2012 Regression Exam Part

B Sample Size = 20
Mean of Sample = 2.1
SE(mean) = $\frac{1.552}{\sqrt{20}}$ = 0.347

95% CI = 2.1 ± 0.347(1.96)
= (1.374, 2.826)
95% confident that the true mean value lies in (1.374, 2.826)

D CI $\hat{y} \pm t_{critical} \sqrt{MSE\left[1/n + (x_i-\bar{x})^2/S_{xx}\right]}$
PI $\hat{y} \pm t_{critical} \sqrt{MSE\left[1 + 1/n + (x^*-\bar{x})^2/S_{xx}\right]}$

- CI used model for estimating the mean value of y, E[y] for a specific value of x.
- PI predicts a particular y value for a given x_p, predicting outcome of single experiment given x value

CI is narrower why?
- The error in estimating the mean value of y, E[y], for a given x_p, is the distance between the least squares line and the true line of mean $E[y] = \beta_0 + \beta_1 x_i$
- Error shown in first $[\hat{y} - E[y]]$

- In contrast the error $(y_p - \hat{y})$ in predicting some future value of y is the sum of 2 errors: the error of estimating the mean of y, E[y] plus the random error that is a component of the value of y to be predicted

- 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

2

2A $\Sigma x_i = 2150$ $\Sigma y_i = 1430$ $n=5$

$\bar{x} = 430$ $\bar{y} = 286$

$\Sigma x_i y_i = 618500$

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

$$b_1 = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}} \qquad \frac{618500 - \frac{2150(1430)}{5}}{931100 - \frac{(2150)^2}{5}} = \frac{6}{11}$$

$$b_0 = \bar{y} - \bar{x}b_1 \qquad 286 - \frac{6}{11}430 = 51.45$$

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

B $$MSE = \frac{SSE}{N-2} = \frac{\Sigma[y_i - \hat{y}_i]^2}{N-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2}$$

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

$$= 2920 - 21600/11$$

$$= \frac{10520}{33} = 318.787 = \sigma^2$$

C $$r^2 = \frac{S_{xy} \times S_{xy}}{S_{xx} S_{yy}} = \frac{(3600)^2}{(6000)(2910)} = 0.6724$$

- R^2 measures the proportion of total variance about the mean $\bar{y}$, explained by the regression
- R^2 explains 67% of the total variation in the data about the average $\bar{y}$, goes from $0 \to 1$

D. $.6724 = R^2$ is a strong figure, roughly 70% of error is due to variation about the mean μ.

- With each increase in one unit the mean value of distribution of sale price rises by 6/11.

202 Regression Exam Rev

2 E Assumptions:
- X_i is the i^{th} value of the predictor variable which is a known constant for all i.
- The observation y_i or ε_i are independent
- At any given X_i, y_i or ε_i is normally distributed
- The observed or ε_i have constant variance
- Mean of y_i can be joined by a straight line:
  $E(y_i) = \beta_0 + \beta_1 x_i$
  β_1 - Slope of regression
  β_0 - Intercept

Q3 A Test if there is a relationship between 2 variables
$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$
Test = -2.55
Prob = 0.018 less than α of 0.05, reject H_0, a relationship exists between the 2 variables

B Slope = -0.00231, The mean distribution value of Sweetness reduced by 0.00231 for each unit increase in perm

Intercept = 6.25, this is the value of the Sweetness index for a perm rate of 0.

C $6.25 - (0.00231)(210) = 5.67$ Sweetness index

D R^2 value of 22.4 suggests a bad fit for the model! Proportion of total variability about the mean is explained by regression $r^2 = 0.23$ implying a bad fit -> unreliable prediction

Regression 202 Exam Paper

1 B.

Variable	N	mean	St dev	SE Mean	95% CI
Benzene	20	2·1	1·552	0·347	(1·374, 2·826)

- Sample size $N=20$
- The mean of the sample is 2·1

$SE\ mean = \frac{1.552}{\sqrt{n}} = 0.347$

95% of time the true mean lies in (1·374, 2·826)
This is ceded by mean $\pm$ t(0.05, 19) SE(mean)

D Confidence versus prediction interval

CI: $\hat{y} \pm t(n-2, 1-\frac{1}{2}\alpha)\ s \cdot \sqrt{\frac{1}{n} + \frac{(x_p-\bar{x})^2}{\sum (x_i-\bar{x})^2}}$

PI: $\hat{y} \pm t(n-2, 1-\frac{1}{2}\alpha)\ s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p-\bar{x})^2}{\sum (x_i-\bar{x})^2}}$

- CI used model for estimating the mean value of y E(y) for a specific value of x.
- 2nd model PI, predicts a particular y value for a given x. Predicts outcome of single experiment given-x value

CI is narrower why?
- The error in estimating the mean value of y, E[y], for a given x_p, is the distance between the least squared line and the true line of means, $E[y] = \beta_0 + \beta x_i$
- Error shown in figure $[\hat{y} - E(y)]$

- In contra, the error $(y_p - \hat{y})$ in predicting some future value of y is the sum of 2 errors: the error of estimating the mean of y, E[y] plus the random error that is a component of the value of y to be predicted

- 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 lied from $\bar{x}$, the larger will be the error of estimation and prediction

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

$n = 5$

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

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

$\Sigma x_i y_i = 618500$

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

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

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

$b_0 = 286 + \frac{6}{11}(430) \quad = 51.45\overline{45}$

51.45

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

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

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

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

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

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

$= \frac{956.363}{3} = 318.787 = \frac{1052.0}{33} = s^2$

c $r^2 = \frac{S_{xy} \times S_{xy}}{S_{xx} \times S_{yy}} = \frac{(3600)^2}{(6600)(2920)} \qquad \frac{12960000}{9520 \; 19272000} = \frac{540}{803} \therefore 0.672$

3 Regression 202 Exam Page

Q7. R^2 mean proporan of total varde about te mean mean $\bar{y}$, explaed by the regression

r^2 explains GY of the total variation in the dde about the average $\bar{y}$. Goes for $0 \to 1$.

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

Q8 67.24 $= R^2$ is a decent figure, roughly 70% of error is due to variation about te mar J.

With each increase in male value, the mean value for distribution of Sale price rise) by 6/11, This to increase the sale price by one, the marker 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.
- 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 at per data

b) Slope is -0.00231, for each unit increase in pectin, the mean value distribution of sweetness decreased 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 suggest a bad fit for the model. Proportion of total variability about the mean is explained about the regression line = 023, implying a bad fit → unreliable results

d) ———