Probability & Statistics
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) $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$

$S_{xy} = \cancel{\Sigma X_i Y_i - \frac{\Sigma X_i \Sigma Y_i}{n}} \quad \Sigma X_i Y_i - \frac{\Sigma X_i \Sigma Y_i}{n}$

$\Sigma X_i Y_i = 182560$
$\Sigma X = 850$
$\Sigma Y = 1737$

$S_{xy} = 34915$

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

$\Sigma x^2 = 92500$

$S_{xx} = 92500 - \frac{(850)^2}{10}$

$= 20250$

$\hat{\beta} = \frac{34915}{20250} = 1.72$

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

$\bar{y} = 173.7$
$\bar{x} = 85$

$\alpha = 173.7 - 1.72 \times 85$

$\alpha = 27.5$

$y = \alpha + \beta x$

$y = 27.5 + 1.72x$

b) The gradient represents the amount of hours per rupee spent

Dell | IBM

Q2 $\Sigma x = 385.2$ | $\Sigma y = 1{,}162.5$

$\Sigma x^2 = 12{,}666.58$ | $\Sigma y^2 = 119{,}026.9$

$\Sigma xy = 38{,}191.41$

$$S_{xy} = \Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}$$

$$= 38191.41 - \frac{385.2 \times 1{,}162.5}{12}$$

$S_{xy} = 875.6$

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

$$= 12666.58 - \frac{(385.2)^2}{12}$$

$$= 301.66$$

$$\beta = \frac{S_{xy}}{S_{xx}} = \frac{875.6}{301.66}$$

$$\beta = 2.902$$

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

$$= 96.875 - 2.902 \times 32.1$$
$$= 3.72$$

$$y = 3.72 - 2.902x$$

ii) Assumptions

1) Distributions for x & y are normal

2) The variables are independent of eachother

Level of significance = 5%

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

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

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

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

$$= 6409.7125$$

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

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

$$= 2.90115 \pm 2.52377$$

95% for confidence Interval for β = (0.37736, 5.42492)

iii) Mean IBM Share Price = 40

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

$x_0 = 40$

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

$\hat{\mu}_0 = 119.78404$

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

$= 10.59894$

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

$= 119.79409 \pm 23.61443$

95% CI for $\hat{\mu}_0 = (96.17866, 143.40852)$

Q3 i) 1) The difference between the mean time taken to commute to office in peak hours and non-peak hours follow an ascending order starting from City A

2) The centre of distribution is different for all 4 cities.

ii) 1) The populations must follow normal distribution
2) The populations have same variance

iii) H_0: Mean difference is same
H_1: mean difference is not same

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

$= 1495 - \frac{103^2}{28}$

$SS_{Tot} = 1116.11$

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

$= 516.11$

$SS_{Res} = 600$

ANOVA Table

Source	Degree of freedom	Sum of Square	Mean Square
Regression	3	516.1	172.04
Residual	24	600	25
Total	27	1116.11	

Under H_0,

$$F = \frac{MSS_{Reg}}{MSS_{Res}}$$

$$= \frac{172.04}{25}$$

$$= 6.88$$

$T.S = 6.88$

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

$\therefore$ Reject H_0

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

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

where

k = number of treatment

n_i = number of response from ith treatment

Y_{ij} is the jth response from ith treatment

τ_i is the ith treatment

μ is the mean response

e_{ij} is the error term

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

b) H_0: mean claim amounts for the five companies are equal

H_1: mean claim amounts are not equal

$n_1 = 4$

$n_2 = 8$

$n_3 = 6$

$n_4 = 7$

$n_5 = 5$

$\therefore n = 33$

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

$$\sum_{i=1}^{s} \sum_{j=1}^{n_i} y_{ij}^2 = 174316$$

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

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

$$SS_{Reg} = \sum_{i=1}^{s} \left(\sum_{j=1}^{n_i} y_{ij}\right)^2 / n_i - CF$$

$$= \frac{354^2}{7} + \frac{386^2}{8} + \frac{87^2}{6} + \frac{645^2}{7} + \frac{384^2}{5}$$
$$- 104385.94$$

$$= 22,325.69$$

$$SS_{Res} = SS_{TOT} - SS_{Reg}$$

$$= 47,604.37$$

Sources of Variation	d.f	SS	MSS
Regression	4	22,326	5581.42
Residual	28	47,604	1700.86
Total	32	69,930	

$$T.S = \frac{MSS_{Reg}}{MSS_{Res}}$$

$= 3.283$

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

$\therefore$ Reject HO.

c) Observed freq^n (Oi)

Paperclears	Salary: 3-5	5-8	8-10	10+2	Total
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

Expected $\text{Freq}^r (E_i)$

Papers Clears	Salary 3-5	5-8	8-10	10 +2
0-3	27.417	23.088	14.450	11.663
4-6	15.1518	12.7594	7.914	6.113
7-9	14.433	12.1518	7.594	5.522

H_0: Salaries are indepent of no. of the actuarial papers cleared.
H_1: Salaries are dependent on no. of actuarial papers cleare.

$$T.S = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= 49.8114$$

Degree of freed $= (r-1)(c-1)$
$= 2 \times 3$
$= 6$

$\chi^2_{6,5} = 12.5$

$\therefore$ Reject H_0.

Q5 i) Assumptions for ANOAV
1) Populations Are normal
2) Populations have common Variance
3) Observations are independent

The k sample Variance are different from each other therefore Common Variance assumption won't apply

ii) mean at Rate 1 for $\sqrt{n}$ $= \frac{\sqrt{28} + \sqrt{13} + \sqrt{21}}{5}$

$= 4.5443$

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

$= 6.1227$

iii) The $\log_e$ transformation must be there fore carrying out one way ANOV.

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

- Y_{ij} is the loge transformed value of the j^{th} observation of the no. of germinations per sqft observed when i^{th} rate was applied
- μ is the overall popu tion 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
$H_0: t_i = 0$
$H_1: t_i \neq 0$

(Q(ii))

Rate	mean	Variance	Y_i^2
1	2.992	0.163	80.582
2	4.827	0.121	209.678
3	5.116	0.186	297.053
4	6.575	0.148	388.060
Total	20.110	0.618	975.373

Here, $y_{i.}^2 = \left(\sum_{j=1}^{3} y_{ij}\right)^2$

$= (3 \times mean_i)^2$

$$SS_{rate} = \frac{973.373}{3} - \frac{(3 \cdot 2\cancel{0}+10)^2}{12}$$

$= 21.153$

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

$$= \sum_{i=1}^{4} (2 \times variance_i^2)$$

$= 2 \times 0.618$

$= 1.236$

Source of variation	df	SS	MSS
Rate	3	21.153	7.051
Residual	8	1.231	0.155
Total	11	22.38	

$$\text{Test } f = \frac{MSS\,Rate}{MSS\,Res}$$

$$= 45.638$$

$$F_{3,8,1\%} = 7.95$$

c) P value for this test is almost near to 0.
$\therefore$ Underlying means are not equal

Q6 i) $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

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

$$S_{xx} = \Sigma x^2 - n\bar{x}^2$$
$$= 54.7$$

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

$$r = 0.94466$$

ii) Fitted regression eq^n

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

$\hat{\alpha} = 9.2285$

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

$= 12.04371$

$\hat{y} = 9.2285 + 12.04371x$

iii) $SS_{TOT} = S_{yy}$

$= 8923.6$

$SS_{REG} = \dfrac{S^2_{xy}}{SS_{x^2}}$

$= 7963.381$

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

$= 960.9250$

$R^2 = \dfrac{S^2_{xy}}{S_{xx} S_{yy}}$

$= 89.238\%$

Model is good fit as 89.238% of variability is seen.

Q1 i) $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$

$= 4788.421$

$S_{xy} = 2176.84$

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

$= 0.45431$

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

$= 10.78058$

$\hat{y} = 10.78058 + 0.454514x$

ii) $H_0 : \beta = 0$

$H_1 : \beta \neq 0$

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

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

$$= \frac{1}{8}\left(S_{yy} - \frac{S^2_{xy}}{S_{xx}}\right)$$

$$\sigma^2 = 22.20136$$

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

$$= 6.67569$$

$$T_{8, 0.025} = 2.262$$

$\therefore$ Linear Relationship exists

Assumptions.

1) Populations follow normal distribution.
2) Population have common Variance
3) The observations are independ.

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

$= 0.9213$

iii) For mean response (μ_0)

$x_0 = 25$

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

$= 22.15331$

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

$= 1.55181$

$\frac{\hat{\mu}_0 - \mu}{se(\hat{\mu}_0)} \sim t_8$

95% CI for $\mu_0 = \hat{\mu}_0 \pm t_{8, 0.025}\, se(\hat{\mu}_0)$

$= 22.15331 \pm 3.5181$

$= (18.6423, 25.6634)$

For individual response (y_0)

$\hat{y}_0 = \hat{\alpha} + \hat{\beta} x_0$
$= 22.15331$

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

$= 4.86079$

$\frac{\hat{y}_0 - y}{se(\hat{y}_0)}$ at ν

95% CI for $y_0 = \hat{y}_0 \pm t_{8, 2.5} \, se(\hat{y}_0)$
$= 22.15331 \pm 11.22131$
$= (10.932, 33.37462)$

iv) There is clearly some ~~correlat~~ relationship seen.

The model is not a good fit for the data.

Q8 i) The main purpose of factor analysis is to reduce many individed ~~a~~ into a fewer number of Dimaps

1) Factor analysis can reduce the number of variables in a regression model.
2) Principal component analysis simplifies the complexity in high dimensional data while retaining trends & patters.

Newly identifie PCA's
1) Not corelated
2) Maximize Variance

ii)

Q9 i) There appears to be a negative correlation b/w the marsk scored and ~~[illegible]~~ number of hours spent on social media

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

$= 75$

$$S_{yy} = \Sigma y^2 - \frac{[\Sigma y]^2}{n}$$

$$= 2482.4$$

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

$$= -296$$

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

$$= -0.686002$$

$\therefore$ Weak neg.

(iii) $H_0: \rho = 0$
$H_1: \rho < 0$

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

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

$$= \left(\tanh^{-1}(r) - \tanh^{-1}(\rho)\right)\sqrt{n-3}$$

$$Z_{cal} = -2.07813$$

$Z_{5\%} = 1.6448$

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

$\therefore$ Reject H_0

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

$= -3.94667$

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

$= 82.16$

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

v) $R^2 = \dfrac{S^2_{xy}}{S_{xx} S_{yy}}$

$= 47.0583\%$

$\therefore$ model explains 47% of the variability. The model is not good for data.

(vi) $x_0 = 1$

$\hat{\mu} = \hat{\alpha} - \hat{\beta}_1 x_0$

expected change $= \hat{\beta}_1 x_0$

$= 78.215333$

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

$= 0.0042$

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

$= 0.00126$

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

$= 0.0022$

$SS_{TOT} = S_{yy}$

$= 0.00126$

$SS_{REG} = \frac{S^2_{xy}}{S_{xx}}$

$= 0.00115$

$SS_{Res} = 0.00011$

ANOVA Table

Sources of Variation	df	SS	MSS
Regression	1	0.00115	0.00115
Residual	1	0.00011	0.00011
Total	2	0.00126	

$H_0 = \beta = 0$
$H_1 = \beta \neq 0$

$TS = \frac{MSS_{Reg}}{MSS_{Res}}$

$= 10.45455$

$F_{1,1,5\%} = 16.4$

$F_{cal} < F_{tab}$

We do not Reject H_0.