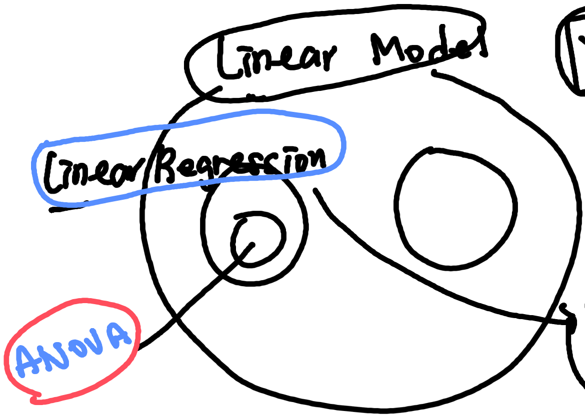
Week 8 {Linear Regression}

Y ~ X

→ parametric modeling



Y : conti num

X : cate / conti

Y : conti numerical

X : categorical → $H_0: \mu_A = \mu_B = \mu_C$

→ H_a:

Goal: Find a relationship between X · Y

& quantify its relationship

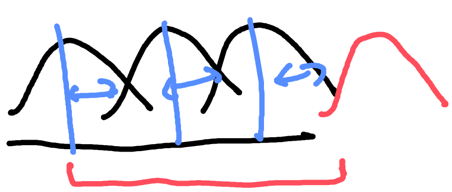
* prediction

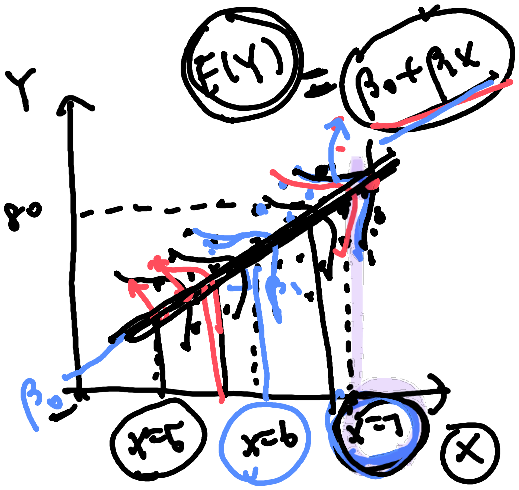
1-way ANOVA

1) Y : conti. X : categorical
2) Normality for each level
3) Equal var
4) indep samples

Linear Regression

1) X, Y linear relationship
2) Normality ← conditioning on X
3) Equal var
4) indep samples





Y: Final score

X: # of hours study

(Assumption 1)

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

$\beta_0 + \beta_1 x$: mean of Y for if $X = x$ → explained by x (model)

ε: Error, individual → can't be expln

① Using Samples, β_0 β_1

goal - Estimate β_0, β_1

Estima: $\hat{\beta}_0$, $\hat{\beta}_1$

1.5

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Estimated reg line, prediction

$\hat{y} = 70.$

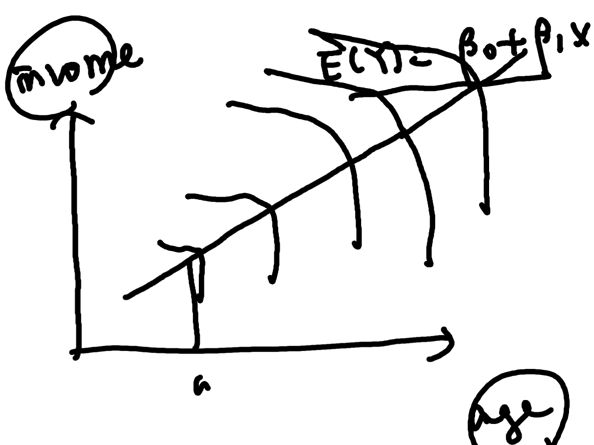
[$H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$]

X, Y no relationship

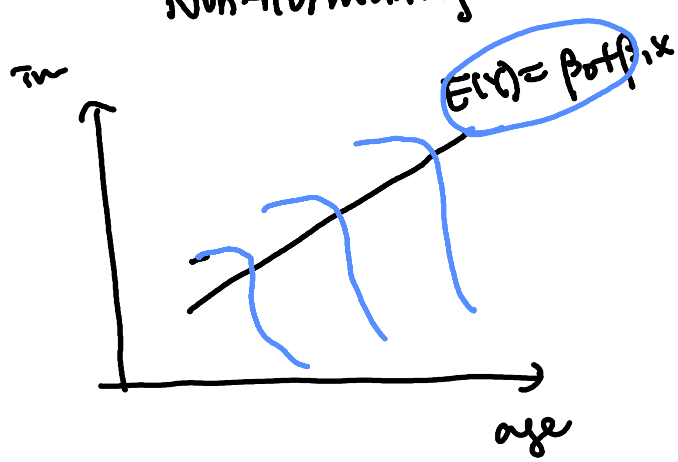
sig. relat

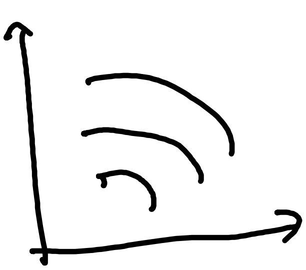
: inferential problem

Heteroscedasticity

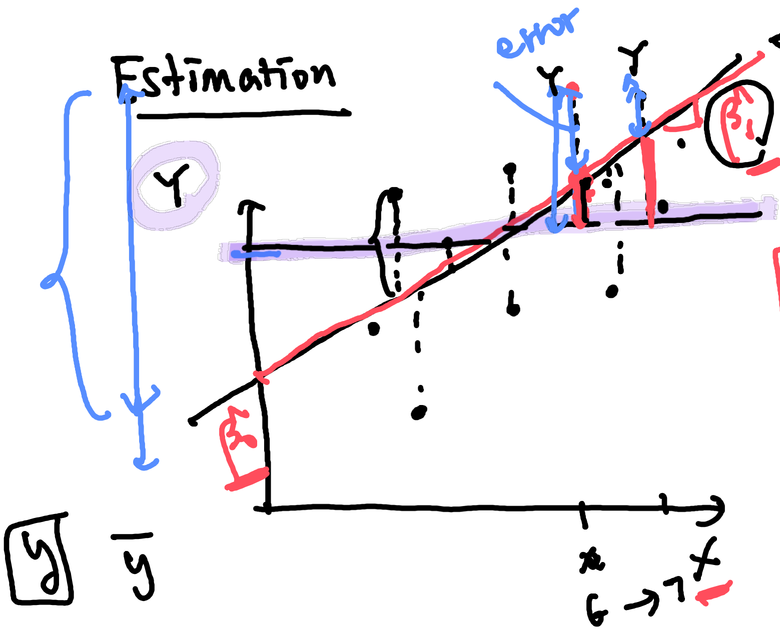


Non-normality





Estimation



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$x \to x+1$

$\hat{\beta}_1$

Least Square Estimation (LSE)

or Ordinary Least Square (OLS)

⇓

minimize squared vertical distance

Y: cirrhosis

X: alcohol

$Y \sim X_1, X_2 \cdots X_5$

β_1 β_2 β_5

- write the estimated regression line

$$\hat{y} = \underbrace{-5.99}_{\hat{\beta}_0} + 1.9779 * \text{alcohol}$$

① Test model significance (F-test)

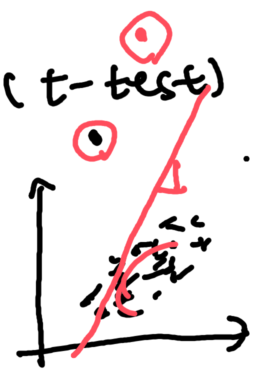
$H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$ H_0: linear model is not useful

$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

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

② Test individual X significance (t-test)

$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$



③ Interpretation of $\hat{\beta}_1$

④ $$R^2 = \frac{SS_{model}}{SS_{Total}}$$

$$SS_{Total} = SS_{Model} + SS_{Error}$$

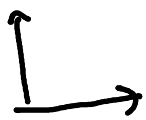
⑤ Model diagnostics ← model assumption

(0)

influential point diagnosis

Exploratory Analysis

< Scatter plot
correlation (ρ) : pearson correlation



X, Y

(1) range. sign + − $|\rho|$

$-1 \le \rho \le 1$

(2) Only measure linearity

$\boxed{\rho}$

$\rho \approx 0$

→ $\rho \approx 0$

(3) Sensitivity to outliers $\rho \approx 0$ → linear relat

sensitive

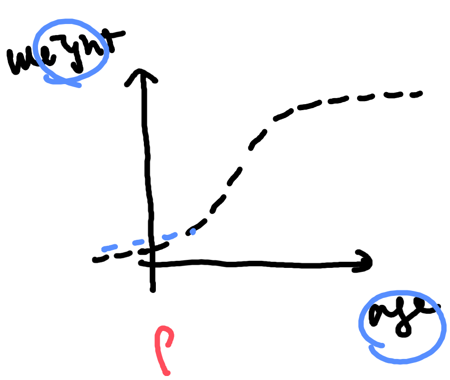
ρ

$$\rho = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2}\sqrt{\sum (y_i-\bar{y})^2}}$$

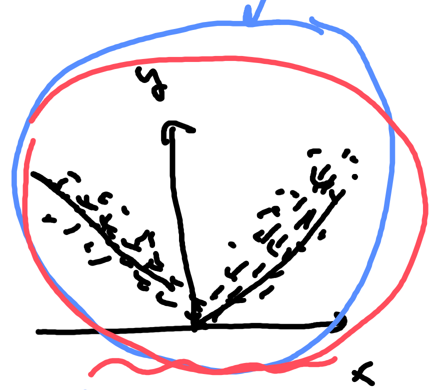
[pearson
spearman corre
robust

correlation

$\rho \approx 0.6$



ρ



$\rho \approx 0$

$\frac{SS_{mol}}{SS_{Total}}$ $SS_{Total} = SS_m + SS_{Err}$

R² and Model significance

→ [$\beta = 0$ or $\beta \neq 0$]

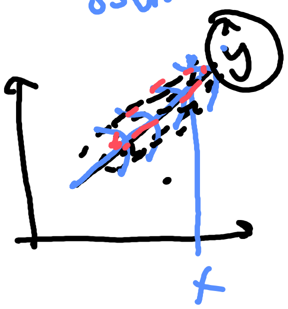
p-value

Goodness-of-fit measure

prediction power

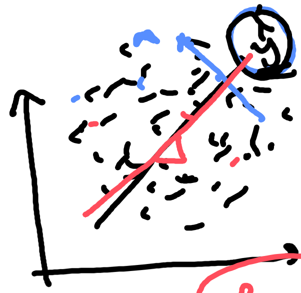
SS Error small

$R^2 \uparrow$



SS Err large

$R^2 \downarrow$



$R^2 = 0.1$

(eg) (Smoking ~ lung-cancer)

(model $R^2 = 0.3$

mod $R^2 = 0.8$

$R^2 = 0.99$

Causal effect ↔ linear relationship

(# of people drown at lake

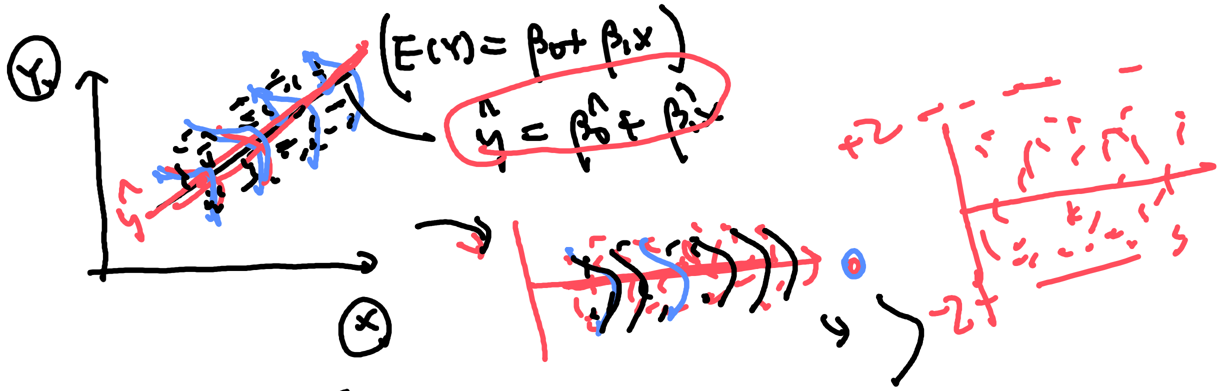
temper

($H_0: \beta = 0$ vs. $H_a: \beta \neq 0$)

year | Y

(Ice cream sales

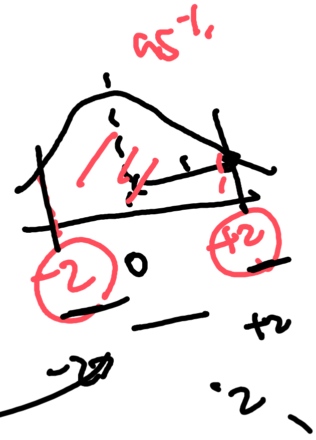
Diagnostics
- (1) Model assumption check
  - linear X. Y
  - Normality
  - Equa
- (2) Influence point check



Residual $= Y - \hat{Y}$

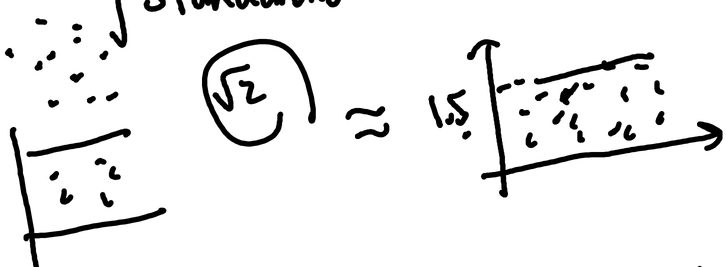
Expect to see
- Normal
- Equal var.

[standardized residual] → standar
0 . 1
$N(0,1)$



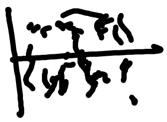
$\sqrt{\text{standardized residual}}$

$\sqrt{2} \approx 1.5$

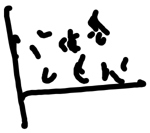


Normality check by
- QQ plot
- square-root standardized residual (below 1.5)

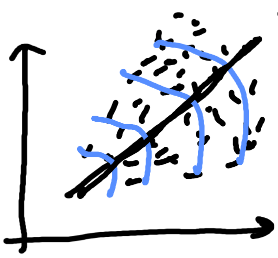
→ Equal variance check by



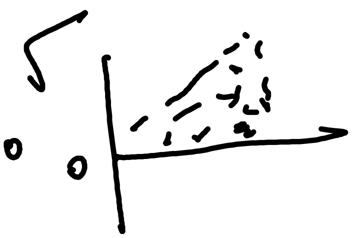
(Residual
$\sqrt{\text{standardized residual}}$



$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$



$\sqrt{\text{Stan Ro}}$



$y - \hat{y}$