Probability & Statistics 2

Assignment 2

Q1.

x	40	10	100	110	120	150	20	90	80	130
y	56	62	195	240	170	270	48	196	214	286

(a) $\Sigma xy = 1,82,560$ $\qquad \Sigma x^2 = 92,500$

$\Sigma x = 850$

$\Sigma y = 1737$

$$S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$$

$$= 34,915$$

$$S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$= 20,250$$

$$\hat{\beta} = \frac{S_{xy}}{S_{xy}}$$

$$= 1.724198$$

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

$\bar{y} = 173.7$

$\bar{x} = 85$

$$\hat{\alpha} = 27.14317$$

$$\hat{y} = 27.14317 + 1.72419\,x$$

(b) The gradient of the regression line refers to ~~the~~ its slope parameter, β. Since β is significantly different from 0, there exists a positive linear relationship b/w x and y.

Here, there is a positive linear relationship b/w the amount withdrawn and the number of hours until the next visit.

Q2. $X =$ share price of Dell
$y =$ share price of IBM

(i) $S_{XY} = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$

$= 875.16$

$S_{XX} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$

$= 301.66$

$\hat{\beta} = \frac{S_{XY}}{S_{XX}}$

$= 2.90115$

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

$= 3.74809$

$\hat{y} = 3.74809 + 2.90115x.$

(ii) Assumptions:
→ The distributions for x & y are normal
→ The ~~[illegible]~~ explanatory variable values (observations) are independent of each other.

$$\frac{\hat{\beta}-\beta}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$$

~~Confidence interval~~ → Level of significance = 5%.

95% confidence interval for $\beta = \hat{\beta} \pm t_{10, 2.5\%} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

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

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

$$= 6409.7125$$

$$\hat{\sigma}^2 = \frac{1}{10}\left(S_{yy} - \frac{S^2_{xy}}{S_{xx}}\right)$$

$$\hat{\sigma}^2 = 387.07447$$

95% Confidence interval for $\beta = \hat{\beta} \pm t_{10, 2.5\%} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

$$= 2.90115 \pm 2.52379$$

95% Confidence interval for β = (0.37736, 5.42494)

PAGE NO.:

DATE: / /

(iii) Mean IBM share price $= \mu_0$

$$\frac{\hat{\mu}_0 - \mu_0}{se(\hat{\mu}_0)} \sim t_{n-2}$$

$x_0 = \mu_0$

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

$\hat{\mu}_0 = 119.79409$

$$se(\hat{\mu}_0) = \sqrt{\left(\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}\right)\hat{\sigma}^2}$$

$= 10.59894$

95% confidence interval for $\mu_0 = \hat{\mu}_0 \pm t_{10, 2.5\%} \, se(\hat{\mu}_0)$

$= 119.79409 \pm 23.61443$

95% confidence interval for $\hat{\mu}_0 = (96.17966, 143.40852)$

Q3. (i) Comments on the plot:

- The centres of the distribution differs for all the 4 cities. Thus, there is a prima facie case 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 (highest) to 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 can't be sure that there is a real underlying difference in variance.

(ii) Assumptions underlying analysis of variance:

- The populations must be normal
- The population have a common variance
- The observations ~~(explanatory variable)~~ are independent.

(iii) H_0: The mean of differences is same for each city
H_1: The mean of differences is not the same.

$$SS_{TOT} = S_{yy}$$
$$= \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$
$$= 1495 - \frac{103^2}{28}$$
$$SS_{TOT} = 1116.11$$
$$SS_{REG} = \cancel{\frac{S_{xy}^2}{S_{xx}}} = \frac{1}{7}(64^2 + 44^2 + 8^2 + (-13)^2) - \frac{103^2}{28}$$
$$= 516.11$$
$$SS_{RES} = SS_{TOT} - SS_{REG} = 600$$

Page No.
Date: / /

ANOVA Table

Source	degree of fredom	Sum of Squares	Mean sum of squares
Regression	3	516.11	172.04
residual	24	600	25
Total	27	1116.11	

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

Test statistic 6.88.

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

Hence, we have sufficient evidence to reject H_0 at 5% level of significance. Thus, ~~they are~~ there are underlying differences b/w the cities.

(iv) Analysis of the mean differences

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

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

Page No.
Date: / /

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

The least significance ~~bto~~ difference b/w any pair 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 least significance difference 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 3 differences are less than 5.52, we underline 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

$$\bar{y}_{1*} - \bar{y}_{3*} = 8$$

There is a significant difference b/w means 1 & 3, so we can't combine the first 2 groups.

Examining to see if the last 2 groups can be combined

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

Hence, we can't combine the last 2 grps.
Therefore, the diagram remains as before.

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

$$Y_{ij} = \mu + T_i + e_{ij}, \quad i = 1, 2, \ldots, k$$
$$, j = 1, 2, \ldots, n_i$$

where,

- k = number of treatments
- n_i = no. of responses from i^{th} treatment
- Y_{ij} is the 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} 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.

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

Page No.______ Date: / /

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

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

$$C.F = \left(\sum_{i=1}^{s}\sum_{j=1}^{n_i} y_{ij}\right)^2 / n = 104385.94$$

$$SS_{TOT} = 174316 - CF$$
$$= 69,930.06.$$

$$SS_{REG} = \sum_{i=1}^{s}\left(\sum_{j=1}^{n_i} y_{ij}\right)^2 / n_i - C.F$$

$$= \frac{354^2}{7} + \frac{386^2}{8} + \frac{87^2}{6} + \frac{645^2}{7} + \frac{384^3}{J} - 104385.94$$

$$= 22,325.69$$

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

$$= 47,604.37$$

Sources of variation	d.f	SS	MSS
Regression	4	22,326	5581.42
Residual	28	47,604	1700.16
Total	32	69,930	

Test statistic $= \dfrac{MSS_{REG}}{MSS_{RES}}$

$= 3.283$

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

We have sufficient evidence to reject H_0.

(C) Observed Frequency (O_i)

Papers Cleared	Salary per annum (in Rs. Lacs) 3-5	5-8	8-10	10-12	Total
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

Expected Frequency (E_i)

Papers Cleared	Salary per annum (in Rs. Lacs) 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	5.882278

No clubbing for E_i's required as ~~well~~ all E_i's are greater than 5.

Contingency table

H_0 : Salaries are independent of the no. of actuarial papers cleared
H_1 : Salaries are dependant on the no. of actuarial papers cleared. (one sided test)

$$\text{Test statistic} = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= 11.27506 + 0.41317 + 4.92817 + 3.32302 + 4.38581 + 4.10871 + 0.13183 + 0.00212 + 6.16284 + 1.41856 + 7.21193 + 6.55324$$

Test statistic = 49.91146

$\cancel{\chi^2_{11, 5\%} = 19.68}$

$\cancel{\chi^2_{11, 1\%} = 24.73}$

degree of freedom = $(r-1)(c-1)$
$= 2 \times 3$
$= 6$

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

Hence, we have sufficient evidence to reject H_0.

The salaries are dependant on no. of actuarial papers cleared.

Page No.
Date: / /

Q5. (i) Assumptions underlying ANOVA:

- Populations are normal
- Populations have common variance
- Observations are independent.

The 4 sample variances are very different from each other.

~~(ii) Mean at Rate 1~~

Hence, the common variance assumption won't hold.

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

$= 4.52443$

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

~~$= \frac{1}{n-1}(\Sigma x^2 - n\bar{x}^2)$~~

~~$= 0.11769$~~

(iii) The scientist was correct in assuming that the $\log_e$ 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.

(iv) (a) $Y_{ij} = \mu + \tau_i + e_{ij}$, $i = 1,2,3,4$, $j = 1,2,3$

- 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 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.

$e_{ij} \sim N(0, \sigma^2)$

(b) For ANOVA,

$H_0 : \tau_i = 0$, $i = 1,2,3,4$

$H_1 : \tau_i \neq 0$ for atleast one i.

$\ln(x)$

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	389.060
Total	20.110	0.618	973.373

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

$$SS_{Rate} = \sum_{i=1}^{4} \frac{Y_i^2}{3} - \frac{Y^2}{12} = \frac{973.373}{3} - \frac{(3 \times 20.110)^2}{12}$$

$$SS_{Rate} = 21.153$$

$$SS_{RES} = \sum_{i=1}^{4}\left(\sum_{j=1}^{3}(y_{ij} - \bar{y}_{i\cdot})^2\right)$$

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

$$= 2 \times 0.618$$

$$= 1.236$$

Source of variation	d.f	SS	MSS
Rate	3	21.153	7.051
Residuals	8	1.236	0.155
Total	11	22.38~~7~~	

Test statistic = $\dfrac{MSS_{Rate}}{MSS_{Res}}$

$= 45.638$

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

(C) Given the observed F statistic value is much larger than that, we can state 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.

Q6

x	5	9	2	3	3	1	1	6	5	4
y	73	120	34	46	35	24	26	93	45	66

$n = 10$

(i) $r = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$

$S_{xy} = \sum xy - n\bar{x}\bar{y}$
$= 661.2$

$S_{xx} = \sum x^2 - n\bar{x}^2$
$= 54.9$

$S_{yy} = \sum y^2 - n\bar{y}^2$
$= 8923.6$

$r = 0.94466$

(ii) Fitted regression eq^n

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
$\hat{\alpha} = 9.2295$

$\hat{\beta} = \dfrac{S_{xy}}{S_{xx}}$
$= 12.04371$

$\hat{y} = 9.2295 + 12.04371x$

Page No.______ Date: / /

(iii) $SS_{TOT} = S_{YY} = 8923.6$

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

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

$R^2 = \frac{S^2_{XY}}{S_{XX} S_{YY}}$

$= 89.238703\%$

Comment:

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

→ Better models can still be considered covering more variability.

Q7. (i) Linear regression eq^n

$S_{XX} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$, $S_{YY} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$

$S_{XX} = 4788.4221$ $S_{YY} = 1189.20767$

$S_{XY} = 2176.84$

$\hat{\beta} = \frac{S_{XY}}{S_{XX}}$

$= 0.45451$

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

$= 10.78056$

$\hat{y} = 10.78056 + 0.45451x$

(ii) $H_0: \beta = 0$
$H_1: \beta \neq 0$

$$\frac{\hat{\beta} - \beta}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$$

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

$$= \frac{1}{9}\left(S_{yy} - \frac{S^2_{xy}}{S_{xx}}\right)$$

$$\hat{\sigma}^2 = 22.20136$$

$$\text{Test statistic} = \frac{\hat{\beta} - 0}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

$$= 6.67568$$

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

Here we have sufficient evidence to reject H_0. $\beta \neq 0$.
Hence a linear relationship exists.

(iii) Assumptions:
→ Populations have a common variance
→ Population follow the normal distribution
→ The observations are independent

Page No.____ Date: / /

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

$= 0.91213$

(iv) <u>For mean response (μ_0)</u>

$x_0 = 25$

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

$= 22.15331$

$se(\hat{\mu}_0) = \sqrt{\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)\hat{\sigma}^2}$

$= 1.55181$

$\frac{\hat{\mu}_0 - \mu_0}{se(\hat{\mu}_0)} \sim t_9$

95% CI for $\hat{\mu}_0 = \hat{\mu}_0 \pm t_{9, 2.5\%} \, se(\hat{\mu}_0)$

$= 22.15331 \pm 3.51018$

$= (18.64313, 25.66349)$

<u>For individual responce (y_0)</u>

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

$= 22.15331$

$se(\hat{y}_0) = \sqrt{\left(\frac{1}{n} + 1 + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\hat{\sigma}^2}$

$= 4.96079$

Page No.
Date: / /

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

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

$= 22.15331 \pm 11.22131$

$= (10.932, 33.37462)$

IV) There ~~exists~~ is clearly some relationship between the reduals and the explaratory variable values. Hence our assumption of normality might be wrong. The observations are not independent as well. The model is not a good fit to the data.

Q8. i)

→ The main purpose of factor analysis is to reduce many individual items into a fewer number of dimensions.

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

→ Principle Component analysis is a technique for reducing the dimensionality of such data sets increasing interpratability but at the same time minimising information loss.

While newly identified principle components are chosen to be:

- Uncorrelated
- Linear combinations of the original variables of the data
- which ~~maxim~~ maximise the variance.

Page No.______ Date: / /

(ii)

Principle component	Diagonal entry (PC_i)	PC_i / (Sum (PC_i) over 1 to 5).
PC1	0.456	65%
PC2	0.137	19.5%
PC3	0.08	11.4%
PC4	0.0165	2.4%
PC5	0.012	1.7%
	0.7015	100%

(iii) As 1st 3 principle components explain over 95% of total variance, dimensionality can be reduced to 3 for this dataset.
The 1st 3 principle components can then be used for building further classification or regression modelling purpose.

Q9. (i) There appears to be a ~~weak~~ negative correlation b/w the marks scored and hours spent per day on social media.

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

$= 75$

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

$= 2482.4$

$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$

$= -296$

Page No.______ Date: / /

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

$$= -0.686002$$

Hence, they have a weak negative correlation.

(iii) $H_0 : \rho = 0$
$H_1 : \cancel{\rho \neq 0} \ \rho < 0$

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

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

Test statistic $\cancel{T.S} = \left(\tanh^{-1}(r) - \tanh^{-1}(\rho)\right)\sqrt{n-3}$

$$Z_{cal} = -2.07893$$

$\cancel{Z_{95\%} = -1.6449}$

$$Z_{5\%} = 1.6449$$

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

Hence, we have sufficient evidence to reject H_0.
$\rho < 0$.

Page No.______
Date: / /

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

$= -3.94667$

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

$= 82.16$

$\hat{y} = 82.16 - 3.94667x$

(V) $R^2 = \frac{S^2_{xy}}{S_{xx}S_{yy}}$

$= 47.05983\%$.

Hence, the model explains only 47% of the variability.
The model is not a good fit to the data.

(Vi) $x_0 = 1$
$\hat{\mu}_0 = \hat{\alpha} - \hat{\beta}x_0$; expected charge $= \hat{\mu}_0$.
$= 78.21333$

Page No.
Date: / /

Q10.

x	0.27	0.36	0.3
y	0.05	0.1	0.08

$S_{XX} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 0.0042$

$S_{YY} = \Sigma y^2 - n\bar{y}^2 = 0.00126$

$S_{XY} = \Sigma xy - n\bar{x}\bar{y} = 0.0022$

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

$SS_{REG} = \frac{S^2_{XY}}{S_{XX}} = 0.00115$

$SS_{RES} = \cancel{0.00108}\ 0.00011$

ANOVA table

Sources of variation	df	SS	MSS
Regression	1	0.00115	0.00115
Residual	1	~~0.00108~~ 0.00011	0.00011 ~~0.00108~~
Total	2	0.00126	

$H_0 : \beta = 0$

$H_1 : \beta \neq 0$ (one sided F-test)

Test statistic $= \frac{MSS_{REG}}{MSS_{RES}} = \cancel{1.06481}\ 10.45455.$

$F_{1,1,5\%} = 161.4$

~~Fcal <~~ $F_{cal} < F_{tab}$,

We do not reject H_0. Thus, industry has no effect on response rate.