Assignment 2 - Probability & Statistics 2

1

DATE / /

Name: Kshiti Vartak Roll No: 86 Section B

1] a) $S_{xy} = \Sigma x_i y_i - \Sigma x_i \Sigma y_i / n = 182560 - 850 \times \frac{1737}{10} = 34916$

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

$y = \hat{a} + \hat{b}x$

$\hat{a} = \left(\frac{1737}{10}\right) - 172\left(\frac{850}{10}\right) = 27.14$

$\hat{b} = 1.72 \quad (\because S_{xy}/S_{xx})$

$\therefore y = 27.14 + 1.72x$

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

2] i) $\Sigma x = 385.2$, $\Sigma y = 1162.5$, $\Sigma xy = 38191.41$, $n = 12$

$\Sigma x^2 = 12666.58$, $\Sigma y^2 = 119026.9$

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

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

$S_{xy} = \Sigma xy - n\bar{x}\bar{y} = 38191.41 - 12\left(\frac{385.2}{12}\right)\left(\frac{1162.5}{12}\right) = 875.16$

$\hat{\beta} = 2.90$

~~$\hat{\alpha} = \bar{y}$~~

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

$$\boxed{y = \hat{\alpha} + \hat{\beta}x = 3.78 + 2.90x}$$

ii) 95% C.I for $\beta \pm t_{n-2}(2.50\%) \cdot s.e(\hat{\beta})$

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

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

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

$\therefore$ 95% C.I. for β : $2.90 \pm 2.228(1.13) = (0.38, 5.42)$

iii) $\hat{y}_{x_0} \pm t_{n-2}(2.5\%)\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$

$x_0 = 40$ (Dell share price in US \$)

$\hat{y}_{x_0} = 3.48 + 2.90(40) = 119.78$

$\therefore$ 95 C.I $= 119.78 \pm 2.228 \times \sqrt{387.07\left[\frac{1}{12}+\frac{(40-32.1)^2}{301.66}\right]}$

$= 119.78 \pm 2.228 \times 10.5989$

$= (96.17, 143.39)$

3) i) The centers of the distributions differ for all the four cities.

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

ii) ★ The populations must be normal

★ The observations are independent

★ The populations have a common variance

iii) H_0: The mean of differences is same for each city

v/s

H_1: The mean of differences are not the same for all cities

$SS_T = 1495 - \frac{103^2}{28} = 1116.11$

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

ANOVA Table:

Source	df	SS	MS	F
Treatments	3	516.11	172.04	6.88
Residual	24	600	25	
Total	27	1116.11		

Under H_0, $F = \frac{172.04}{25.00} = 6.88$ $(\sim F_{3,24})$

The 5% critical point is 3.009, so we have sufficient evidence to reject H_0 at the 5% level. Therefore it is resonable to ~~at~~ conclude that there are underlying differences between the cities.

5] • i) The following assumptions are required for one-way analyses of variance (ANOVA):
- The populations must be normal.
- The populations have a common variance.
- The observations are independent.

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

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

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

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

$$= 0.1213$$

iii] The scientist was correct in asserting 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] We will perform an ANOVA on $\log_e x$ data. We would assume the following model:

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

Here:
- Y_{ij} is $\log_e$ transformed value of the j^{th} observation of the no. of germinators 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 $\Sigma \tau_i = 0$
- e_{ij} are the independent error terms which follows Normal distribution with mean 0 and common unknown variance σ^2

For ANOVA, the null hypothesis being tested here is:

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

Sum of squares table: ($\log_e(x)$)

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
	20.110	0.618	973.373

Here: $Y_i^2 = (\Sigma_{j=1}^{3} Y_{ij})^2 = (3 * \text{Mean}_i)^2$

Now:

$SS(\text{Rate}) = \Sigma_{i=1}^{4} \frac{Y_i^2}{3} - \frac{Y_{..}^2}{12} = \frac{973.373}{3} - \frac{(3 \times 20.110)^2}{12}$

$= 21.163$

$SS(\text{Residuals}) = \Sigma_{i=1}^{4} \{\Sigma_{j=1}^{3} (Y_{ij} - \bar{Y}_i)^2\}$
$= \Sigma_{i=1}^{4} \{2 \times \text{Variance}_i\} = 2 \times 0.618 = 1.236$

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		

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

Consedering the F statistic value is much larger than this, we can observe that the p-value for this test is almost near to zero. This is an evidence against the null hypothesis H_0. Thus it can be concluded that the underlying means are not equal.

v) A 95% confidence interval for the i^{th} rate mean is given by:

$$\bar{y}_i \pm t_{8;0.975} \times \frac{\hat{\sigma}}{\sqrt{3}} \quad \text{for } i = 1, 2, 3, 4$$

Here: $t_{8;0.975}$ is the 97.5% critical point of a t-distribution with 8 degrees of freedom and equals 2.306.

The common error variance σ^2 is estimated as:

$$\hat{\sigma}^2 = \frac{\text{Residual Sum of Squares}}{8} = 0.155$$

The 95% confidence intervals are tabulated in the following table.

Rate	Observed Mean	95% C.I (Transformed Scale)		95% C.I. (Original Scale)	
		LCI	UCI	LCI	UCI
1	2.992	2.469	3.516	11.810	33.635
2	4.827	4.303	5.350	73.954	210.623
3	5.716	5.193	6.239	179.950	512.505
4	6.575	6.052	7.098	424.770	1209.763

6] i]

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

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

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

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

$$= \frac{661.20}{\sqrt{54.90}\sqrt{89236}} = 0.945$$

ii] Fitted linear regression Equation

The coefficients of the regression eqn are:

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

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \times \bar{X} = \left(\frac{562}{10}\right) - 12.04\left(\frac{39}{10}\right) = 9.23$$

$$\boxed{\therefore Y = \hat{\alpha} + \hat{\beta}X = 9.23 + 12.04X}$$

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.

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

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

ii] $S_{xx} = 277.5 - \frac{45^2}{10} = 75$

$S_{yy} = 43956 - \frac{664^2}{10} = 2482.4$

$S_{xy} = 2602 - 664\left(\frac{45}{10}\right) = -296$

$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = -0.686$

iii] Null hypothesis $H_0 : \rho = 0$ against $H_1 : \rho < 0$
We will assume that the data comes from a bivariate normal distribution.

$\therefore r = 0.5\left[\ln\left(\frac{1-0.686}{1.686}\right)\right] = -0.8404$

Fisher's standardized statistic $= \frac{(-0.8404 - 0)}{\left(\sqrt{1/7}\right)}$

$= -2.22$

From above we get p-value as,

$p\text{-value} = P(z < -2.22) = 0.013$

The p-value obtained is quite small and hence shows a strong evidence to reject the null hypothesis with 95% C.I.

Hence, the marks obtained and hours spent on social media are negatively correlated.

iv) Beta $= \frac{S_{xy}}{S_{xx}} = \frac{-296}{75} = -3.9467$

Alpha = mean of y - beta × mean of x
$= \frac{644}{10} + 3.9467\left(\frac{45}{10}\right) = 82.16$

Therefore, the fitted line is $\boxed{y = 82.16 - 3.9476x}$

v) $R^2 = -0.686^2 = 0.4706$

vi) From the fitted line equation we can conclude that, for ~~ee~~ every additional hour spent on social media per day, the total marks reduce by 3.95.

10] H0 : There is no difference among industries
H1 : At least one industry differs significatly from the overall mean
$SS_R = 19(5^2 + 10^2 + 8^2) = 3591$
Mean of resignation $= \frac{(27+36+30)}{3} = 31$
$SS_B = 20((27-31)^2 + (36-31)^2 + (30-31)^2) = 840$
$F_{2,57} = (840/2)/(3591/57) = 6.667$

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.

8] i) Factor analysis is:
A method for reducing dimensionality of data
It seeks to identify key components necessary to model and understand data.

Original variables may be correlated with each other while newly identified principal components are chosen to be uncorrelated, linear combinations of the original variables of the data which maximise the variance.

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

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

Q7] i] $S_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 4789.42$

$S_{xy} = 2176.84$

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

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

Hence, the fitted regression equation is

$\boxed{\hat{y} = 10.79 + 0.4545x}$

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

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

$S_{xy} = 2176.84$

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

$= 22.20$

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

To test $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

the test statistic is $\frac{\hat{\beta} - 0}{s.e.(\hat{\beta})} = \frac{0.4545}{0.0681} = 6.674$

Beta has a t-distribution with n-2 d.o.f. if we assume the errors of the regression are i.i.d $N(0, \sigma^2)$ random variables

Critical value : $t_{9,0.025} = 2.68$

Since the critical value at 95% ~~C.I. is~~ level of significance is less than the test statistic, there is sufficient evidence to reject the null hypothesis.

iii) $S_{yy} = 1189.21$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{2176.84}{\sqrt{4789.42 \cdot 1189.21}} = 0.91$$

iv) The estimated value of y corresponding to $x^3 = 25$

$$\hat{\sigma}^2 = 10.79 + 0.4545 \times 25 = 22.15$$

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

$$\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]\hat{\sigma}^2 = \left[\frac{1}{11} + \frac{84.0889}{478942}\right] 22.20 = 2.41$$

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

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

Using t_9 distribution, 95% C.I for intervals for mean and individual responses are:

$22.20 \pm 2.262\sqrt{2.41}$ and $22.20 \pm 2.262\sqrt{24.61}$
$= (18.7, 25.7)$ & $(11.0, 33.4)$

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