classmate
Date
Page

P & S - II Assignment - 2

Q1 a) $\hat{y_i} = \hat{\alpha_i} + \hat{\beta} x_i$

b) Gradient of regression line $= \hat{\beta} = 1.724197531$

Q2 i $\hat{\beta} = \frac{S_{XY}}{S_{XX}}$ and $\hat{\alpha_i} = \bar{y} + \hat{\beta}\bar{x}$

So $S_{XX} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 12666.58 - \frac{(385.2)^2}{12} = 301.66$

$S_{YY} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 119026.9 - \frac{(1162.5)^2}{12} = 6409.7125$

$S_{XY} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 38191.41 - \frac{(385.2)(1162.5)}{12}$

$= 875.16$

$\hat{\alpha_i} = 3.748181$ $\hat{\beta} = 2.901146$

$\Rightarrow$ $\hat{y_i} = \cancel{\hat{\alpha_i}}\ 3.748181 + 2.901146 x_i$

ii $PQ = \frac{\hat{\beta} - \beta}{se(\hat{\beta})} \sim t_{n-2}$

$\Rightarrow se(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}$

$\Rightarrow$ $\hat{\sigma}^2 = \frac{1}{10}\left[6409.7125 - \frac{(875.16)^2}{301.66}\right] = 387.0744$

$\Rightarrow$ $se(\hat{\beta}) = 1.132$

$$\Rightarrow P\left[-2.228 < \frac{2.901-\beta}{1.132} < 2.228\right] = 0.95$$

$$\Rightarrow P\left[-2.228(1.132)-2.901 < -\beta < 2.228(1.132)-2.901\right] = 0.9$$

$$\Rightarrow 95\% \text{ CI for } \beta = \{2.901 - 2.228(1.132),\ 2.901+2.228(1.132)\}$$

$$= \{0.378904,\ 5.423096\}$$

iii

$$P\phi = \frac{\hat{\mu_i} - \mu_i}{Se(\hat{\mu_i})} \sim t_{n-2}$$

$$\Rightarrow \hat{\mu_i} = 3.748181 + 2.901146(40)$$

$$= 119.79$$

$$\Rightarrow Se(\mu_i) = \sqrt{\left[\frac{1}{12} + \frac{(7.9)^2}{301.66}\right]38707} = 10.5988$$

$$\Rightarrow P\left[-2.228 < \frac{119.79-\mu_i}{10.5988} < 2.228\right] = 0.95$$

$$\Rightarrow 95\% \text{ CI for } \mu_i = \{119.79 - 2.228(10.5988),\ 119.79+2.228(10.5$$

$$= \{96.17,\ 143.40\}$$

Q3 i Comments on the plot

- The center of the distribution differ for all the 4 cities. Thus, there is a prime facie for suggesting that the underlying means are different

- The difference b/w the mean time taken to commute to office in peak hours & non-peak hours are in the order City A to city D

- The variation in the data for city C is largest compared to City D appears to be highest. However with only 7 observation for each city, we can't be sure that there is a real underlying difference in variance

ii Assumptions underlying analysis of variance:

- The population must be normal
- The population have a common variance
- The observation are independent

iii
H0: The mean of differences is same for each city
H1: The mean of difference is not the same

$$SSTOT = S_{YY} = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n} = 1495 - \frac{103^2}{28}$$

$$= 1116.11$$

$$SSREG = \frac{1}{7}\left(64^2 + 44^2 + 8^2 + (-13)^2\right) - \frac{103^2}{28} = 516.11$$

$$\Rightarrow SSRES = SSTOT - SSREG = 600$$

ANOVA table

Source of Variation	df	Sum of sqrs.	Mean sum of sqrs
Regression	3	516.11	172.04
Residual	24	600	25
Total	27	1116.11	

Under H_0, $F = \frac{MSSREG}{MSSRES} = \frac{172.04}{25} = 6.88$

Test Statistic = 6.88

$F_{3,24,5\%} = 3.009$ {Right tailed test}

Hence we have sufficient evidence to reject H_0 at LS of 5%.
Thus, there are underlying difference b/w the cities.

IV Analysis of mean differences

Since, $\bar{y}_{1*} = 9.14$
$\bar{y}_{2*} = 6.29$
$\bar{y}_{3*} = 1.14$
$\bar{y}_{4*} = -1.86$

$\bar{y}_{1*} > \bar{y}_{2*} > \bar{y}_{3*} > \bar{y}_{4*}$

classmate

Date
Page

$\Rightarrow \hat{\sigma}^2 = \frac{SS_{RES}}{n-2} = 25$

The least significance difference b/w any of means is

$t_{24, 0.025} \times \hat{\sigma}^2 \sqrt{\frac{1}{7}+\frac{1}{7}} = 2.064 \times \sqrt{25} \times \sqrt{\frac{1}{7}+\frac{1}{7}} = 5.52$

Now, we can examine the difference b/w each of the pairs of means. If the difference is less than the level of significance, then there is no significant difference b/w the means.

We have,

$\bar{y}_{1*} - \bar{y}_{2*} = 2.85$
$\bar{y}_{2*} - \bar{y}_{3*} = 5.15$
$\bar{y}_{3*} - \bar{y}_{4*} = 3$

All the three difference are less than 5.52, we will these pairs to show that they have no significant difference

$\bar{y}_{1*} > \bar{y}_{2*} > \bar{y}_{3*} > \bar{y}_{4*}$

Examining to see if the first 2 groups can be combined,

$\bar{y}_{1*} - \bar{y}_{2*} = 8$

There is significant diff. b/w mean 12.3 so we can't combine the first two groups.

Examining to see if ~~least~~ last 2 groups can be combined

$\bar{y}_{2*} - \bar{y}_{4*} = 8.15$

Hence we can't combine the last 2 groups

Therefore, the diagram remains as before.

Q4 a) The mathematical model of one-way ANOVA is given by,

$y_{ij} = \mu + \tau_i + e_{ij}$, $i = 1, 2, \ldots, k$

$j = 1, 2, \ldots, n_i$

where,

k = number of treatments

n_i = no. of responses from i^{th} treatment

y_{ij} = j^{th} response from i^{th} treatment

τ_i = is the i^{th} treatment

μ = is the over mean response

e_{ij} = is the error term.

Assumption: e_{ij} are iid $\sim N(0, \sigma^2)$

b) H_0: Mean claim amounts of five companies are equal

H_1: Mean claim amounts of five companies are not equal

$n_1 = 7$ $n_4 = 7$

$n_2 = 8$ $n_5 = 5$

$n_3 = 6$ ~~N~~ $n = 33$

$$\sum_{i=1}^{5} \sum_{j=1}^{n_i} y_{ij} = 1856$$

$$\sum_{i=1}^{5} \sum_{j=1}^{n_i} y^2_{ij} = 174316$$

$$CF = \left\{ \sum_{i=1}^{5} \sum_{j=1}^{m_j} y_{ij} \right\}^2 \Big/ n = 104385.94$$

$SS_{TOT} = 174316 - CF$
$= 69930.06$

$$SS_{REG} = \frac{354^2}{7} + \frac{386^2}{8} + \frac{87^2}{6} + \frac{643^2}{7} + \frac{384^2}{3} - 104385.94$$

$= 22325.69$

$SS_{RES} = SS_{TOT} - SS_{REG} = 47604.37$

Sources of Variation	df	Sum of Sqrs	Mean sum of sqrs.
Regression	4	22326	5581
Residual	28	47604	1700
Total	32	69930	

$$\text{Test Statistic} = \frac{MSS_{REG}}{MSS_{RES}} = 3.283$$

$F_{4, 28, 0.05} = 2.714$

We have sufficient evidence to reject H_0

c)

Oi

	3-5	5-8	8-10	10-12	
0-3	45	20	6	5	= 76
4-6	7	20	9	6	= 42
7-9	5	8	15	12	= 40
	57	48	30	23	158

Ei

	3-5	5-8	8-10	10-12
0-3	27.41772	23.08861	14.43038	11.06329
4-6	15.15189	12.75949	7.97468	6.11392
7-9	14.43038	12.15189	7.59494	5.82278

No clubbing for Ei's is required as all Ei's are greater than 5

Contingency Table

H_0: Sales are independent
H_1: Sales are dependent

$$TS = \sum \frac{(O_i - E_i)^2}{E_i} = 49.91146$$

degree of freedom = $2 \times 3 = 6$

$\Rightarrow$ $\chi^2_{6, 5\%} = 12.59$

We have sufficient evidence to reject H_0

classmate
Date
Page

Q5 i Assumptions underlying ANOVA
- Populations are normal
- Population have common variance
- Observation are independent

The sample variance are very different from each other
Hence, the common variance assumption won't hold

ii Mean at Rate 1 for $\sqrt{x}$ $= \frac{\sqrt{29} + \sqrt{13} + \sqrt{21}}{3}$

$= 4.52443$

Variance at Rate 2 for $\log_e(x)$ $= \frac{\sum(x_i - \bar{x})^2}{n-1} = 0.12127$

iii The test was correct asserting that the large transformation must be done before carrying out one-way ANOVA as for the transformation it can be claimed that the assumption of common variance for the underlying proportion holds.

IV-a) $y_{ij} = \mu + T_i + e_{ij}$, $i = 1, 2, 3, 4$ $j = 1, 2, 3$

- y_{ij} is the $\log_e$ transformed value of the j^{th} observation of the no. of germination per square food observed when the i^{th} rate was applied

- μ is the overall pop^n mean
- T_i is the deviation of the i^{th} rate mean such that $\sum T_i = 0$

- e_{ij} are the independent error terms
$e_{ij} \sim N(0, \sigma^2)$

b) For ANOVA

$H_0: \tau_i = 0$

$H_1: \tau_i \neq 0$

lm(rw)

Rate	Mean	Variance	Y_i^2
1	2.992	0.163	80.582
2	4.827	0.121	209.678
3	5.716	0.186	294.053
4	6.575	0.148	389.063
Total =	20.11	0.618	973.373

Here, $Y_i^2 = \left(\sum_{j=1}^{3} Y_{ij}\right)^2 = (3 \times \text{Mean}_i)^2$

$$SS_{Rate} = \sum_{i=1}^{4} \frac{Y_i^2}{3} - \frac{Y^2}{12} = 21.153$$

$$SS_{RES} = \sum_{i=1}^{4} \left\{ \sum_{j=1}^{3} (Y_{ij} - \bar{Y}_i)^2 \right\}$$

$$= \sum_{i=1}^{4} \{2 \times \text{variance}\}$$

$$= 2 \times 0.618$$

$$= 1.236$$

Source of Variation	d.f.	SS	MSS
Rate	3	21.153	7.051
Residuals	8	1.236	0.155
	11	22.389	

$TS = \frac{MSS_{rate}}{MSS_{res}} = 45.638$

$F_{3,8,1\%} = 7.951$

⇒ Given the observed F statistic value is much larger than this, we can tele the P-value for this test is almost near to 0 or in other words there is overwhelming evidence against H_0. Thus it can be concluded that the underlying means are not equal.

6) $\Sigma X = 39$, $\Sigma X^2 = 207$, $\Sigma Y = 562$, $\Sigma Y^2 = 40508$,
$\Sigma XY = 2853$, $n = 10$

i) $S_{XX} = 207 - \frac{39^2}{10} = 54.9$

$S_{yy} = 40508 - \frac{562^2}{10} = 8923.6$

$S_{xy} = 2853 - \frac{562 \times 39}{10} = 661.2$

$r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} = \frac{661.2}{\sqrt{54.9 \times 8923.6}} = 0.94466$

ii) $\hat{B} = \frac{S_{XY}}{S_{XX}} = \frac{661.2}{54.9} = 12.04371$

$\hat{\alpha} = \bar{Y} - \hat{B}\bar{X} = 56.2 - 12.04371 \times 3.9$
$= 9.2295$

iii) $SS_{TOT} = SS_y = 8923.6$

$SS_{REG} = \frac{S^2_{XY}}{S_{XX}} = \frac{661.2^2}{54.9} = 7963.30492$

$SS_{RES} = SS_{TOT} - SS_{REG} = 960.29508$

iv) $R^2 = \frac{S^2_{xy}}{S_{xx} \cdot S_{yy}} = 89.24\%$

Comment →

1. Model is a good fit as it explains 89% of the variability in the output.

2.

classmate

Date
Page

7) i) $\Sigma X = 174.13$, $\Sigma Y = 197.84$, $\Sigma X^2 = 7545.9$,

$\Sigma Y^2 = 4747.45$, $\Sigma(X-\bar{X})(Y-\bar{Y}) = 2176.84$.

$n = 11$

$$S_{XX} = 7545.9 - \frac{174.13^2}{11}$$

$$= 4789.4221$$

$$S_{yy} = 4747.45 - \frac{197.84^2}{11}$$

$$= 1189.20767$$

$$S_{xy} = 2176.84$$

$$\hat{B} = \frac{S_{xy}}{S_{xx}} = 0.45451$$

$$\hat{\alpha} = \bar{Y} - \hat{B}\bar{X} = 10.78056$$

$$\hat{Y} = 10.78056 + 0.45451 X$$

ii) $H_0 : B = 0$, $H_1 : B \neq 0$

$$T.S \rightarrow \frac{\hat{B} - B}{S_E(\hat{B})} \sim t_{n-2}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 22.20136$$

$$S_E(\hat{B}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

classmate
Date
Page

$$T.S = \frac{0.45451 - 0}{\sqrt{\frac{22.20136}{4789.4221}}} \sim t_{9}$$

$6.6755 \sim t_9$

$t_{9, 2.5\%} = 2.262$

$\therefore$ We have sufficient evidence to Reject H_0 at 5% LS
$\therefore$ $B \neq 0$, hence there is a linear relationship.

Assumptions $\rightarrow$

① Pop^n have a common Variance
② Pop^n follow the Normal dist^n.
③ The observations are independent.

iii) $$r = \frac{S_{xy}}{\sqrt{S_{xx}.S_{yy}}} = 0.91213$$

iv) $X_0 = 25$

Mean Response $(\mu_0) \rightarrow$ P.Q $\rightarrow$ $\frac{\hat{\mu} - \mu}{S_E(\hat{\mu})} \sim t_{n-2}$

$$SE(\hat{\mu}) = \sqrt{\left(\frac{1}{n} + \frac{(X_0 - \bar{x})^2}{S_{xx}}\right)\hat{\sigma}^2} \quad \left(\hat{\mu} = \hat{\alpha} + \hat{B} \times 25\right)$$

$$SE(\hat{\mu}) = \sqrt{\left(\frac{1}{11} + \frac{(25 - 174.13/11)^2}{4789.4221}\right) \times 22.02136} = 1.55$$

95% CI for $\mu_0 = \hat{\mu} \pm t_{9, 2.5\%} \times S.E(\hat{\mu})$

$= 22.15331 \pm 3.51018$

$= (18.64313, 25.66349)$

For Individual Response,

$\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0$

$= 22.15331$

$$S.E(\hat{y}_0) = \sqrt{\left[\frac{1}{11} + 1 + \left(\frac{(25 - 174.131111)^2}{4789.4221}\right)\right] \times 22.702136}$$

$= 4.96079$

$$\frac{\hat{y}_0 - y_0}{SE(\hat{y}_0)} \sim t_9$$

95% CI for $y_0 = \hat{y} \pm t_{9, 2.5\%} \times SE(\hat{y}_0)$

$= 22.1531 \pm 11.22131$

$= (10.932, 33.3746)$

v) There is clearly some relationship b/w residuals, and the explanatory variable values. Hence, our assumption of normality might be wrong. ∴ The model is not a good fit for the data.

Date
Page

8) i) The main purpose of factor analysis is to reduce many individual items into a fewer no. of dimensions.

Factor Analysis can be used to simplify data, such as reducing the no. of variables in regression model.

Principal Component Analysis is a technique for reducing the dimensionality of such data sets increasing interpretab-ility but at the same time minimising information loss.

While newly identified principle components are choosen to be →

1) Uncorrelated
2) Linear Combination of the original variables of the data.
3) Which maximise the variance.

ii)

Principal Component	Diagonal Entries	PCi / Sum (PCi) over 1 to 5
PC1	0.456	65%
PC2	0.137	19.5%
PC3	0.080	11.4%
PC4	0.0165	2.4%
PC5	0.012	1.7%
	0.7015	

iii) As the 1st 3 P.C explains over 95% of the variance, demonstrating can be reduced to these dataset. The 1st 3 P.C can be then used for building further classification or regression modelling purpose.

9) $\sum X = 45,\ \sum X^2 = 277.5,\ \sum Y = 644,\ \sum Y^2 = 43956,$
$\sum XY = 2602,\ n = 10$

i) There appears to be a negative correlation b/w the marks scored and hours spent per day on social media.

ii) $S_{XX} = 277.5 - \dfrac{45^2}{10} = 75$

$S_{YY} = 43956 - \dfrac{644^2}{10} = 2482.4$

$S_{XY} = 2602 - \dfrac{644 \times 45}{10} = -296$

$r = \dfrac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} = -0.686002$

Hence, we have a weak -ve correlation.

iii) $H_0 : \rho = 0,\quad H_1 : \rho \overset{<}{\cancel{\neq}} 0$

$\tanh^{-1}(r) - \tanh^{-1}(\rho) \sim N\left(0, \dfrac{1}{n-3}\right)$

$T.S \rightarrow \dfrac{\tanh^{-1}(r) - \tanh^{-1}(\rho)}{\sqrt{1/n-3}} \sim N(0,1)$

$= -2.07893$

$Z_{5\%} = 1.6448$

$|Z\ cal| > Z_{tab}$

Hence, we have sufficient evidence to Reject H_0.

$\Rightarrow \beta < 0.$

iv) $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = -3.9467$

$\hat{\alpha} = \bar{Y} - \bar{X}.\hat{\beta} = 82.16$

$\hat{y} = 82.16 - 3.9467X$

v) $R^2 = \frac{S_{xy}^2}{S_{xx} \times S_{yy}} = 47.0598\%$

Hence model explain only 47% of the variability
The model is not a good fit for the data.

vi) $X_0 = 1$

Expected Change $= \hat{\beta}x_0$
$= -3.9467 \times 1$
$= -3.9467$

$\boxed{\text{additional}}$

$\therefore$ For every 1 hr spent on social media, the expected ~~change~~ decrease in marks is 3.9467.

10) $\sum X = 0.93$, $\sum X^2 = 0.2925$, $n = 3$
$\sum Y = 0.23$, $\sum Y^2 = 0.0189$, $\sum XY = 0.0735$

$$S_{XX} = 0.2925 - \frac{0.93^2}{3} = 0.0042$$

$$S_{YY} = 0.0189 - \frac{0.23^2}{3} = 0.00126$$

$$S_{XY} = 0.0735 - \frac{0.93 \times 0.23}{3} = 0.0022$$

$SS_{TOT} = S_{YY} = 0.00126$

$SS_{REG} = S_{XY}^2 / S_{XX} = 0.00115$

$SS_{RES} = 0.00011$

ANOVA Table →

Sources of Var.	D.o.F	SS	MSS
Regression	1	0.00115	0.00115
Residual	1	0.00011	0.00011
Total	2	0.00126	

$H_0 : B = 0$, $H_1 : B \neq 0$

T.S → $\dfrac{0.00115}{0.00011} \sim F_{1,1} = 10.4545 \sim F_{1,1}$

$F_{1,1,5\%} = 161.4$, $F_{cal} < F_{tab}$, ∴ We do not reject H_0
∴ Industry has no effect on resignation rate.