Week 9

Linear Regression

- $Y$ : continuous response — Final
- $X$ : continuous or categorical predictors — # of hours

Goal

(i) identify linear relationship btw $Y$ . $X$ — Inference

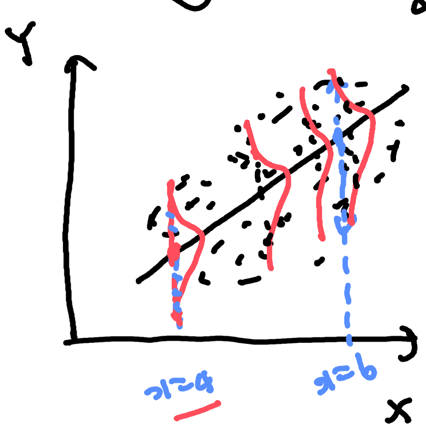
if significant → quantify

$H_0$: $X$ . $Y$ no linear rela

$H_a$: $X$ - $Y$ linear relation

(ii) prediction — $X$ → $Y$ — $\tilde{Y}$

$Y$ - $\tilde{Y}$




simple linear regression

$E(Y) = \boxed{\beta_0 + \beta_1 X}$

$X_2$

$Y = \beta_0 + \beta_1 X + \varepsilon$ — mean, Intercept — Error

$\varepsilon \sim N(0, \sigma^2)$

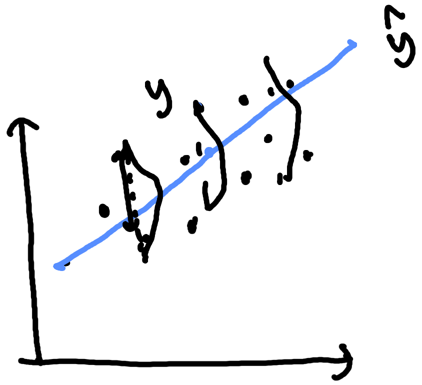
⇒ $E(Y) =$ ↘ $\hat{\beta}_0$ $\hat{\beta}_1$

(1) Linear relationship btw $X$ . $Y$

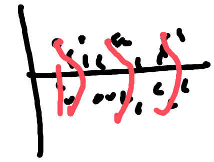
(2) conditional Normality of $Y$ given $X = x$

(3) Equal variance

(4) indep samples




LSE (OLS) $y - \hat{y}$



$\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$

$\beta_1 \ \beta_2 \cdots \ \beta_p$

p

simple: $\boxed{Y} = \beta_0 + \beta_1 X_1 + \varepsilon$ (→ 0)

multiple: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$

(1) Model significance

- $H_0$: model is not useful
- $H_a$: moden is useful

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

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$
$H_a$: at least one $\beta \neq 0$

(2) Individual term significance

- $H_0$: $\beta_{alcohol} = 0$
- $H_a$: $\beta_{alcohol} \neq 0$

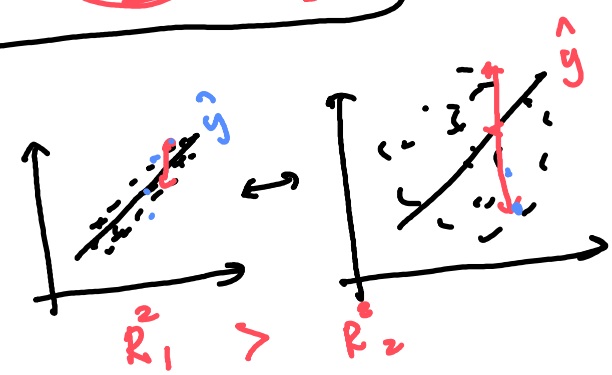
$H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$
$\vdots$
$H_0: \beta_p = 0$ vs $H_a: \beta_p \neq 0$

(3) Estimated regression line $\hat{\beta_0}$ $\hat{\beta_1}$

$* \ \hat{Y} = -5.99 + 1.919 * \text{alcohol}$

prediction power

$\hat{y}$ $y$ (4) $R^2 = \dfrac{SS_{model}}{SS_{Total}}$



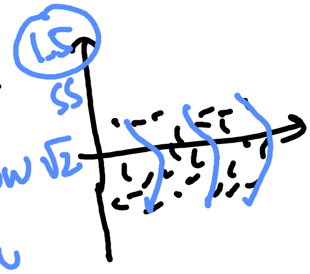

(5) Model diagnostics

Diagnostics

① Model assumption

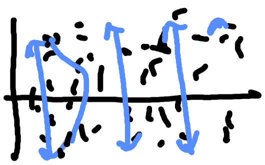
- Normality — Normal Q-Q / $\sqrt{\text{std re}}$ below $\sqrt{2}$
- Homoscedasticity — residual
- linear relationship — residual

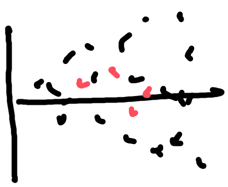
② Influential points check

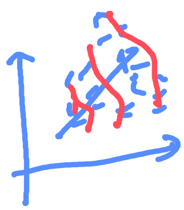



$Y - \hat{Y}$

= residual







$Y_1$ $Y_2$ $Y_3$ $Y_4$

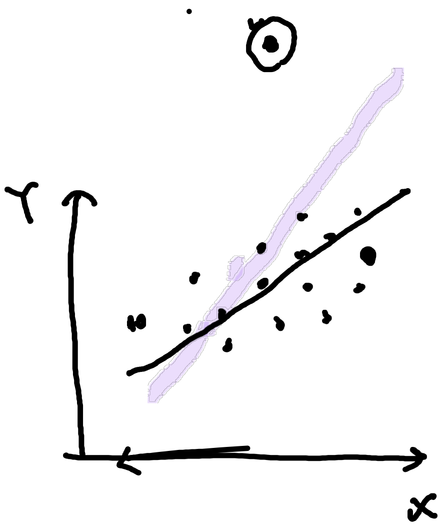
Simulation studies

| | linearity x-Y | Normality | Homoscedasticity | |
|---|---|---|---|---|
| 1 | O | O | O | lm($Y_1$ ~ x) |
| 2 | O | X | O | lm($Y_2$ ~ x) |
| 3 | O | O | X | lm($Y_3$ ~ x) |
| 4 | X | O | O | lm($Y_4$ ~ x) |

Influential points

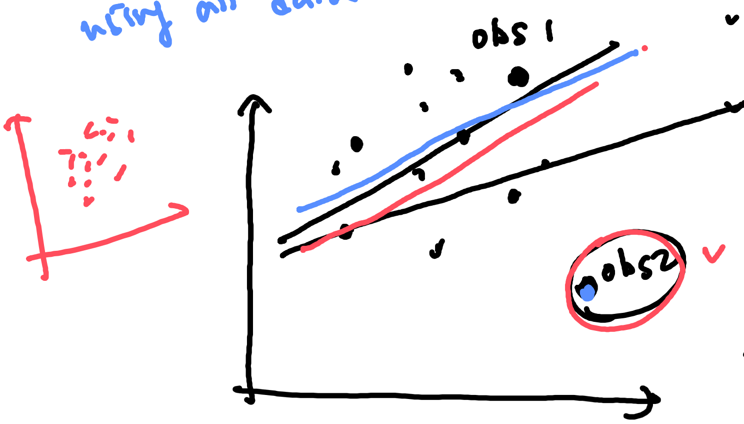
⇒ observations that greatly affect the slope of regression line.

- outliers : abnormal behv in y
- leverage points : " in x



Cook's distance : influence of individual data points on model fitting

$$\underset{\text{fitted value } \hat{y} \text{ using all data}}{\overset{100}{\hat{Y}}} - \underset{\text{fitted } \hat{y} \text{ without ith}}{\overset{99}{\hat{Y}_{(i)}}}$$



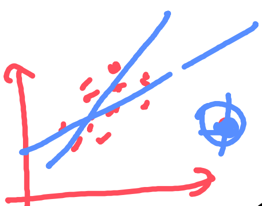

obs 1 : $\hat{Y} - \hat{Y}_{(1)}$

obs 2 : $\hat{Y} - \hat{Y}_{(2)}$

⋮

obs 100 .

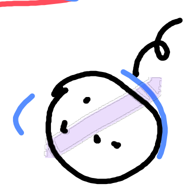
Cook's D



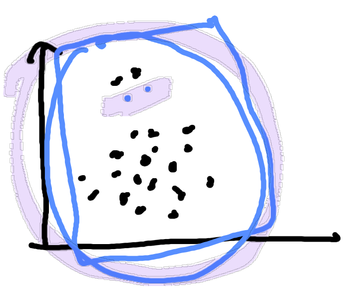
large Cook's D : influential points

small Cook's D :

$\frac{4}{n}$



NOTE



(Q) several influential obs removed.. all done?

- drinking data example

using all data

$\hat{y} = -5.99 + 1.97 * \text{alcohol}$

without France

$\hat{y} = -3.61 + 1.64 * \text{alcohol}$

- crime.csv data

$\hat{y} = -81.20 + 49.03 * \text{poverty}$ 25%

w/o 51st dc $\hat{y} = 209 + 25.45 * \text{po}$ 13.6%

SStotal

SStotal

S

40 + 50 = 90

50% 50%

35 → 15% +6 → 41

25 → 25% * 40% → +10

47