Chap. 12 - Simple Linear Regression (SLR)

- looks @ a linear relationship btwn variables

eg. 1 - wish to determine a student's grade in (Y) want to predict stats course @ the end @ the 1st yr of university before the student has been enrolled and, based on the math test results (X) that student took prior to university entrance. (SLR) 2 variables

eg. 2 - Want to estimate wheat yield (Y) based on the # of cm of rainfall (X_1) # of hrs of sunshine (X_2) + soil type (X_3) (MLR) → 2 or more variables

y = response/(dependent) variable Depends on values of x

x = independent/(explanatory)/predictor variable

x, y can be quantitative / qualitative; discrete / continuous

If a relationship btwn x + y is linear it can be written down by the equation of a straight line

$y = a + bx$ — not used in course

used: $y = \beta_0 + \beta_1 x$

β_0 = y intercept

β_1 = slope

β_0:

- part of average value of y that doesn't chg w/ x (i.e. when x=0)

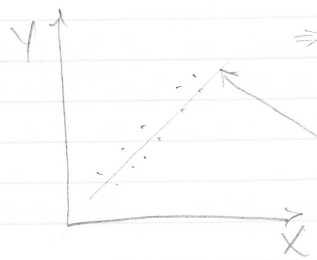
- β_1 = slope. pos. relationship; neg. rel.

prt of average value of y that chngs w/ a unt incrs of x

- Deterministic model - each pnt lies on the line
- values of y determined by x

- model is deterministic because the values of y are *exactly* determined from values of x
- unfortunately in practice this model is suitable only for explaining physical phenomena when the error is negligible.

I.e In practice, we always get an error.



$\Rightarrow$ error $\Rightarrow$ probablistic (or statistical) model

$$y = \beta_0 + \beta_1 X + \varepsilon$$

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

where ε is the error term, it takes into acct all unpredictable & unknown factors which have not been included in the relationship (because they are unimportant, unmeasurable, or unknown)

- We assume that errors average to zero, ie $E(\varepsilon) = 0$ have some standard variance & that they are normally distributed

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

$\Rightarrow$ if $E(\varepsilon) = 0 \Rightarrow y = \beta_0 + \beta_1 X + \varepsilon$

$$\Rightarrow E(y) = E(\beta_0 + \beta_1 X + \varepsilon)$$

$$= E(\beta_0 + \beta_1 X) + \underbrace{E(\varepsilon)}_{=0}$$

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

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

We assume:

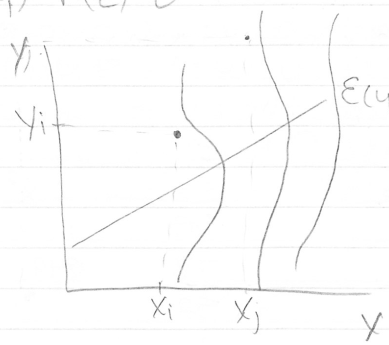
1. That X's are observed w/out error
2. Y's are independently distributed w/ the mean $E(y) = \beta_0 + \beta_1 X$
3. Y's have constant variance σ^2, $\forall X$ ("$\forall$" "for every value")
4. $Y \sim N(E(y), \sigma^2)$, $\forall X$

pf $V(y) = V(\beta_0 + \beta_1 + \varepsilon) =$ $\quad V(x+y) = V(x) + V(y) + 2\,cov(x,y)$

$\hookrightarrow \underbrace{V(\beta_0 + \beta_1 X)}_{V(const) = 0} + \underbrace{V(\varepsilon)}_{\sigma^2} + 2\cancel{cov(\beta_0 + \beta_1 X + \varepsilon)}$

since assumption #1 says that x's are observed w/out error

$V(y) = V(\varepsilon) = \sigma^2$



$E(y) = \beta_0 + \beta X$

i.e. assumptions are
① equivalent to x's observed w/o error
② ε's are indep. distributed w/ mean $E(\varepsilon) = 0$
③ Errors have const. var. σ^2 $\forall x$
④ $\varepsilon \sim N(0, \sigma^2)$, $\forall x$

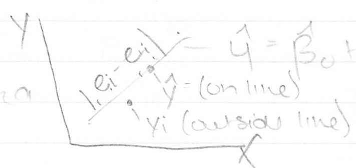
Assumptions must be stated using Y's or ε's

Because $y = \beta_0 + \beta_1 X + \varepsilon$ is the model for the pop., the values of $\beta_0 + \beta_1$ are unknown, so we need to estimate them

unknown pop. parameters $\beta_0 + \beta_1$ are estimated using sample statistics $\hat{\beta}_0 + \hat{\beta}_1$

i.e. $\hat{E}(y) = \boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X}$ (based on sample pts)

pop.- uses greek
sam.- uses reg. alpha
^ - for sample



$\hat{\beta}_0$ = y intercept = part of estimated average value of y that doesn't chng w/ X

$\hat{\beta}_1$ = slope = part of estimated average value of y that changes w/ the unit incrs of X

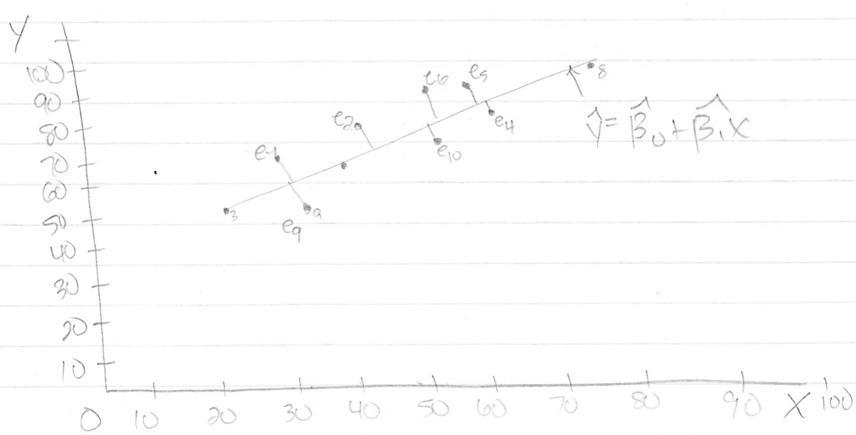
y_i = actual sample value of y when $X = X_i$
$\hat{y}_i$ = value given by estimated regression line when $X = X_i$
$e_i = y_i - \hat{y}_i$ = residual (i.e. error)

cont. on p. 4

Ex Assume n=10 students constitute a random sample from a pop. of 1st yr. university students who have already entered the university or will do so in the immediate future:

Students	math test Results (X_i)	final grade Stats (Y_i)	X_i^2	X_iY_i	Y_i^2
1	39	65	1521	2535	4225
2	43	78	1849	3354	6084
3	21	52			
4	64	82			
5	57	92			
6	47	89			
7	28	73			
8	75	98			
9	34	56			
10	52	75			

First we plot a scatter plot or (i.e. scatter diagram)

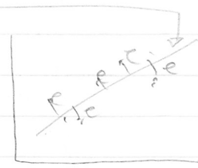


scatter plot indicates a pos. linear relationship btwn math test results + Final grade in stats

i.e SLR model $y = \beta_0 + \beta_1 X + \varepsilon$ or
$E(y) = \beta_0 + \beta_1 X$, $\varepsilon \sim N(0, \sigma^2)$ } assumptions

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

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X \Rightarrow$ estimates for pop



-Method of Least Squares
-it minimizes the errors (so-called residuals)
$e_i = y_i$ (observed) $- \hat{y}_i$ (Predicted)

"Choose as the best fitting line, the line that minimizes the sum of squares of the deviations of the observed values of y from those predicted"

→ sum of squares for error
i.e. $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$

$(a+b)^2 = a^2 + b^2 + 2ab$ (Derivative)

$\frac{\partial SSE}{\partial \hat{\beta}_0} \Rightarrow 2 \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)](-1) = 0$

$= \sum y_i = \sum_{i=1}^{n} \hat{\beta}_0 + \sum \beta_1 x_i$

$= n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \ldots (1)$

$\frac{\partial SSE}{\partial \beta_1} \Rightarrow 2 \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)](-x_i) = 0$

i.e. $\sum x_i y_i = \sum \beta_0 x_i + \sum \hat{\beta}_1 x_i^2$

$\sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 \ldots (2)$

- equations 1 + 2 are called normal equations (N.E.'s)

Solving these 2 N.E.'s will give us

$$\boxed{\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_i x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum_i x^2 - \frac{(\sum_i x_i)^2}{n}}\end{aligned}}$$

$\bar{y}$ = mean of y $\bar{x}$ = mean of x

$$\underbrace{\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}}_{S_{xx}} = \sum_i (x_i - \bar{x})^2 = V(x)$$

same

$$(\sum x_i - \bar{x})(y_i - \bar{y}) \quad \underbrace{\sum_i x_i y_i - \frac{(\sum_i x_i)(\sum y_i)}{n}}_{S_{xy}} = cov(xy)$$

$$\boxed{\text{i.e. } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

i.e. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$ ← least squares line, or best fitting line, or regression line, or prediction line, or (line = equation)

Least squares method guarantees that the estimates $\hat{\beta}_0$ + $\hat{\beta}_1$ are socalled BLUES "Best Linear Unbiased Estimators"

i.e. they are linear estimators ✓

- unbiased because ~~expected~~ value is such that expected value

E(estimator) = parameter

$E(\hat{\beta}_0) = \beta_0$

$E(\hat{\beta}_1) = \beta_1$

$E(\bar{x}) = \mu$

$E(s^2) = \sigma^2$

and variances of $\hat{\beta}_0$ + $\hat{\beta}_1$ are the smallest ⇒ best ✓

$$V(\hat{\beta}_0) = \frac{\sigma^2 \sum_i x_i^2}{n S_{xx}} \qquad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

- if unknown must estimate w. s^2 p. 4

from p.2

Ex. $\sum_{1}^{} y_i = 760$ $\sum_{1}^{} x_i = 460$ $\sum x_i^2 = 23{,}634$ $\sum x_i y_i = 36{,}854$ $\sum y_i^2 = 59{,}816$

$$\bar{Y} = \frac{\sum_{1}^{} y_i}{n} = \frac{716}{10} = 76$$

$$\bar{X} = \frac{\sum x_i}{n} = \frac{460}{10} = 46$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{1}^{} x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{36{,}854 - \frac{(460)(7\overset{60}{\cancel{16}})}{10}}{23{,}634 - \frac{(460)^2}{10}}$$

$$= \frac{1{,}894}{2{,}474} = 0.7655618 = .766$$ → can only round if this is final answer

↓ if using further, use whole # don't round

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 76 - (.7655618)46 = \underline{40.7841572} = 40.784$$

least squares line is given by

$\hat{y} = \underline{40.784 + 0.766X}$ (Represents linear relationship)

if we want to predict new students stats grade knowing that the student scored 50 on a math test (i.e. $x = 50$)

x-must be w/in range to predict y

$\hat{y} = 40.784 + 0.766(50)$

$= \underline{79.08}$

-Estimation of σ^2

-not only β_0 + β_1 are unknown pop. parameters that need to be estimated using $\hat{\beta}_0$ + $\hat{\beta}_1$, but σ^2 (i.e. pop. variance) needs to be estimated as well using s^2 (sample variance)

→ over

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [\underbrace{y_i}_{a} - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{b}]^2$$

$(a \pm b)^2 = a^2 \pm b^2 \pm 2ab$

$$= \sum y_i^2 + \sum(\hat{\beta}_0 + \hat{\beta}_1 x_i)^2 - 2\sum y_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) =$$

$$= \sum y_i^2 + \sum \hat{\beta}_0^2 + \sum \hat{\beta}_1^2 x_i^2 + 2\sum \hat{\beta}_0 \hat{\beta}_1 x - 2\sum \hat{\beta}_0 y_i - 2\sum \hat{\beta}_1 x_i y_i$$

constants can come out

$$= \sum y_i^2 + n\hat{\beta}_0^2 + \hat{\beta}_1^2 \sum x_i^2 + 2\hat{\beta}_0\hat{\beta}_1 \sum x_i - 2\hat{\beta}_0 \sum y_i - 2\hat{\beta}_1 \sum x_i y_i$$

= sub in $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ =

$$= \underbrace{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}_{S_{yy}} + \underbrace{\hat{\beta}_1^2 \sum x_i^2 - \hat{\beta}_1^2 \frac{(\sum x_i)^2}{n}}_{\hat{\beta}_1^2 S_{xx}} - \underbrace{2\hat{\beta}_1\left(\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right)}_{2(\hat{\beta}_1 S_{xy})}$$

will need

$$= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2(\hat{\beta}_1 S_{xy})$$ sub $\hat{\beta}_1 = \frac{S_{xy}}{S_{xy}}$

$$= S_{yy} + \hat{\beta}_1 \left(\frac{S_{xy}}{\cancel{S_{xx}}}\right) \cdot \cancel{S_{xx}} - 2\hat{\beta}_1 S_{xy}$$

$$= S_{yy} + \hat{\beta}_1 (S_{xy}) - 2\hat{\beta}_1 S_{xy}$$

$$= S_{yy} - \hat{\beta}_1 S_{xy}$$

sub $\hat{\beta}$ $= S_{yy} - \left(\frac{S_{xy}}{S_{xy}}\right) S_{xy} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$

sum of squared errors

$$\therefore SSE = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

var of y — cov of xy — var of x

$$\boxed{s^2 = \frac{SSE}{n-2}}$$ i.e. $$\boxed{s^2 = \frac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{n-2}}$$ ⇒ estimate for σ^2 ⇒ $s = \sqrt{s^2}$

2 parameter estimates ($\hat{\beta}_0 + \hat{\beta}_1$)

$$v(x) = \frac{\sum (x_i - \bar{x})^2}{n-1}$$ ⇒ 1 parameter estimates

degrees of freedom for t test

NOTE $s = \sqrt{s^2}$

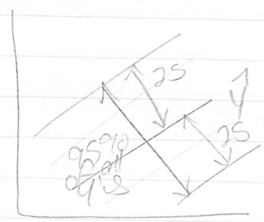
standard error of estimation + it is an estimate of how much y varies about the true mean $E(y) = \beta_0 + \beta_1 x$

- the smaller s, the more accurate our estimation / prediction

- σ measures the spread of the y values about their line of means $E(y) = \beta_0 + \beta_1 x_i$ ie. from empirical rule we would expect about 95% of ~~all~~ the y values to fall w/in $E(y) \pm 2\sigma$

* Empirical Rule: if bell shaped data
- approximately 68% of all observations will lie w/in $\mu \pm \sigma$
  - 95% $\mu \pm 2\sigma$
  - 99% $\mu \pm 3\sigma$

i.e. Since $E(y)$ + σ are unknown, we estimate them w/ $\hat{y}$ + s



Ex $s^2 = \dfrac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{n-2} = \dfrac{\left[\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}\right] - \dfrac{\left[\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}\right]^2}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}}{n-2}$

$= \dfrac{\left[59816 - \frac{(760)^2}{10}\right] - \frac{(1894)^2}{2474}}{8} = \dfrac{2056 - \frac{(1894)^2}{2474}}{8} =$

$= 75.7532$

$\Rightarrow s = \sqrt{s^2} = \sqrt{75.7532} = 8.7036 \approx 8.7$

Is s small enough?

θ - parameter (unknown)
$\hat{\theta}$ - estimator
$c.v(\hat{\theta}) = \frac{se(\hat{\theta})}{\theta}$
coefficient of variation

ex $\theta = E(y)$ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$ $V(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n S_{xx}}$

$\hat{\theta} = \hat{y}$

$CV(\hat{y}) = \frac{Se\,\hat{y}}{E(y)} = \frac{\sqrt{s^2}}{\bar{y}}$ $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

Relative coefficient of variation

$\frac{s}{\bar{y}} = \frac{8.70}{76} = 0.1144$ or 11.44 % - Determines how good estimator is.

not covered in this course

$y = \beta_0 + \beta_1 + \varepsilon$ - pop line
$E(y) = \beta_0 + \beta_1 X$, $\varepsilon \sim N(0, \sigma^2) \forall x$

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

Testing the significance of the linear relationship
- Is pop. line linear?
  - Is slope (β) $= 0$; not linear
  - " " $\neq 0$; linear

we've seen that there is a linear relationship in a sample, now we need to test wheather pop has a linear relationship as well.

Remember $\boxed{\sum x_i^2 \neq (\sum x_i)^2}$

To test for a linear relationship btwn x + y is = to testing wheather the slope, β_1, of the line is equal to zero or not.

null ↱ no hats, because test is for the pop.

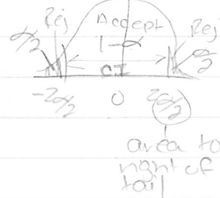
i.e. $H_0: \beta_1 = 0$ (no linear relationship)

alternative want to prove $H_a: \beta_1 \neq 0$ (linear relationship)

two tailed test, or 1 tail

↳ never =

α - type I error; reject H_0 when true



use hats for estimates of pop.

since N dis. use z test

test stat.: $z = \dfrac{\text{estimate} - E(\text{estimate})}{\sqrt{Var(\text{estimate})}}$

$$= \frac{\hat\beta_1 - E\hat\beta_1}{\sqrt{Var(\hat\beta_1)}} = \frac{\hat\beta_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} = \frac{\hat\beta_1}{\sqrt{\sigma^2/S_{xx}}}$$

$E(\hat\beta_1) = \beta_1$

$V(\hat\beta_1) = \dfrac{\sigma^2}{S_{xx}}$

since σ^2 is unknown, estimate it w/ S^2 + $z \xrightarrow{\text{changes}} t_{(n-2)}$

$S^2 = \dfrac{SSE}{n-2}$

z used when σ^2 is known; t when unknown

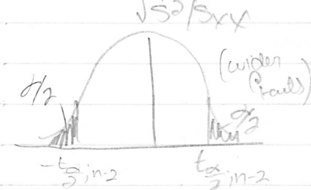
(I) t-test - must be approx. normal

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

test-stat: $t = \dfrac{\hat\beta_1}{\sqrt{S^2/S_{xx}}}$

R.R.



We reject H_0 is test-stat falls in either rejection reg. $t > t_{\frac{\alpha}{2};n-2}$ $t < -t_{\frac{\alpha}{2};n-2}$

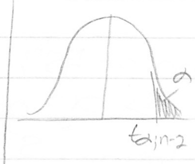
(ii) 1 sided

negate H_a $H_0: \beta \leq 0$

$H_a: \beta_1 > 0$

test stat. same

R.R.



We rej. if $t > t_{\alpha;n-2}$

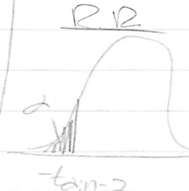
or

(iii) 1 α

$H_0: \beta \geq 0$

$H_a: \beta_1 < 0$

R.R.



We reject if $t < -t_{\alpha;n-2}$

(Ex) con'd: Is there a linear relationship btwn a 1st yr. student's result of a math test (x) and a final grade in stat's course (y)?

* $\alpha = .05$ (use if not given)

$H_0: \beta_1 = 0$ (no lin. rel.) $\quad \alpha = .05 \Rightarrow \alpha/2 = .025$

$H_a: \beta_1 \neq 0$ (have lin. rel.)

test-stat. $t = \dfrac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} = \dfrac{.76556618}{8.73/\sqrt{2474}} = \underline{4.375}$ use full value

R.R. we rej. H_0 if $t > t_{\frac{\alpha}{2};n-2}$ or $= (t_{.025;8}) = 2.306$

$t < -t_{\frac{\alpha}{2};n-2} = -t_{.025;8} = -2.306$

Always state conclusion significance level

* Since $t = 4.375 > 2.306$, we rej H_0 and conclude that @ 5% level of significance there is a linear relationship btwn results of math test & final grade in stats course.

— (using C.I to see if linear rel. exists)

$(1-\alpha)100\%$ C.I for β_0 &/or β_1

"lies in"

β_0: $\beta_0 \in \left(\hat{\beta}_0 \pm t_{\frac{\alpha}{2};n-2}\sqrt{\dfrac{s^2 \Sigma x^2}{n S_{xx}}}\right)$ sample

β_1: $\beta_1 \in \left(\hat{\beta}_1 \pm t_{\frac{\alpha}{2};n-2}\sqrt{s^2/S_{xx}}\right)$ sample

(Ex) con'd: Find a 95% CI for β_1

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

$\beta_1 \in \left(\hat{\beta}_1 \pm t_{\frac{\alpha}{2};n-2}\sqrt{s^2/S_{xx}}\right) = .766 \pm \underbrace{t_{.025;8}}_{2.306} \dfrac{8.73}{\sqrt{S_{xx}}}$

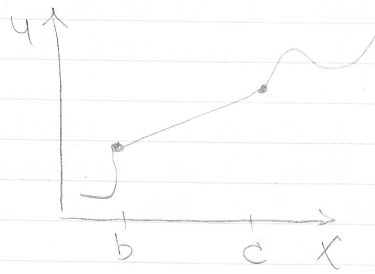
$= (0.766 \pm 0.403) = (.363, 1.169)$

if includes 0; not linear

* state assumptions (N)

i.e. We are 95% confident that in repeated sampling the true pop. slope β_1 would be in $(0.363^{\circ}, 1.169)$

NOTE: If data has x values btwn b + c ($b \leq x \leq c$) then prediction equation/line is valid only for this region ∴ don't extrapolate!



- Even if we don't reject H_0 (ie $H_0: \beta_1 = 0$ (is true)), it doesn't necessarily mean that x + y are unrelated
  - it only means that x + y aren't linearily related but they may be related in a none linear way.

- If we reject H_0 (i.e. $H_a: \beta \neq 0$ (is true)), we don't conclude that the true relationship btwn x + y is linear. It only means that a linear relationship is resonable (since y may depend on other variables as well)

- If we accept H_a (i.e. $H_a: \beta_1 \neq 0$), we don't conclude that there is a CAUSAL relationship btwn x + y.
  ex. x = % of grey hair a person has
  y = blood pressure

one would expect a positive relationship btwn x + y BUT even though % of grey hair may be a good predictor of blood pressure, this does not imply that an additional grey hair causes your blood pressure rise. It is most likely that a 3rd factor (Age) causes both % of grey hair + blood press ↑

※ - We reject H_0 + conclude that @ α% level of significance there's an evidence that the statement in Ha is true

- We do not reject H_0, + conclude that @ α% level of significance there's not enough evidence to say that the statement in Ha is true.

II Analysis of variance (ANOVA)
for testing of a linear relationship

(variance of y_i) $V(y) = \sum_i (y_i - \bar{y})^2$

$a^2 + b^2 + 2ab$

$$y_i - \bar{y} = \overset{a}{(y_i - \hat{y}_i)} + \overset{b}{(\hat{y}_i + \bar{y})}$$ / both sides

$$(y_i - \bar{y})^2 = (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$ | $\sum_i$ both sides

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\underbrace{\cancel{\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}}_{\text{sum of cross product terms is always } = 0} \to 0$$

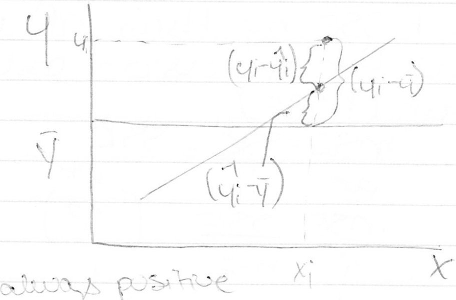
$$\underbrace{\sum_i (y_i - \bar{y})^2}_{TSS} = \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{SSE}$$

→ sum of squares due to error. measures variation of the actual observation to the prediction line. ↑

TSS = tot sum of squares
SSE = sum of squared errors

TSS measures the tot variation in y_i's ie $V(y)$

SSR: sum of squares for regression - measures the variation of the pnts on the line to the line of means



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

* always positive

i.e. $\underbrace{TSS}$ = SSR + SSE

Computation formulas

$TSS = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} = S_{yy}$

$SSR = \frac{\left[\sum_i x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right]^2}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{(S_{xy})^2}{S_{xx}}$

$SSE = TSS - SSR = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$

if small, line is gd fit

NOTE: If we have perfect fit ⇒ SSE = 0 ⇒ TSS = SSR

If there's no linear relationship btwn x & y ⇒ SSR = 0 ⇒ TSS = SSE

$df_{TSS} = n-1$ — always

$V(x) = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$

$df_{SSR} = K$ (K = # of parameters associated w/ x's)

$= 1$ (in SLR case) $Y = \beta_0 + \beta_1 X + \varepsilon$

$df_{SSE} = n-2$ (in SLR case)

$= n -$ tot. # of parameters in the model, including β_0)