- by dividing SSR + SSE by their df we obtain mean sums for regression called (MSR) + for error (MSE)

ie. $MSR = \frac{SSR}{df/SSR}$ $\quad MSR = \frac{SSR}{1} = SSR$ (test)

$MSE = \frac{SSE}{df/SSE}$ $\quad MSE = \frac{SSE}{n-2}$ $\left(= S^2\right)$ (t-test)

$t = \frac{\hat{\beta}_1}{\sqrt{\frac{S^2}{S_{xx}}}}$

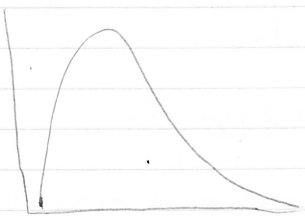
identical q's

i.e. since S^2 is an unbiased estimator of σ^2 (ie. $E(S^2) = \sigma^2$) and since $S^2 = MSE \Rightarrow MSE$ is an unbiased estimator of σ^2 (ie. $E(MSE) = \sigma^2$)

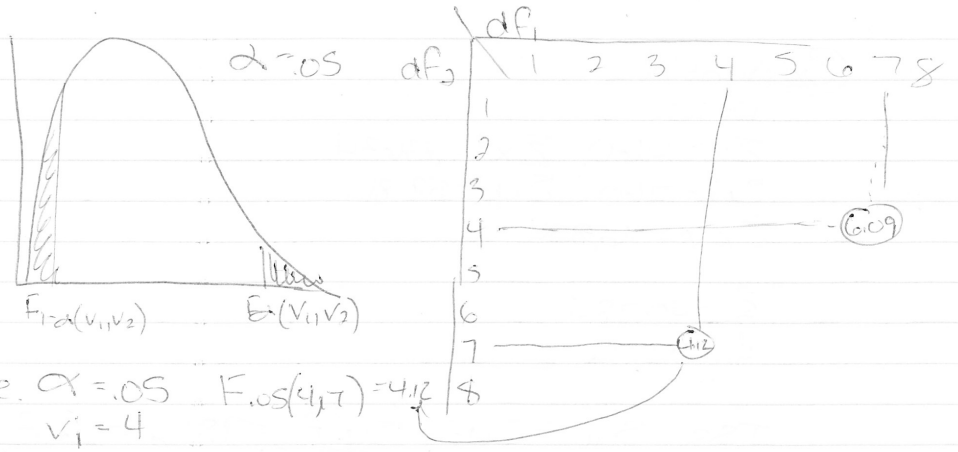
- it can be shown that under $H_0: \beta_1 = 0$ MSR is also an unbiased estimator of σ^2 where expected value of $MSR = \sigma^2 + \hat{\beta}_1^2 (\sum x_i - \bar{x})^2$

$\Rightarrow$ test-statistic $\quad F = \frac{MSR}{MSE}$ $\quad$ R.R. Reject H_0 if $F > F_{\alpha(1,n-2)}$ accept H_0 if $F = 1$

※ F-Dist



- skewed to the right (i.e. positively skewed)
- not symmetric + it (only has) takes pos. values
- has 2 different df (V_1 = numerator df and V_2 = denominator df)



ie. $\alpha = .05$ $F_{.05}(4,7) = 4.12$

$v_1 = 4$

$v_2 = 7$

put in this order $\frac{1}{F_{\alpha(v_1,v_2)}} \Rightarrow \frac{1}{F_{\alpha(v_2,v_1)}}$

$$F_{.95}(4,7) = \frac{1}{F_{.05}(7,4)} = \frac{1}{6.09} = 0.16$$

ANOVA Table

$TSS = SSR + SSE$

$df_{TSS} = df_{SSR} + df_{SSE}$

source of variation	df	SS	mean sums ms	F
Regression	1	SSR	msR	F
Error	n-2	SSE	msE	
Total	n-1	TSS		

$H_0: \beta_1 = 0$; α

$H_a: \beta_1 \neq 0$

test-stat. $F = \frac{msR}{msE}$

R.R

We reject H_0 if $F > F_{\alpha(1,n-2)}$

Ex cont'd

$n=10$

$\sum x_i = 460 \quad \sum x_i^2 = 23634$

$\sum y_i = 760 \quad \sum y_i^2 = 59,816$

$\sum x_i y_i = 36,854$

$\hat{\beta}_0 = 40.784$

$\hat{\beta}_1 = 0.176$

$$TSS = S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 59816 - \frac{760^2}{10} = 2,056$$

$$SSR = \frac{(S_{xy})^2}{S_{xx}} = \frac{\left[\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right]^2}{S_{xx}\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{\left(36854 - \frac{(460)(760)}{10}\right)^2}{23634 - \frac{(460)^2}{10}} =$$

$$\frac{(1,894)^2}{2,474} = 1,449.9741$$

$SSE = TSS - SSR = 2,056 - 1,449.9741 = 606.02587$

$MSR = \frac{SSR}{1} = \frac{1449.9741}{1}$

$MSE = \frac{SSE}{n-2} = \frac{606.02587}{8} = 75.7532 \quad (\hat{\sigma}^2)$

F-test-stat.

$F = \frac{MSR}{MSE} = 19.14$

Source of Variation	d.f.	SS	MS	F
Regression	1	1449.9741	1449.9741	19.14
Error	$n-2=8$	606.02587	75.7532	
Tot.	$n-1=9$	2056		

always have to add up, even if using rounded values.

$H_0: \beta_1 = 0$ - not linear
$H_a: \beta_1 \neq 0$ - linear $\alpha = .05$

R.R. we reject if
$F > F_\alpha(1, n-2) = F_{.05}(1,8) = 5.32$

test stat.
$F = \frac{MSR}{MSE} = 19.14$

Since $F = 19.14 > 5.32$, we reject H_0 & conclude that @ 5% level of significance there is evidence that math test results & final stats grade are linearly related.

F & t test must give same ~~results answers~~

NOTE: t-test & F-test must give you identical answers
- in fact t-test & F-test are equivalent only if we are testing 1 parameter & $H_a: \beta_1 \neq 0$
$H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$
otherwise we can only use t-test for testing that $H_a: \beta_1 < 0$ &/or $H_0: \beta_1 \leq 0$
$H_a: \beta_1 < 0$ $H_a: \beta_1 > 0$

i.e. $t^2 = F$
$(t_{\alpha/2, n-2})^2 = F_\alpha(1, n-2)$

e.g. $t^2 = (4375)^2 \Rightarrow 19.14 = F$
$t^2_{.025;8} = (2.306)^2 = 5.3176 = 5.32 \approx F_{.05}(1;8)$

$* \; t_{(n-2)} = \left(\frac{\hat{\beta}_1}{\sqrt{\frac{S^2}{S_{xx}}}}\right)^2 = \frac{\hat{\beta}_1^2}{S^2} = \frac{\hat{\beta}_1^2}{MSE} = \frac{\hat{\beta}_1^2 \cdot S_{xx}}{MSE} = \frac{S_{xy}^2}{S_{xx}^2} \cdot \frac{S_{xx}}{MSE}$

$= \frac{(S_{xy})^2}{S_{xx}} = \frac{SSR}{MSE} = \frac{MSR}{MSE} = F_{(1;n-2)}$

Review: x vs y (scatter plot)

② SLR model $y = \beta_0 + \beta_1 x + \varepsilon$

or $E(y) = \beta_0 + \beta x$, $\varepsilon \sim N(0, \sigma^2)$, $\forall x$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ - least squares line

$H_0: \beta_1 = 0$ (no linear)
$H_a: \beta_1 \neq 0$ (linear)

$s^2 = \frac{SSE}{(n-2)}$

(I) $t = \frac{\hat{\beta}_1}{\sqrt{\frac{s^2}{S_{xx}}}} = \frac{\hat{\beta}_1}{\frac{s}{\sqrt{S_{xx}}}}$ R.R. - Reject H_0 if $t > t_{\alpha/2, n-2}$ or $t < -t_{\alpha/2, n-2}$

or

(II) $F = \frac{MSR}{MSE}$ R.R. - we Reject H_0 if $F > F_{\alpha}(1, n-2)$

$MSR = \frac{SSR}{1}$

$MSE = \frac{SSE}{n-2} = s^2$

$t^2 = F$ only if we are testing one parameter
$H_0: \beta_1 = 0$ v $H_a: \beta_1 \neq 0$

4

Inferences Concerning $E(y)$

- assuming $H_0: \beta_1 = 0$ has been rejected (i.e. we have a linear relationship btwn x & y),
  ⇒ we might be interested in following/estimating
  - the average sales ($E(y)$) per month for a given level of expenditure (x)
  - mean response ($E(y)$) for a given drug dose (x)
  - average grade ($E(y)$) in stats for a given math test score (x)

$\hat{y} = \widehat{E(y)}$ → average mean of expected value

-here model is $y = \beta_0 + \beta_1 x + \varepsilon$
or $E(y) = \beta_0 + \beta_1 x$

$\Rightarrow \hat{y} = \hat{\beta} + \hat{\beta}_1 x$ ← fitted line (regression) can be used to estimate the unknown pop. line $E(y) = \beta_0 + \beta_1 x$

$E(\text{estimator}) = \text{parameter}$

$$E(\hat{y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x) = E(\hat{\beta}_0) + E(\hat{\beta}_1)^{(\text{constant})} = \beta_0 + \beta_1 x = E(y)$$

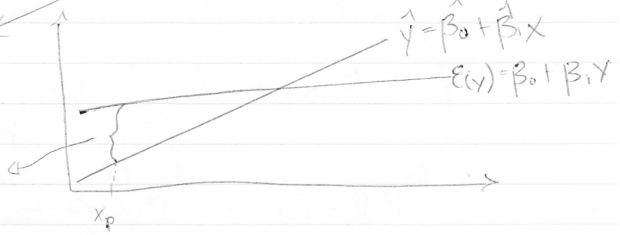
$\hat{y}$ ↓ unbiased estimator

σ^2 (sigma²) unknown → plug

$$V(\hat{y}) = \sigma^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right]$$

$\hat{y} \sim N(E(y), V(\hat{y}))$ (mean, variance)

Error of estimating $E(y)$ (when using regression line)



$$\hat{V}(\hat{y}) = s^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right] = MSE \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right]$$

-We might interested, for example, whether the mean corn yield per plot is 16 when we apply some amnt of fertilizer (say 6 lbs)

$H_0: E(y) = 16$, α $x_p = 6 \Rightarrow \hat{y}_{|x_p=6} = \hat{\beta}_0 + \hat{\beta}_1(6)$

$H_a: E(y) \neq 16$

σ^2 unknown can't use z, use t

test stat

$$t = \frac{\text{estimate} - E(\text{estimator})}{\sqrt{\text{var}(\text{estimator})}} \quad \frac{\hat{y} - E(y)}{\sqrt{s^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right]}} =$$

$$= \frac{\hat{y} - \mu_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{(6-\bar{x})^2}{S_{xx}}\right)}} \rightarrow \text{plug + solve}$$

- In general $H_0: E(y) = \mu_0$; α
  $H_a: E(y) \neq \mu_0$

test stat: $t = \dfrac{\hat{y} - \mu_0}{\sqrt{s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right]}}$

R.R. - Reject H_0 if $t > t_{\alpha/2, n-2}$ or $t < -t_{\alpha/2, n-2}$

ii) $H_0: E(y) > \mu_0$; α
$H_a: E(y) < \mu_0$

test stat: $t = \dfrac{\hat{y} - \mu_0}{\sqrt{s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right]}}$

R.R. - Reject if $t < -t_{\alpha, n-2}$

iii) $H_0: E(y) \leq \mu_0$; α
$H_a: E(y) > \mu_0$

R.R. - Reject if $t > t_{\alpha, n-2}$

$(1-\alpha)100\%$ C.I for $E(y)$

$$E(y) \in \left(\hat{y} \pm t_{\alpha/2, n-2}\sqrt{s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right]}\right) = \left(t_{\alpha/2, n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right)}\right)$$

$\hat{y}|x_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$

(Ex) cont

- Find a 95% C.I. for the expected ($E(y)$ "average") final grade in stats course given that the math test score was 50 (i.e. $x_p = 50$)

$1-\alpha = .95 \Rightarrow \alpha = .05 \Rightarrow \alpha/2 = .025$

$$E(y) \in \left(\hat{y} \pm t_{\alpha/2, n-2} \sqrt{MSE\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right]}\right)$$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_p =$
$= 40.784 + (0.766)(50)$
$= 79.08$

$MSE = 75.7532$
$S_{xx} = 2,474$
$\bar{x} = 46$

$$= 79.08 \pm \underbrace{t_{.025,8}}_{2.306} \sqrt{75.7532\left[\frac{1}{10} + \frac{(50-46)^2}{2,474}\right]}$$

$= 79.08 \pm 6.55 = \underline{(72.53, 85.63)}$ "on average" the grade will be btwn"

i.e. We are 95% confident (that in repeated sampling) that for students who scored 50 on a math test, the average final grade in stats course will fall in the interval (72.53, 85.63)

$\hat{y}$ < estimate $E(y)$
predict y

Predicting y for a given value of x (say x_p)

- suppose instead of estimating average stats course grade, we want to predict the final grade in stats for a new student, selected from pop. of interest, who scored 50 on a math test ($x_p = 50$)

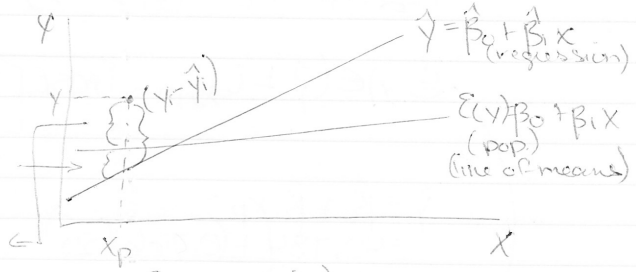
$E(y) = \beta_0 + \beta_1 x$

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

- Error of estimating $E(y)$: $V(\hat{y})$

- Variance of y's

- Variation of y about $E(y) = V(y)$



$E(\hat{y}) = E(y)$

$$V(y_i - \hat{y}_i) = V(y) - V(\hat{y}) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right)$$

$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right)$$

(for prediction error is greater than for estimation)

- CI will be wider)

$\hat{y} \sim N(E(y), V(y - \hat{y}))$

- but σ^2 unknown, we use s^2

$$\Rightarrow V(y - \hat{y}) = s^2\left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right) = MSE\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xy}}\right]$$

$(1-\alpha)100\%$ Prediction Interval (PI) for y

$$y \in \left(\hat{y} \pm t_{\alpha/2;n-2}\sqrt{s^2\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right]}\right)$$ can sub MSE for s^2

where $\hat{y}|_{x=x_p} = \hat{\beta}_0 + \hat{\beta}_1 x_p$

(Ex) cont'd - Find a 95% PI for y when $x_p = 50$

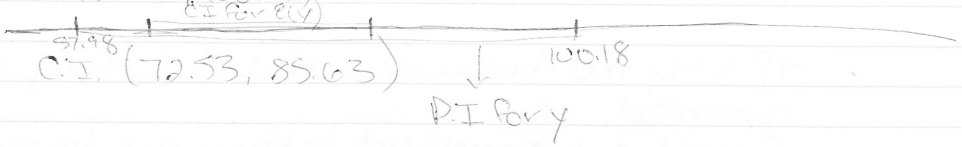
$$y \in \left(\hat{y} \pm t_{\alpha/2, n-2} \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right)} \right) \Bigg| \begin{array}{l} 1-\alpha = .95 \Rightarrow \alpha = .05 \\ \alpha/2 = .025 \end{array}$$

$\hat{y} = 40.784 + (.766)50 = 79.08$

$$= \left(79.08 \pm 2.306 \sqrt{75.7532\left(1 + \frac{1}{10} + \frac{(50-46)^2}{2474}\right)}\right)$$

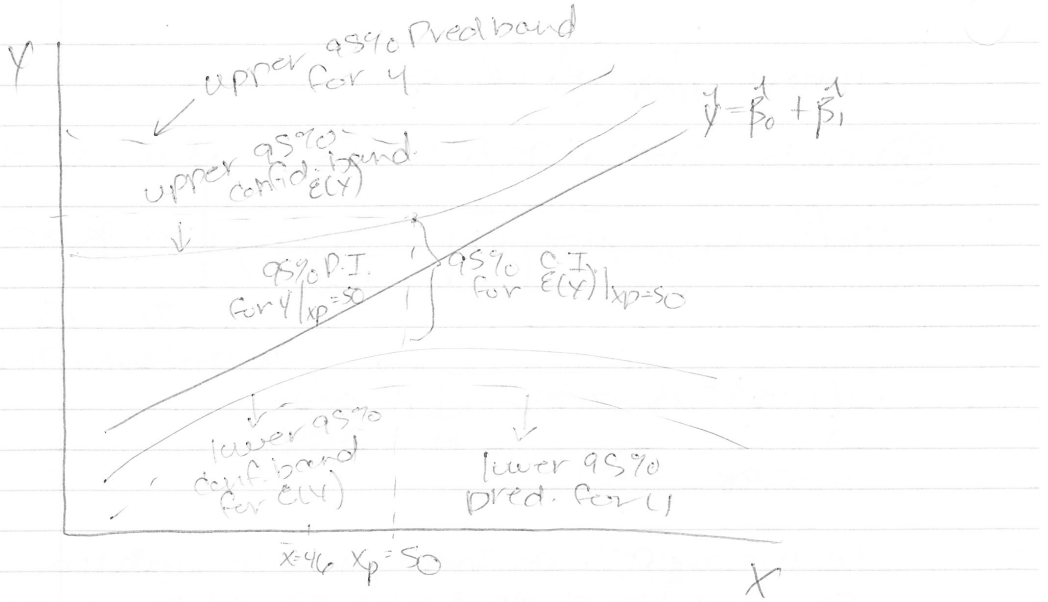
$$= (57.98, 100.18)$$

i.e. We are 95% confident, that in repeated sampling, for a student who scored 50 on the math test, his/her final grade in stats will lie in (57.98, 100.18)



NOTES • P.I for an individual is always greater than C.I. for a group

- Both P.I & C.I have their min. value at the pnt where $X = \bar{X}$
- By plotting the end pnts of the C.I's or P.I.'s for different values of x_p & connecting these end pnts we obtain a general $100\%(1-\alpha)$ confidence bands for $E(y)$ or (Prediction bands for y)



Correlation - measures the strength & direction of a linear relationship btwn x & y

- if x & y are linearly related ⇒ they are correlated
- if x & y are correlated ⇒ they are linearly related

N - the sample "coefficient of correlation", also known as "Pearson product-moment coefficient of correlation"

$r = \frac{S_{xy}}{\sqrt{(S_{xx})(S_{yy})}}$ or $r = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}}$ or

$= \frac{S_{xy}}{S_{xx}} \sqrt{\frac{S_{xx}}{\frac{S_{xy}^2}{S_{yy}}}}$

$= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

"correlation coefficient"

- Properties of r

① $-1 \leq r \leq 1$

② $r = 0 \Rightarrow$ x & y are not linearly related (ie they are not correlated)

③ $r > 0 \Rightarrow$ pos. linear relation.

④ $r < 0 \Rightarrow$ neg. linear relation.

→ cont

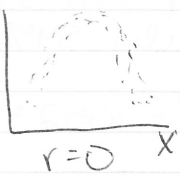
⑤ r^2 "correlation of determination"

$= \frac{SSR}{TSS}$ - measures how much of the tot. variation in the data can be explained by the regression line (& hence how much is due to error.)

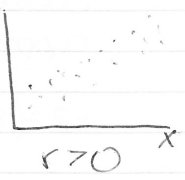
i.e. $0 \leq r^2 \leq 1$ @ 0 model is iffy

a) 

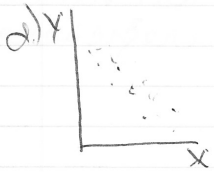
$r = 0$

b) 

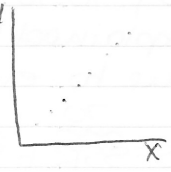
$r = 0$

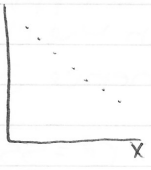
c) 

$r > 0$

d) 

$r < 0$

e) 

f) 

a) + b) - no lin. rel.
$\Rightarrow SSR = 0$
$\Rightarrow TSS = SSE$
$r^2 = \frac{SSR}{TSS} = \frac{0}{TSS} = 0$

$r = +1$ ↘ ↙ $r = -1$
perfect fit $\Rightarrow$ no error
$SSE = 0$
i.e. $TSS = SSR$
i.e. $r^2 = \frac{SSR}{TSS} = 1$

Ex cont'd

- we wish to calculate the coeff of correlation btwn math test results & the final grade in Stats course & also calculate the coefficient of determination, r^2
- scatter plot suggests a pos. linear relation. btwn X & Y

$S_{xx} = 2474$
$S_{yy} = 2056$
$S_{xy} = 1894$

$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{1,894}{\sqrt{(2474)(2056)}} = 0.83978 = .84$

i.e. math test results & final grade in stats are positively correlated (or positively linearly) related w/ the strength of their relationship approximately 84%.

$r^2 = (.83978)^2 = .7052 = .70$

i.e. approximately 70% of the tot. variation in the data is explained by the regression line, hence 30% is due to error

$r^2 = \frac{SSR}{TSS}$ $\left(TSS = \overbrace{SSR}^{70} + \overbrace{SSE}^{30}\right)$

NOTES
① • r = sample coefficient of correlation and is an estimate of a unknown pop. coefficient of correlation
• ρ (rho) = pop. coeff. of correlation
② $H_0: \rho = 0$ (no correlation) $\Leftrightarrow H_0: \beta_1 = 0$ (no lin. relat.)
$H_a: \rho \neq 0$ (correlation) $H_a: \beta_1 \neq 0$ (lin. relat.)

If one test is done in a prob, you don't have to do the other, they are the same

(for correlation) test-stat. $\Leftrightarrow t = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1}{\sqrt{\frac{s^2}{S_{xx}}}}$

$\downarrow t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ (equivalent)

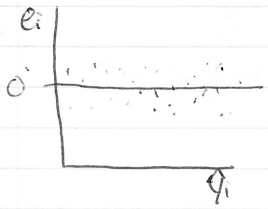
Examining Lack of Fit in Linear Regression (Residual Analysis)

- So far we've been working w/ the model $Y = \beta_0 + \beta_1 X + \varepsilon$ when checking of its validity + the validity of its assumptions
- ∴ we need to perform residual analysis to check whether the model + its assumptions have no violations.
- Should be performed immediately after finding the fitted regression line + before any hypothesis testing, estimation, +/or prediction
- 2 ways of performing residual analysis:
  ① graphical ② numerical

① graphical residual analysis

(i) scatter plot (plot X vs Y) ← to check for violations of linearity
- to check for any outliers ect.

(ii) Plot of predicted values vs. residuals (i.e. $\hat{y}_i$ vs e_i) $e_i = (y_i - \hat{y}_i)$ ← To check for any violations of assumptions of Y's or ε's being independently distributed + /or violations of linearity.

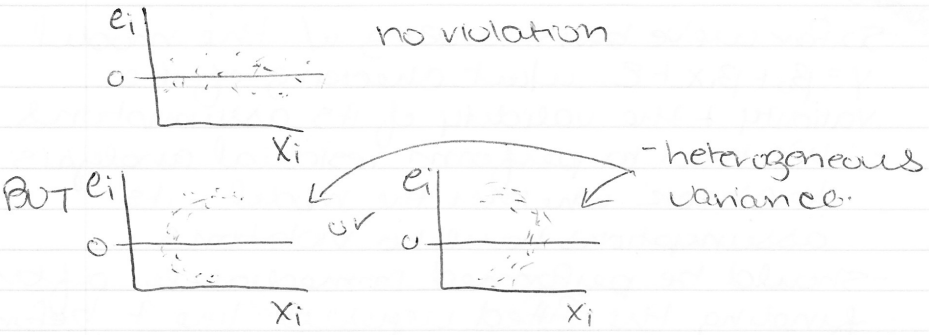


- If no pattern ⇒ No violations

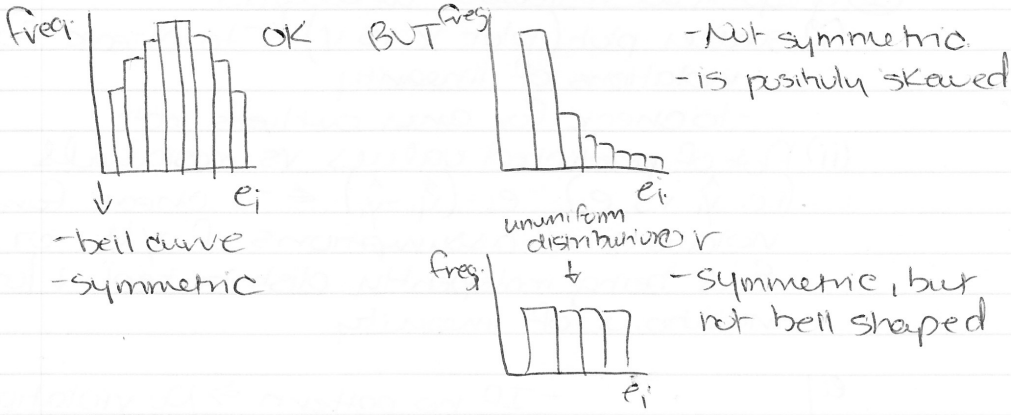
BUT

- curved linear pattern ⇒ violation
⇒ Higher order terms need to be added to the model

(iii) Plot X vs e_i ← To check for validity of the assumption of constant variance



(iv) Plot histogram of errors → To check validity of the assumption of errors being normally distributed



- Good fit = no violations + high r^2

missing Lecture