2

- Consequently the error of predicting a particular value of y will always be larger than the error of estimating the mean value of y for particular x

- The further x_p lies from $\bar{x}$, the larger will be the errors of estimation and prediction

Q2 a $b_0 = \bar{y} - b_1\bar{x}$

$n = 5$

$\Sigma x_i = 2150$ $\bar{x} = 430$

$\Sigma y_i = 1430$ $\bar{y} = 286$

$\Sigma x_i y_i = 618500$

$\Sigma x_i^2 = 931100$ $\Sigma y_i^2 = 411900$

$$b_1 = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}$$

$$\frac{618500 - \frac{2150(1430)}{5}}{931100 - \frac{(2150)^2}{5}} \quad \frac{3600}{6600} = \frac{6}{11} \quad 0.5454$$

$$b_0 = 286 + \frac{6}{11}(430) = 51.4545$$

51.45

$$y_i = 51.45 + \frac{6}{11}x_i$$

b $MSE = \frac{SSE}{n-2}$ $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$

$$\Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - n\bar{y}^2$$

$$\Sigma y_i^2 - n(\bar{y})^2 - b_1[\Sigma x_i y_i - n\bar{x}\bar{y}]$$

$$411900 - 5(81796) - \frac{6}{11}[618500 - 5(430)(286)]$$

$$2920 - \frac{21600}{11}$$

$$= \frac{956.363}{3} = 318.787 = \frac{10520}{33} = s^2$$

c $$r^2 = \frac{S_{xy} \times S_{xy}}{S_{xx} \times S_{yy}} = \frac{(3600)^2}{(6600)(2920)}$$

$$\frac{12960000}{19272000} = \frac{540}{803} = 0.6724$$

3

3 Regression 2012 Exam Part

Q2. R^2 mean proportion of total variance about the mean mean $\bar{y}$, explained by the regression

r^2 explained 67% of the total variation in the data about the average $\bar{y}$. Goes from $0 \to 1$.

If r^2 was 20% I would be concerned, 20% of the error accounted for is due to chance, the other 80% is due to our model

2 R 67.24 $= R^2$ is a decent figure, roughly 70% of error is due to variation about the mean $\bar{y}$.

With each increase in market value, the mean value for distribution of sale price rises by 6/11, Thus to increase the sale price by one, the market value would need to increase by 2.

E. Assumptions:
- X_i is the i^{th} value of the predictor variable, which is a known constant for all i.
- ε_i is a random error term with properties:
  - $E(\varepsilon_i) = 0$.
  - $Var(\varepsilon_i) = \sigma^2$
  - ε_i and ε_j are uncorrelated so $Cov(\varepsilon_i, \varepsilon_j) = 0$ for i, j $i \neq j$
  - ε_i are normally distributed $N(0, \sigma^2)$
- The means of Y_i can be joined by a straight line, given as:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

Where β_0 and β_1 are unknown parameters such that:
- β_1 is the slope of regression line and indicates the change in mean of y for unit increase in x.

- B_0 is the intercept of the regression model. B_0 gives the mean distribution of y at $x=0$,

Q3 Test if there is a relationship between the two variables or not

$H_0: B_1 = 0$ vs. $H_1: B_1 \neq 0$

P value = 0.018 less than 0.05, reject H_0, there is a relationship between sweetness index and pectin.

Are related as per data

b) Slope is -0.00231, for each unit increase in pectin, the mean value distribution of sweetness decreases by 0.00231 units.

$b_0 = 6.2521$, this is the sweetness index value when pectin is set to zero

c) $= 6.25 - 0.00231(250) = 5.67235$ sweetness

R^2 value of 27.9 suggests a bad fit for the model. Proportion of total variability about the mean as explained about the regression line = 0.23, implying a bad fit → unreliable prediction

D) ___

Regression 2011 Exam Paper

a **Assumptions:**

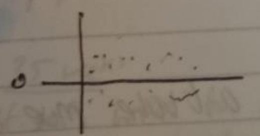
- x_i is the i^{th} value of the predictor variable, which is a known constant for all i.
- The observations y_i (or ε_i) are independent.
- At any given x_i, y_i (or ε_i) is normally distributed
- The observation y_i (or ε_i) have constant standard deviation
- The means of y_i can be joined by a straight line given as:

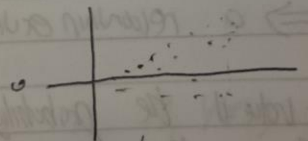
$$E(y_i) = \beta_0 + \beta_1 x_i$$

where β_0 and β_1 are unknown parameters

- β_1 is the slope of the regression ~~model~~ line and indicates the change in the mean of y for one unit increase in x.
- β_0 is the intercept of the regression model. If it is sensible to think of a value of $x=0$ for a particular application, then β_0 gives the mean of the distribution of y at $x=0$, it is not always possible to have a physical explanation for this parameter

B 1 constant variance for ε_i for all i:

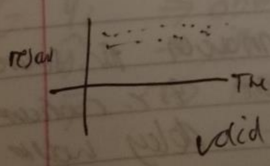


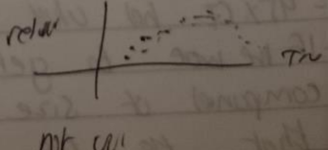


mean of 0 ✓

2 Independent of ε_i

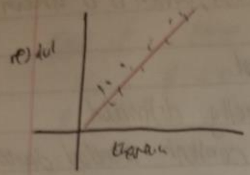
Time order plot

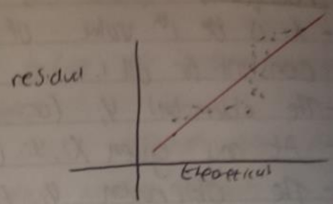




2

3. Normality of ε_i:
Q-Q plot





1.C $H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$
Compare t_{calc} with t_{crit}.
$t_{critical} = 2.306$
$|t_{calc}| = 11.69 > t_{critic}$, reject H_0, thus we reject that time taken to adopt innovation is 0 for a company of size = 4

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$.
$t_{critical} = 2.306$
$|t_{calc}| = 4.43 > t_{critical}$ reject H_0.
Reject that there is a zero change in time taken to adopt innovation for a 1 unit increase in size
$\Rightarrow$ a relationship exists

P-value is the probability of getting a t value more extreme than the one we got in the test

D- The point estimate for the average value of delay in innovation for company of size 60m is given by $y = b_0 + b_1(60)$.
- 95% CI has closed interpretation
- If we were to get the delay in innovation for several companies of size 60m, then we are 95% confident that the true mean (average of delay would lie within this interval

Regression 2011 Exam Paper

A point estimate for the delay time for a 'single' company of size 60m is also $y' = b_0 + b_1(60)$

But a 95% PI refers to the interval within which the true delay time of that single company of size=60m will lie with 95% confidence

They are different since the variability of delay time for a single firm is much greater than the variability of the avg

Single observation = mean + chance variation

This is why 95% PI are wider than 95% CI.

A 60m company, book: single intern

i. 20 weeks PI = (4.93, 24.02)

If we were to test for $H_0: y'=20$ vs $H_1: y' \neq 20$ then from PI we see that null hypothesis will not be rejected

→ 20 can be accepted as a good point estimate for true population

ii. 35 weeks

If we were to test $H_0\ y'=35$ vs $H_1\ y' \neq 35$, from PI we can reject H_0

- 35 cannot be accepted as a good estimate for the prediction of delay time

G. y_i ⇒ observed y

$\hat{y}_i$ ⇒ ith fitted val

$\bar{y}$ ⇒ mean of y_i for all y

SSTO = SSR + SSE

$$\Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2$$

$$R\text{-}Sq = \frac{SSR}{SSTO}$$

R^2 is the proportion of total variability about the mean $\bar{y}$, a) explained by the regression line