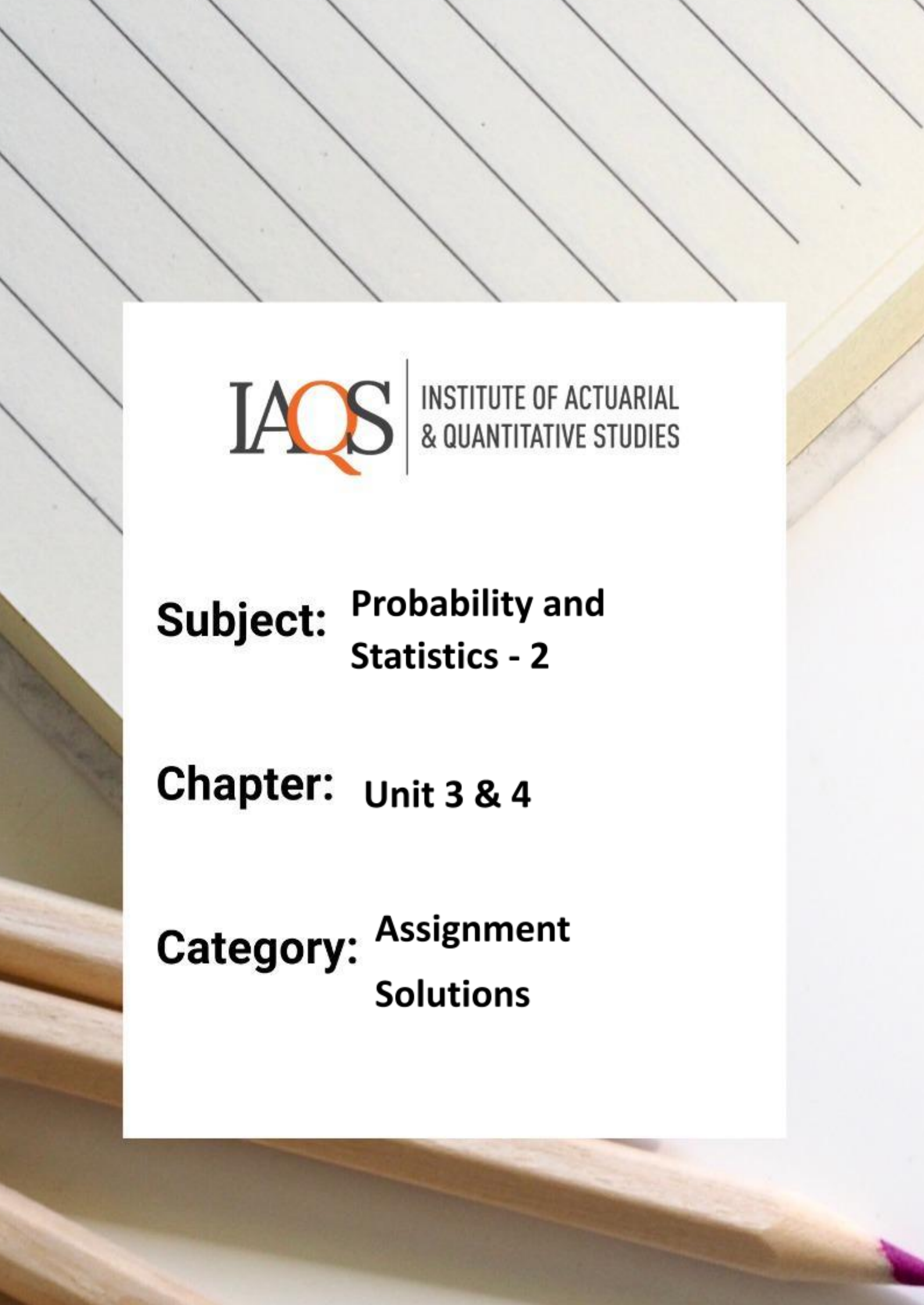

INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: Probability and Statistics - 2

Chapter: Unit 3 & 4

Category: Assignment Solutions

1.

	1	2	3	4	5	6	7	8	9	10	Total
x	40	10	100	110	120	150	20	90	80	130	**850**
y	56	62	195	240	170	270	48	196	214	286	**1,737**
xy	2,240	620	19,500	26,400	20,400	40,500	960	17,640	17,120	37,180	**182,560**
x^2	1,600	100	10,000	12,100	14,400	22,500	400	8,100	6,400	16,900	**92,500**

$S_{xy} = \sum x_i y_i - \sum x_i \sum y_i / n = 182560 - 850 * 1737/10 = 34915$

$S_{xx} = \sum x_i^2 - (\sum x_i)^2 / n = 92500 - 850^2/10 = 20250$

$\hat{b} = S_{xy} / S_{xx} = 1.72$

$\hat{a} = \bar{y} - b\bar{x} = (1737/10) - 1.72 * (850/10) = 27.14$

$y = 27.14 + 1.72x$

(b) Gradient represents the amount of hours per rupee spent

2.

i. Fitted Linear Regression Equation

The relevant summary statistics to fit the equation are:

$\sum x$ = 385.2; $\sum x^2$ = 12,666.58;
$\sum y$ = 1,162.5; $\sum y^2$ = 119,026.9;
$\sum xy$ = 38,191.41; n = 12.

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 12666.58 - 12 * \left(\frac{385.2}{12}\right)^2 = 301.66$$

$$S_{xy} = \sum xy - n\overline{xy} = 38191.41 - 12 * \left(\frac{385.2}{12}\right)\left(\frac{1162.5}{12}\right) = 875.16$$

$$S_{yy} = \sum y^2 - n\bar{y}^2 = 119026.90 - 12 * \left(\frac{1162.5}{12}\right)^2 = 6409.71$$

The coefficients of the regression equation are:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{875.16}{301.66} = 2.90$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x} = \left(\frac{1162.5}{12}\right) - 2.90 * \left(\frac{385.2}{12}\right) = 3.78$$

Therefore, the fitted regression line is: $\mathbf{y = \hat{\alpha} + \hat{\beta} x = 3.78 + 2.90\,x}$

ii. Confidence interval for β

Assuming normal errors with a constant variance:

95% confidence interval for β: $\hat{\beta} \pm t_{n-2}(2.50\%) * s.e.(\hat{\beta})$

$$\text{Here: } s.e.(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left[S_{yy} - \frac{{S_{xy}}^2}{S_{xx}}\right] = 387.07$$

$$s.e.(\hat{\beta}) = \sqrt{\frac{387.07}{301.66}} = 1.13$$

95% confidence interval for β: $2.90 \pm 2.228 * 1.13 = (0.38, 5.42)$

iii. <u>95% confidence intervals for the mean IBM share price</u>

$$\hat{y}_{x_0} \pm t_{n-2}(2.50\%)\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

The Dell Share price is US $ 40 (x_0).

$$\hat{y}_{x_0} = 3.78 + 2.90 * 40 = 119.78$$

Thus, 95% Confidence interval:

$$= 119.78 \pm 2.228 * \sqrt{387.07 * \left[\frac{1}{12} + \frac{(40 - 32.1)^2}{301.66}\right]}$$

$$= 119.78 \pm 2.228 * 10.5989$$

$$= (96.17, 143.39)$$

3.
i) Comments on the plot
The centers of the distributions differ for all the four cities. Thus there is a prima facie case for suggesting that the underlying means are different.

The difference between the mean time taken to commute to office in peak hours and nonpeak hours are in the order City A (highest), City D (lowest).

The variation in the data for City C is *lowest* compared to City D which appears to be *highest*. However, with only 7 observations for each city, we cannot be sure that there is a real underlying difference in variance.

ii. Following are the assumptions underlying analysis of variance:
The populations must be **normal**.
The populations have a **common variance**.
The observations are **independent**.

iii) We are carrying out the following test:
H0: The mean of differences is same for each city
against
H1: The mean of differences are not the same for all of the cities

To carry out the ANOVA, we must first compute the Sum of Squares

$$SS_T = 1{,}495 - \frac{103^2}{28} = 1{,}116.11$$

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

$$SS_R = SS_T - SS_B = 600.00$$

The ANOVA table is:

Source	df	SS	MS	F
Treatments	3	516.11	172.04	6.88
Residual	24	600.00	25.00	
Total	27	1,116.11		

Under $H_0, F = \frac{172.04}{25.00} = 6.88$, using the $F_{3,24}$ distribution.

The 5% critical point is 3.009, so we have sufficient evidence to reject H0 at the 5% level.

Therefore it is reasonable to conclude that there are underlying differences between the cities.

iv. Analysis of the mean differences

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

we can write:

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

$$\hat{\sigma}^2 = \frac{SS_R}{n-k} = 25$$

The least significant difference between any pair of means is:

$$t_{24,0.025} * \hat{\sigma} \sqrt{\left(\frac{1}{7} + \frac{1}{7}\right)} = 2.064 * \sqrt{25} * \sqrt{\left(\frac{1}{7} + \frac{1}{7}\right)} = 5.52$$

Now we can examine the difference between each of the pairs of means. If the difference is less than the least significant difference then there is no significant difference between the means.

We have

$\bar{y}_{1*} - \bar{y}_{2*} = 2.85;\ \bar{y}_{2*} - \bar{y}_{3*} = 5.15;\ \bar{y}_{3*} - \bar{y}_{4*} = 3.00$

Observing that all these 3 differences are less than 5.52, we underline these pairs to show that they have no significant difference:

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

Examining to see if the first two groups can be combined

$$\bar{y}_{1*} - \bar{y}_{3*} = 8.00$$

There is a significant between means 1 and 3, so we cannot combine the first two groups.

Examining to see if the last two groups can be combined:

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

There is a significant between means 2 and 4, so we cannot combine the last two groups.

Therefore the diagram remains as before.

4.

a)

The mathematical model of one-way ANOVA is given by

$$Y_{ij} = \mu + \tau_i + e_{ij};\ \text{i= 1,2,.....,k} \qquad \text{j = 1,2,.....,}n_i$$

where,

k = number of treatments

n_i = number of responses from i^{th} treatment

Y_{ij} is the 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} is i.i.d $N(0,\sigma^2)$

b)

H_0: Mean claim amounts of five companies are equal

H_1: Mean claim amounts of five companies are not equal

We have $n_1 = 7$, $n_2 = 8$, $n_3 = 6$, $n_4 = 7$, $n_5 = 5$, $n = 33$

$$\sum_{i=1}^{5}\sum_{j=1}^{ni} y_{ij} = 1{,}856$$

$$\sum_{i=1}^{5}\sum_{j=1}^{ni} y_{ij}^2 = 174{,}316$$

$$\text{C.F.} = \left(\sum_{i=1}^{5}\sum_{j=1}^{ni} y_{ij}\right)^2 / n = 104{,}385.94$$

SS_T = 174316- C.F. = 69,930.06;

$$SS_B = \sum_{i=1}^{5}\left(\sum_{j=1}^{ni} y_{ij}\right)^2 / n_i - C.F. = 354^2/7 + 386^2/8 + 87^2/6 + 645^2/7 + 384^2/5 - 104{,}385.94$$

= 22,325.69

$SS_R = SS_T - SS_B$ = 47,604.37

Sources of Variation	d.f	SS	MSS	F
Companies	4	22,326	5,581.42	3.283
Residual	28	47,604	1,700.16	
Total	32	69,930		

$F_{observed} = 3.283$; $F_{(4,\ 28,\ 0.05)} = 2.714$;

Reject H_0

c)

H_0: Salaries are independent of number of actuarial papers cleared

H_1: Salaries are dependent on number of actuarial papers cleared

Observed Values (O_i)

Papers cleared	Salary per annum (in Rs. lacs)				Total
	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
Total	57	48	30	23	158

Under H_0, Expected Values (E_i)

Papers cleared	Salary per annum (in Rs. lacs)				Total
	3 - 5	5 - 8	8 - 10	10 - 12	
0 - 3	27.42	23.09	14.43	11.06	76.00
4 - 6	15.15	12.76	7.97	6.11	42.00
7 - 9	14.43	12.15	7.59	5.82	40.00
Total	57.00	48.00	30.00	23.00	158.00

$$\chi^2 = \sum_{i=1}^{12} (O_i - E_i)^2 / E_i = (27.42 - 45)^2 / 27.42 + \ldots\ldots + (5.82 - 12)^2 / 5.82 = 49.919$$

$\chi^2_{Observed} = 49.91$ $\chi^2_{(6,\ 0.05)} = 12.59$,where d.f. 6 = (3-1) × (4-1)

Reject H_0

5.
[i] The assumptions required for one-way analyses of variance (ANOVA) are:
The populations must be **normal**
The populations have a **common variance**
The observations are **independent**. [1]
The sample variance observed for the four rates appear very different from each other. Thus, we can clearly see that the assumption that the underlying populations have a common variance assumption will not hold for the data as they are. [1]

[ii]

For the transformation $x \to \sqrt{x}$, the value of sample mean for rate 1 will be:

$$\frac{1}{3}(\sqrt{29} + \sqrt{13} + \sqrt{21}) = 4.52$$

[1]

For the transformation $x \to log_e x$, the value of sample variance for rate 2 will be

$$\frac{1}{2}[(log_e 180 - 4.827)^2 + (log_e 90 - 4.827)^2 + (log_e 120 - 4.827)^2] = 0.1213$$

[2]

[iii]
The scientist was correct in asserting that the loge transformation must be done before carrying out a one-way ANOVA as for this transformation it can be claimed that the assumption of common variance for the underlying population holds. [1]

To justify this, a quick check can be done on the ratio of maximum to minimum sample variance among the four rates data. A smaller ratio and close to 1 would indicate that the variances are close enough which in turn implies that the assumption of common variance for the underlying population holds

Variance	x	$\sqrt{x}$	$\log_e(x)$	$1/x$
Min	64	0.79	0.1213	0.0000004
Max	63,300	23.96	0.1861	0.0004721
Ratio	**989.06**	**30.17**	**1.53**	**1,268.14**

Clearly, the transformation $\log_e x$ produces the minimum ratio of maximum to minimum observed sample variance and that too close to 1. [1]

[iv]

We will perform an ANOVA on the $\log_e$ x data. We would assume the following model:

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

Here:

- Y_{ij} is the $\log_e$ transformed value of the j^{th} observation of the number of germinations per square foot observed when the i^{th} rate was applied
- μ is the overall population mean
- τ_i is the deviation of the i^{th} rate mean such that $\sum \tau_i = 0$
- e_{ij} are the independent error terms which follows Normal distribution with mean 0 and common unknown variance σ^2 [1]

We have already argued that we can assume equal underlying variances for this transformation. So, all requisite assumptions hold here.

For ANOVA, the null hypothesis being tested here is:

$$H_0: \tau_i = 0, \quad i = 1,2,3,4 \quad against \quad H_1: \tau_i \neq 0 \; for \; at \; least \; one \; i$$ [1]

To carry out the ANOVA, we must first compute the Sum of Squares. We have the following table using the information given in the question:

Rate	$\log_e(x)$		
	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	389.060
	20.110	**0.618**	**973.373**

Here: $Y_{i.}^2 = \left(\sum_{j=1}^{3} Y_{ij}\right)^2 = (3 * Mean_i)^2$

Now:

- $SS(\text{Rate}) = \sum_{i=1}^{4} \frac{Y_{i.}^2}{3} - \frac{Y_{..}^2}{12} = \frac{973.373}{3} - \frac{(3*20.110)^2}{12} = 21.153$
- $SS(\text{Residuals}) = \sum_{i=1}^{4} \left\{\sum_{j=1}^{3} \left(Y_{ij} - \bar{Y}_{i.}\right)^2\right\} = \sum_{i=1}^{4} \{2 * Variance_i\} = 2 * 0.618 = 1.236$

[3]

The ANOVA table is as follows:

Source of Variation	d.f.	Sum of Squares	Mean Squares	F
Rates	3	21.153	7.051	45.638
Residuals	8	1.236	0.155	
Total	**11**	**22.389**		

[2]

The 1% critical value for F (3, 8) distribution is 7.951.

Given the observed F statistic value is much larger than this, we can state the p-value for this test is almost near to zero or in other words there is overwhelming evidence against the null hypothesis H0. Thus it can be concluded that the underlying means are not equal. [1]

6.

i)

The relevant summary statistics to compute correlation coefficient are:

$$S_{xx} = \sum x^2 - n\bar{x}^2 = 207 - 10 * \left(\frac{39}{10}\right)^2 = 54.90$$

$$S_{xy} = \sum xy - n\overline{xy} = 2853 - 10 * \left(\frac{39}{10}\right)\left(\frac{562}{10}\right) = 661.20$$

$$S_{yy} = \sum y^2 - n\bar{y}^2 = 40508 - 10 * \left(\frac{562}{10}\right)^2 = 8923.60$$

$$\text{Correlation Coefficient } r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{661.20}{\sqrt{54.90}\sqrt{8923.60}} = 0.945$$

ii)

<u>Fitted Linear Regression Equation</u>

The coefficients of the regression equation are:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{661.20}{54.90} = 12.04$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x} = \left(\frac{562}{10}\right) - 12.04 * \left(\frac{39}{10}\right) = 9.23$$

Therefore, the fitted regression line is: $\mathbf{y = \hat{\alpha} + \hat{\beta}x = 9.23 + 12.04x}$

iii)

Relation: $SS_{TOT} = SS_{REG} + SS_{RES}$

$$SS_{TOT} = S_{yy} = 8923.60$$

$$SS_{RES} = S_{yy} - \frac{S_{xy}^{\ 2}}{S_{xx}} = 8923.60 - \frac{(661.20)^2}{54.90} = 960.30$$

$$SS_{REG} = S_{TOT} - S_{RES} = 8923.60 - 960.30 = 7963.30$$

iv)

Coefficient of Determination:

$$R^2 = \frac{S_{xy}^{\ 2}}{S_{xx}\,S_{yy}} = \frac{SS_{REG}}{SS_{TOT}} = \frac{7963.30}{8923.60} = 0.8924$$

For the simple linear regression model, the value of the coefficient of determination is the
square of the correlation coefficient for the data, since,

$$0.945 = r = \frac{S_{xy}}{(S_{xx} * S_{yy})^{0.5}} = \sqrt{R^2} = \sqrt{0.8924}$$

7.

i) $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 4789.42$

$S_{xy} = 2176.84$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{2176.84}{4789.42} = 0.4545$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 17.895 - 0.4545 * 15.83 = 10.79$

The fitted regression equation is $\hat{y} = 10.79 + 0.4545 * x$

ii) $S_{xx} = 4789.42$ from result of part i

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 1189.21$$

$S_{xy} = 2176.84$ from result of part i

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{1}{9} * \left(1189.21 - \frac{2176.84^2}{4789.42}\right) = 22.20 \quad [2]$$

$$s.e.(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{22.20}{4789.42}} = 0.0681 \quad [1]$$

To test $H_o: \beta = 0$ *v* $H_1: \beta \neq 0$, *the test statistic is*

$$\frac{\hat{\beta} - 0}{s.e.(\hat{\beta})} = \frac{0.4545}{0.0681} = 6.674 \quad [1]$$

Under the assumption that the errors of the regression are i.i.d $N(0, \sigma^2)$ random variables, beta has a t distribution with n-2 degrees of freedom. [0.5]

Critical value for t-distribution with 9 degrees of freedom: $t_{9,0.025} = 2.68$. [0.5]

Since the critical value at 95% level of significance is less than the test statistic, there is sufficient evidence to reject the null hypothesis. Hence, it cannot be concluded that there is no statistically significant relationship between x and y. [0.5]

[6]

iii) Pearson's correlation coefficient is computed as: $\frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$ [0.5]

$S_{yy} = 1189.21$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{2176.84}{\sqrt{4789.42 * 1189.21}} = 0.91 \quad [1.5]$$

iv) The estimated value of y corresponding to $x^3 = 25$ *is* $10.79 + 0.4545 * 25 = 22.15$ [1]

$\hat{\sigma}^2 = 22.20$ ... from earlier workings

The variance of the estimator of the mean response is given by

$\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = \left[\frac{1}{11} + \frac{84.0889}{4789.42}\right] * 22.20 = 2.41$ [2]

The variance of the estimator of the individual response is given by

$\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = [1 + 0.1085] * 22.20 = 24.61$ [2]

Using t_9 distribution, the 95% confidence intervals for mean and individual responses are:

22.20 ± 2.262 *sqrt(2.41) and 22.20 ± 2.262 *sqrt(24.61) = (18.7,25.7) & (11.0,33.4) [2]

[7]

v) The residual plot shows a definite pattern. Although the correlation coefficient is high, the model does not seem to be appropriate. [1]
Using this model leads to underestimation of premium rates at low and high mortality ratings. [1]
[2]

8. i) Factor analysis / Principal Component Analysis is - A method for reducing the dimensionality of data
It seeks to identify key components necessary to model and understand data [0.5]

Original variables may be
· correlated with each other [0.5]

While Newly identified principal components are chosen to be
· uncorrelated [0.5]
· linear combinations of the original variables of the data [0.5]
· which maximise the variance

ii)

Principal Component	Diagonal entry (PCi)	PCi/ (Sum(PCi) over 1 to 5)	
PC1	0.456	65.0%	of total variance explained by PC1
PC2	0.137	19.5%	of total variance explained by PC2
PC3	0.08	11.4%	of total variance explained by PC3
PC4	0.0165	2.4%	of total variance explained by PC4
PC5	0.012	1.7%	of total variance explained by PC5
	0.7015	**100.0%**	

Correct formula (1 mark)
Sum of PCi (0.5 marks)
Correct calculation (2.5 Marks)

iii) As 1st 3 Principal components explain over 95% of total variance, dimensionality can be reduced to 3 for this dataset [1]

The 1st 3 Principal Components can then be used for building further classification or regression modelling purpose [1]

[2]

9.

i) The scatter plot suggests an inverse relation between marks obtained and hours spent on social media per day [1]

ii) Sxx = 277.5 – 45^2/10 = 75
Syy = 43,956 – 644^2/10 = 2482.4
Sxy = 2,602 – 644*45/10 = -296
r = Sxy/$\sqrt{(Sxx * Syy)}$ = -0.686 [3]

-69% correlation co-efficient also implies a moderate negative linear relation between the two variables as visible from the scatterplot. [1]
[4]

iii) Null hypothesis H0: ρ = 0 against H1: ρ < 0 [1]

Need to assume that data come from a bivariate normal distribution. [1]

From page 25 of tables, r = 0.5 * ln(1-0.686/1.686) = -0.8404 [1]

And under H0, this should be a value from the N(0, 1/7) distribution. [1]

Fisher's standardized statistic = (-0.8404 – 0)/($\sqrt{1/7}$) = -2.22 [1]

This gives the p-value = P(z<-2.22) = 0.013 which is quite small and hence shows a strong evidence to reject the null hypothesis with 95% confidence. We can conclude that marks obtained and hours spent on social media are negatively correlated. [2]
[7]

iv) Beta = Sxy/Sxx = -296/75 = -3.9467
Alpha = mean of y – beta * mean of x = 644/10 + 3.9467 *45/10 = 82.16
Fitted line is y = 82.16 – 3.9476x
[3]

v) R^2 = -0.686^2 = 0.4706
This gives the proportion of total variation explained by the model. [2]

vi) For every additional hour spent on social media per day, the total marks reduce by 3.95 (~4 marks) basis the fitted equation. [1]
[18 Marks]

10.

H0: There is no difference among industries [1]

H1: At least one industry differs significantly from the overall mean

$SS_R = 19\ (5^2 + 10^2 + 8^2) = 3591$ [1.5]

Mean of resignation = (27+36+30)/3 = 31

$SS_B = 20((27-31)^2 + (36-31)^2 + (30-31)^2)$ [1.5]

= 840

$F_{2,57} = (840/2)/(3591/57) = 6.667$ [1]

The 1% point from $F_{2,60}$ is 4.977 and since the test statistic is higher than this, the null hypothesis is rejected. We conclude that resignation rate is different across different industries. [1]

[6 Marks]

11.

i) The PF of Z is

$$f(z) = \binom{n}{z}\mu^z(1-\mu)^{(n-z)}$$

The PF function of Y can be obtained by replacing z with ny :

$$f(y) = \binom{n}{ny}\mu^{ny}(1-\mu)^{(n-ny)}$$

This can be written as :

$$f(y) = \exp\left\{\ln\binom{n}{ny} + ny\ln\mu + n\ln(1-\mu) - ny\ln(1-\mu)\right\}$$
$$= \exp\left\{ny\ln\left(\frac{\mu}{1-\mu}\right) + n\ln(1-\mu) + \ln\binom{n}{ny}\right\}$$
$$= \exp\left\{\frac{y\ln\left(\frac{\mu}{1-\mu}\right) + \ln(1-\mu)}{1/n} + \ln\binom{n}{ny}\right\}$$

Comparing this to the generalized form of exponential family of distributions:

$\theta = \ln\left(\frac{\mu}{1-\mu}\right)$. Rearranging this gives $\mu = \frac{e^\theta}{1+e^\theta}$

$b(\theta) = -\ln(1-\mu) = -\ln\left(1-\frac{e^{\theta}}{1+e^{\theta}}\right) = -\ln\left(\frac{1}{1+e^{\theta}}\right) = \ln(1+e^{\theta})$

$\varphi = n$,

$a(\varphi) = \frac{1}{\varphi}$

$c(y, \varphi) = \ln\binom{n}{ny} = \ln\binom{\varphi}{\varphi y}]$

[4]

ii) Using the properties of exponential distributions

$E(Y) = b'(\theta) = \frac{d}{d\theta}(\ln(1+e^{\theta})) = \frac{e^{\theta}}{1+e^{\theta}} = \mu$

$V(Y) = a(\varphi)\, b''(\theta) = \frac{e^{\theta}(1+e^{\theta}) - e^{\theta}e^{\theta}}{n(1+e^{\theta})^2} = \frac{e^{\theta}}{n(1+e^{\theta})^2}$

Substituting $\theta = \ln\left(\frac{\mu}{1-\mu}\right)$

(

$V(Y) = \frac{\frac{\mu}{1-\mu}}{n(1+\frac{\mu}{1-\mu})^2} = \frac{\mu}{n(1-\mu)}(1-\mu)^2 = \mu(1-\mu)/n$

[3]

iii) Using the model output, we can see that

$\beta_1 > 2 \times standard\ error(\beta)$

i.e $0.5459 > 2 \times 0.08352 = 0.16704$

Since

$\beta_1 > 2 \times standard\ error(\beta)$, it can be concluded that the parameter β_1 for the variable "no. of assignment" is significant in the model.

[2]

iv) Using binomial canonical link function,

$\eta(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) = \alpha_i + \beta_1 N + \beta_2 S$

So for $\alpha_Y = -1.501$, $\beta_1 = 0.5459$, $\beta_2 = 0.0251$ and N = 4, S = 65

$\ln\left(\frac{\mu}{1-\mu}\right) = -1.501 + 0.5459 \times 4 + 0.0251 \times 65 = 2.3141$

$\mu = (1 + e^{-2.3141})^{-1} = 91\%$

Hence probability of passing students in the given scenario is 91%

12.

i)

a) Correct Option is Option D

steps not required

theta= log(mu) hence, mu = e^{theta}
E(Y) = b(theta) = mu = e^{theta}

b`(theta) = e^{theta} = mu
V(mu) = b``(theta) = e^{theta} = mu
Hence correct option is Option D

b) Correct option is Option C (2)

steps not required

b(theta) = -log(-theta)
mean = E(Y)= b`(theta) = -1/theta
as theta=1/mu, b`(theta) = mu
variance function is V(mu) = b``(theta) = $1/theta^2 = mu^2$
Hence, variance is V(mu) / a(phi) = $mu^2/1 = mu^2$
Hence, correct option is Option C i.e. mean=mu and variance = mu^2

ii)

a) **Interaction** term means that effect of age band on accidental hospitalisation claims depends on the gender of the insured and **significant** indicates that accidental hospitalisation claims are better modelled with interaction term, when claims for any age group change with respect to gender of the insured.

E.g. accidental hospitalisation claims are expected to be higher for males compared to females for any age group.
This can be achieved in the model by having Beta_males > Beta_females across age groups. This will result in higher expected accidental hospitalisation claims for males compared to females for any age group (assuming only other parameter is for age band which is same for males and females. (4)

b) Policy renewal is binary event for a single policy. Hence, predicted policy renewal rate (say, mu) for a group of policies can vary between 0 to 1.

If logit link function is used, then
eta = log(mu/(1-mu))
mu/(1-mu) = exp(eta)
mu = exp(eta) – mu*exp(eta)
mu(1+exp(eta)) = exp(eta)
mu = exp(eta)/ (1+exp(eta)) = 1/(1+exp(-eta)) = (1+exp(-eta))^-1

This is expected to result in the range of 0 to 1 for mu as required - renewal rate for a group of policies.
Hence, logit link function can be used for renewal rate. (4)

iii)

a)

	Model 1	Model 2	
SSREG	1.380	2.380	Given
A	1.380	2.380	as A/1 = SS_{REG}
B	4.000	3.000	B=5.38-A
C	0.333	0.250	C= B/12
F	4.140	9.520	F=SS_{REG}/C

b)

H_0 = Beta parameter is zero (no linear relationship)

H_1 = Beta parameter is not equal to zero (linear relationship is present)

For Model 1, F = 4.14, this is between 5% critical value of 4.747 and 10% critical value of 3.177
Hence, we have sufficient evidence to reject null hypothesis that beta parameter is zero at 10% level (but cannot be rejected at 5% level) indicating linear relationship between response variable and predictor variable at 10% level

For Model 2, F=9.52 is greater than critical value at 1% level as well.
Hence, we have sufficient evidence to reject null hypothesis even at 1% level indicating linear relationship between response variable and predictor variable at 1% level

(3)

c) $R^2 = SS_{REG} / SS_{TOT}$

For Model 1, R^2 = 1.38/5.38 = 25.7% and
For Model 2, R^2 = 2.38/5.38 = 44.2%

As % of variation explained by model 1 and Model 2 is low (based on low value of R^2 of Model 1 and Model 2), we can conclude that none of the model is good fit to the data and hence, models are not suitable for prediction purpose (though there is linear relationship between response and predictor variable for Model 1 and Model 2 as indicated in part b)

(3)

iv)

a)

Based on Result 1

- Model using x as predictor is significant improvement over null model
- As reduction in deviance (by 1769) is significantly more compared to 2 times the loss in degrees of freedom by 1 when x is used as predictor over null model
- Result is significant even at 0.000001 level as p value is smaller than that

Based on Result 2

- Model using interaction term between x and region is significant improvement over model using just x as predictor
- As reduction in deviance (by 87) is significantly more compared to 2 times the loss in degrees of freedom by 2 when interaction between x and region is considered over model using just x as predictor
- Result is significant at 0.005 level as p value is smaller than 0.005

(4)

b)

Based on Result 3

- Model using only main effect of x and region is not significantly different compared to model using interaction between x and region
- As reduction in deviance (by 0.037) is less than 2 times the loss in degrees of freedom by 1 when interaction between x and region is considered over model using just the main effect of x and region
- As p value (0.92) is much more than 90% and

Comparing Result 2 and Result 3

- Model using only main effects of x and region is significantly better than model using x as predictor

- as reduction in deviance is significantly more compared to 2 times the loss in degrees of freedom by 1 when main effect of x and region are used as predictor over model using only x
- and p value to be significant at 0.005 level

- We prefer simpler model i.e. Model with only main effects of x and region over complex model i.e. Model having interaction between x and region as additional complexity is not justified by better prediction