

$$Q1 a) \hat{y}_i = \hat{\alpha}_i + \hat{\beta} x_i$$

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

$$Q2 i) \hat{\beta} = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\alpha}_i = \bar{y} - \hat{\beta} \bar{x}$$

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

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

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 38191.41 - \frac{(385.2)(1162.5)}{12} = 875.16$$

$$\hat{\alpha}_i = 3.748181 \quad \hat{\beta} = 2.901146$$

$$\Rightarrow \hat{y}_i = \hat{\alpha}_i + \hat{\beta} x_i = 3.748181 + 2.901146 x_i$$

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

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

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

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

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

$$\Rightarrow P[-2.228(1.132) - 2.901 < -\beta < 2.228(1.132) - 2.901] = 0.95$$

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

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

$$\text{iii} \quad PQ = \frac{\hat{\mu}_i - \mu_i}{Se(\hat{\mu}_i)} \sim t_{n-2}$$

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

$$= 119.79$$

$$\Rightarrow Se(\mu_i) = \sqrt{\left[\frac{1}{n-2} + \frac{(7.9)^2}{301.66} \right] 387.07} = 10.5988$$

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

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

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

03

Comments on the plot

- The center of the distribution differ for all the 4 cities. Thus, there is a prime face 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

H_0 : The mean of differences is same for each city
 H_1 : The mean of difference is not the same

$$SSTOT = S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 1495 - \frac{103^2}{28} = 1116.11$$

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

$$SSRES = SSTOT - SSREG = 600$$

ANOVA table

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

$$\text{Under } H_0, F = \frac{MSS_{REG}}{MSS_{RES}} = \frac{172.04}{25} = 6.88$$

$$\text{Test Statistic} = 6.88$$

$$F_{3,24, 5.1.} = 3.009 \quad \{ \text{Right tailed test} \}$$

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

IV

Analysis of mean differences

$$\text{Since, } \bar{y}_{1*} = 9.14$$

$$\bar{y}_{2*} = 6.29$$

$$\bar{y}_{3*} = 1.14$$

$$\bar{y}_{4*} = -1.86$$

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

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

The least significance difference b/w any of means is

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

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,

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

$$\bar{y}_{2*} - \bar{y}_{3*} = 5.5$$

$$\bar{y}_{3*} - \bar{y}_{4*} = 3$$

All the three difference are less than 5.52, we will treat 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 ~~last~~ ^{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.

Q4a) The mathematical model of one-way ANOVA is given by,

$$Y_{ij} = \mu + T_i + e_{ij}, \quad \begin{matrix} i = 1, 2, \dots, k \\ j = 1, 2, \dots, m_i \end{matrix}$$

where,

k = number of treatments

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

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

T_i = is the i^{th} treatment

μ = is the over mean response

e_{ij} = is the error term.

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

b) H_0 : Mean claim amounts of five companies are equal
 H_1 : Mean claim amounts of five companies are not equal

$$m_1 = 7$$

$$m_4 = 7$$

$$m_2 = 8$$

$$m_5 = 5$$

$$m_3 = 6$$

$$m = 33$$

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

$$\sum_{i=1}^5 \sum_{j=1}^{m_i} y_{ij}^2 = 174316$$

$$CF = \left\{ \sum_{i=1}^5 \sum_{j=1}^{m_i} y_{ij} \right\}^2 / m = 104385.94$$

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

$$SS_{REG} = \frac{354^2}{7} + \frac{386^2}{8} + \frac{87^2}{6} + \frac{643^2}{7} + \frac{384^2}{3} - 104385.94 \\ = 22325.69$$

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

Sources of Variation	df	Sum of Sqr	Mean sum of sqs.
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)

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

 E_i

	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 E_i 's is required as all E_i '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 = $2 \times 3 = 6$

$\Rightarrow \chi^2_{0.05, 6} = 12.59$

We have sufficient evidence to reject H_0

Q5i 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

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

for $\sqrt{m}$

$$= 4.52443$$

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

for $\log_e(m)$

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) $Y_{ij} = \mu + T_i + e_{ij}, \quad i=1,2,3,4 \quad j=1,2,3$

- Y_{ij} is the loge 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 popⁿ 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(m)

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

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

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

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

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

$$= 2 \times 0.618$$

$$= 1.236$$

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

$$TS = \frac{MSS_{\text{Rate}}}{MSS_{\text{res}}} = 45.638$$

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

c) Given the observed F statistic value is much larger than this, we can tell 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.

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

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

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

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

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{661.2}{\sqrt{54.9 \times 8923.6}} = 0.94466$$

$$ii) \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.2}{54.9} = 12.04371$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} = 56.2 - 12.04371 \times 3.9 = 9.2295$$

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

$$SS_{REG} = \frac{S_{xy}^2}{S_{xx}} = \frac{661.2^2}{54.9} = 7963.30492$$

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

Date _____
Page _____

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

Comment →

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

$$7) i) \sum X = 174.13, \sum Y = 197.84, \sum X^2 = 7545.9, \\ \sum Y^2 = 4747.45, \sum (X - \bar{X})(Y - \bar{Y}) = 2176.84, \\ n = 11$$

$$S_{xx} = 7545.9 - \frac{174.13^2}{11} \\ = 4789.4221$$

$$S_{yy} = 4747.45 - \frac{197.84^2}{11} \\ = 1189.20767$$

$$S_{xy} = 2176.84$$

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

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

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

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

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

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

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

$$T.S = \frac{0.45451 - 0}{\sqrt{\frac{22.20136}{4789.4221}}} \sim t_{n-2}$$

$$6.6755 \sim t_9$$

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

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

Assumptions $\rightarrow$

- ① Popⁿ have a common variance
- ② Popⁿ follow the Normal distⁿ.
- ③ The observations are independent.

$$\text{iii)} \quad r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = 0.91213$$

$$\text{iv)} \quad X_0 = 25$$

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

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

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

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 interpretability but at the same time minimising information loss.

While newly identified principle components are chosen to be $\rightarrow$

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

ii)

Principal Component	Diagonal Entries	$PC_i / \text{Sum}(PC_i)$ over 1 to 5
PC 1	0.456	6.5%
PC 2	0.137	19.5%
PC 3	0.080	11.4%
PC 4	0.0165	2.4%
PC 5	0.012	1.7%
	<u>0.7015</u>	

iii) As the 1st 3 P.C explains over 95% of the variance, demonstrating can be reduced to this 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) $S_{xx} = 277.5 - \frac{45^2}{10} = 75$

$$S_{yy} = 43956 - \frac{644^2}{10} = 2482.4$$

$$S_{xy} = 2602 - \frac{644 \times 45}{10} = -296$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = -0.686002$$

Hence, we have a weak -ve correlation.

iii) $H_0: \rho = 0, H_1: \rho \neq 0$

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

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

$$= -2.07893$$

$$Z_{s1} = 1.6448$$

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

Hence, we have sufficient evidence to Reject H_0 .

$$\Rightarrow \rho < 0.$$

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

$$\hat{\alpha} = \bar{y} - \bar{x} \cdot \hat{\beta} = 82.16$$

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

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

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

$$vi) X_0 = 1$$

$$\begin{aligned} \hat{\beta} \times X_0 &= \text{Expected Change} = \hat{\beta} X_0 \\ &= -3.9467 \times 1 \\ &= -3.9467 \end{aligned}$$

additional

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

Date _____
Page _____

$$\begin{aligned}
 95\% \text{ CI for } \mu_0 &= \hat{\mu} \pm t_{9, 2.5\%} \times S.E(\hat{\mu}) \\
 &= 22.15331 \pm 3.51018 \\
 &= (18.64313, 25.66349)
 \end{aligned}$$

For Individual Response,

$$\begin{aligned}
 \hat{y}_0 &= \hat{\alpha} + \hat{\beta}x_0 \\
 &= 22.15331
 \end{aligned}$$

$$\begin{aligned}
 S.E(\hat{y}_0) &= \sqrt{\left[\frac{1}{11} + 1 + \frac{(25 - 174.13/11)^2}{4789.4221} \right]} \times 22.202136 \\
 &= 4.96079
 \end{aligned}$$

$$\frac{\hat{y}_0 - y_0}{S.E(\hat{y}_0)} \sim t_9$$

$$\begin{aligned}
 95\% \text{ CI for } y_{00} &= \hat{y} \pm t_{9, 2.5\%} \times S.E(\hat{y}_0) \\
 &= 22.1531 \pm 11.22131 \\
 &= (10.932, 33.3746)
 \end{aligned}$$

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

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 $\rightarrow$

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

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

$$T.S \rightarrow \frac{0.00115}{0.00011} \sim F_{1,1} = 10.4545 \sim F_{1,1}$$

$$F_{1,1,5\%} = 161.4, F_{cal} < F_{tab}, \therefore \text{We do not reject } H_0$$

$\therefore$ Industry has no effect on resignation rate