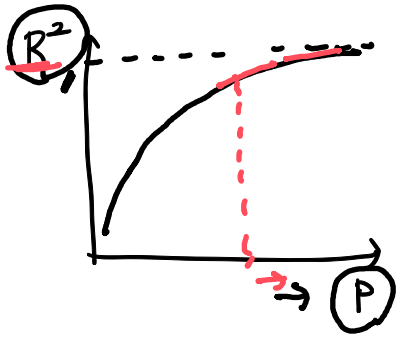
Error $\sum (y_i - \hat{y}_i)^2$ small

penalized goodness-of-fit



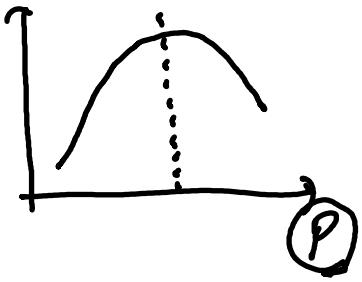


☒ adj-R^2

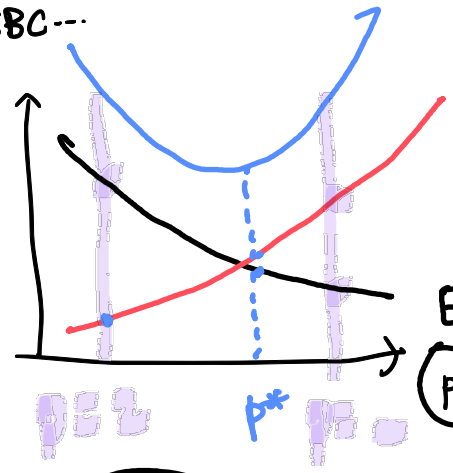


100 ∵ $p = n$ 100 $R^2 = 1$

adj-R^2



AIC, BIC, SBC…



$= -2 \log L(M) + \text{penalty}$

$\sum (\hat{y}_i - y_i)^2$

$C_p \leq p$

model 1 $C_p : 5$ $p = 3$

model 2 $C_p : 5.1$ $p = 6$

Data Example

- adj-R^2 → model 8 → M Ed Pol M.F (V1) V2 Intg Prob
- AIC → model 8
- BIC → model 6 → M Ed Pol — — V2 Intg Prob.
- Cp → model 6

↓ stepwise

Multicollinearity Issue in model selection

- ① Automatic step selection (p-value criteria)
- ② goodness-of-fit measure search