Anova : ~~R~~esiduals $\sim N(0; \sigma^2)$.

$\rightarrow$ S. residuals $\sim N(0, 1)$.

Anova is done $\rightarrow$ can get S.R $\rightarrow$ check if $SR \sim N(0,1)$

Fit model M $\rightarrow$ get $\rightarrow$ check assumption.

$\rightarrow f(x) = 2x - x^2 + 3$. ; $x \in R$.

$\Rightarrow$ how to find $\min f(x)$?

$\Rightarrow \underline{f'(x)} = \boxed{2 - 2x = 0} \Rightarrow \boxed{x = 1}$ is a stationary point.

$g(x, y) = x^2 - y^2 + \dots$, $x \in [\quad]$

$y \in [\quad]$

$\Rightarrow \min g(x, y)$

$\Rightarrow \begin{cases} g'(x, y)\big|_x = 0 \\ g'(x, y)\big|_y = 0 \end{cases} \Rightarrow$ solve both to get the stationary point.

---

degree of freedom in modelling :

response : y $\rightarrow$ no regressor $\rightarrow$ then model has $df = 0$.

$\rightarrow$ has 1 regressor : $\underline{y \sim x} \Rightarrow$ model has $\underline{df = 1}$

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

$\rightarrow$ has k regressors : $y \sim \underset{\text{age}}{x_1} + \underset{\text{height}}{x_2} + \dots + \underset{\text{income}}{x_k}$

$\hat{y} = \hat{\beta_0} + \hat{\beta_1} * x + \dots + \hat{\beta_k} * x_k$

then model has $df = k$.

a parameter is estimated by an estimator → need to check
How good the estimator is ?

- estimator is unbiased ?
  yes → good.
- variance / SD of estimator ?
  small → good.

example

population mean : μ → estimate μ by estimator $\boxed{\bar{X}}$

$E(\bar{X}) = \mu$ → $\bar{X}$ is unbiased estimator.

$\bar{X} \overset{CLT}{\sim} N(\mu; \sigma^2/n)$ $\quad Var(\bar{X}) = \sigma^2/n$ ←

in modeling : parameter : $\beta_0 \to \hat{\beta_0}$
$\beta_1 \to \hat{\beta_1}$

→ these estimators good or not?
⇒ mean & var. ?
distn of ?

W ~ H :

Sample, size 1000 ⇒ $\hat{W} = 20 + 25 * H$ ⇒ $\hat{\beta_1} = 25$

2^{nd} sample, size 1000 ⇒ $\hat{W} = 19 + 26 * H$ ⇒ $\hat{\beta_1} = 26$

⋮

N^{th} sample, size 1000 ⇒ $\hat{W} = 21 + 24.8 * H$ ⇒ $\hat{\beta_1} = 24.8$

⇒ value of the slope / intercept vary from sample to sample.
$\hat{\beta_1} / \hat{\beta_0}$ is variable.

$\frac{SS_{Res}}{n-2}$ → to estimate σ^2. why $(n-2)$?

In simple model : $y \sim x$ ⇒ model has 2 coef including intercept ⇒ $(n-2)$

$W \sim H$, fitted model:

$$\widehat{W} = \underbrace{-592.6}_{\hat{\beta}_0} + \underbrace{11.19}_{\hat{\beta}_1} * H.$$

σ^2 is estimated by MS_{Res}

σ ———————— $\sqrt{MS_{Res}} = 11.86$

F statistic $= 84.85 = F_0$ to test the sig. of model.

null distn $\rightarrow F_{1,5}$, $F_{1,n-2}$ $\Rightarrow$ p-value $= 0.000256$

$n = 7$

$W \sim H + age + income$.

test the sig. of variable income:

$$W = \beta_0 + \beta_1 * H + \beta_2 * A + \beta_3 * Income + \varepsilon.$$

$H_0: \beta_3 = 0$ vs $H_1: \beta_3 \neq 0$.

Income is NOT significant

it is significant.

---

simple model: $W \sim H$: $W = \beta_0 + \beta_1 * H + \varepsilon$.

test the sig of variable H:

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

$$t_0 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \overset{H_0}{\sim} t_{\boxed{n-2}}$$ bc2 simple model.

in R: $SE(\hat{\beta}_1) = 1.218$;

$t_0 = 9.19 \overset{H_0}{\sim} t_5$ bc2 $n = 7$

$\rightarrow$ p-value $= 0.000256$.

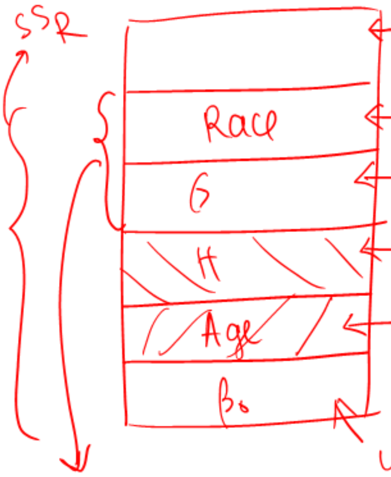
p-value of t test & F test in SIMPLE model are the same.

$* \; F_{1,df} \approx (t_{df})^2 \Leftarrow$

$W \sim A + \boxed{H} + G$

t_test for H : A and G are included in the model.

$W \sim H \Leftarrow$



Race when β_0, Age, H, G are added.

Gender when β_0, Age, H are added.

contribution of H when β_0, Age are added.

contribution of Age when β_0 is added $\Rightarrow SS_R(\beta_1 | \beta_0) \rightarrow$?

contribution of β_0 to the model.

contribution of Race & G when β_0, Age, H are already added: $SS_R(\beta_3, \beta_4 | \beta_0, \beta_1, \beta_2)$

SS_R

---

assumptions of a model:

- linearity : $y \sim x \Leftarrow$
- normality : $\varepsilon \sim N$
- constant var: ε has constant var: σ^2

$\Big\}\; \varepsilon \sim N(0; \sigma^2)$

- error are uncorrelated. $\Rightarrow$

⊗ the regressor $(x_i, x_j \dots)$ are uncorrelated.

$W \sim H + A + G$ || collinearity

ST3131

W H Model: $\underbrace{\hat{W} = 20 + 25 * H}$ $\hat{W}$ e

$1 \to 48 \quad 1.6 \longrightarrow \underbrace{20 + 25 * 1.6} \leftarrow \hat{W}_1$ | $\underline{48 - \hat{W}_1 = e_1}$

$2 \to 65 \quad 1.72 \longrightarrow \underbrace{20 + 25 * 1.72} \leftarrow \hat{W}_2 \to 65 - \hat{W}_2 = e_2$

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

$\Downarrow$

represented by $\underbrace{e_1, \dots, e_n} \sim N(0; \sigma^2)$

$\Downarrow$

$\boxed{\text{standardized} \to \sim N(0, 1)}$

$SR \sim N(0,1) \to$ what we expect.

---

build a model $\to$ get SR $\to$ check if $\underline{SR \sim N(0,1)}$.

---

a model: $\boxed{\text{adaquate}}$? $\qquad$ $\varepsilon \sim N(0, \sigma^2)$

good fitting? $\leftarrow R^2 = 0.9$. $\qquad$ $\{e_i\} \sim N(0, \sigma^2)$ $\qquad$ $r_i \sim N(0,1)$

Adequate = model satisfies the assumptions
- linearity
- normality
- constant var
- error are uncorrelated
- regressors are uncorrelated

$W \sim \underline{H + A + G}$

X_i are correlated $\to$ $\boxed{\text{collinearity}}$ ST3131

$\Leftarrow$ more $SR > 0 \Leftrightarrow$ more $e_i > 0$

$\Leftrightarrow$ more $y_i - \hat{y}_i > 0$

$\Rightarrow$ model $\to$ underestimating.

$48 \to 44$

$46 \to 45$

$70 \to 60$.

$\boxed{\sum e_i = 0}$

---

band of SR: $-3 \to 3$ or $-2 \to 2$ | $SR \sim N(0,1)$: 95% of SR: $-2, 2$. $\Leftarrow$

99.9% of SR: $-3, 3$. $\leftarrow$

$n = 7$; $n = 100 \to$ can accept 4-5 points: $|SR| > 2$.

1-2 points: $|SR| > 3$

$W \sim \underset{\text{quant}}{\underline{\text{width}}} + \underset{\text{3 cate}}{\underline{\text{spine}}} \rightarrow S_1 \,\&\, S_2 \rightarrow$ spine = 2.

$S_1 \downarrow$ cate spine = 1

width → t_test → sig ?

S_1 → t_test → sig ?

S_2 →

$\alpha = 0.05$

S_1 is sig but S_2 is not.

⇒ can we drop S_2, but keep S_1.

$W \sim$ width + S_1

$S_1 \downarrow$ spine = 1

⇒ if so: spine = 2 categories:

1 or (2 & 3)

1, 2, 3

→ ~~spine~~ $S_1 = \begin{cases} 1 & \text{if spine} = 1 \\ 0 & \text{o.w.} \end{cases}$

8 cates → $\underbrace{1 \rightarrow 4}_{\text{low}}$ $\underbrace{5 \rightarrow 6}_{\text{med}}$ $\underbrace{7\text{-}8}_{\text{hi}}$

→