Probability and Statistics

Sarthak 43

Assignment 2

Q1]

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

$\Sigma x = 850$

$\Sigma y = 1737$

$$S_{XY} = \Sigma xy - \frac{\Sigma x \Sigma y}{n} \qquad S_{xx} = \Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2$$

$$= 34{,}915 \qquad = 20{,}250$$

$$\hat{\beta} = \frac{S_{XY}}{S_{xx}} = 1.724198$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \bar{y} = 173.7 \quad \bar{x} = 85$$

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

$$\hat{y} = 27.143 + 1.7242x$$

b) The gradient of line which is $\hat{\beta} = 1.724$ shows that there exist a positive relationship between x and y. That is when the amount withdrawn increases, the number of hours to visit also increases.

Q2]

~~X = Share prices of Dell~~

i) $S_{xx} = \Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2$ $S_{xy} = \Sigma xy - \frac{\Sigma x \Sigma y}{n}$

$= 301.66$ $= 875.16$

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = 2.90115 \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$= 3.74809$$

$$\hat{y} = 3.748 + 2.901x$$

ii) Assumptions:

a) It follows a Normal distribution

b) The explanatory variabe value all are independent of each other

c) All error terms are constant

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

$$= \cancel{38}\ 387.07447$$

95% CI of β

$$\hat{\beta} \pm t_{10, 2.5\%} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$= 2.90115 \pm 2.52379$$

$$= (0.37736\ ,\ 5.42494)$$

iii) Mean IBM share price $= \mu_0$

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

$\hat{\mu}_0 = 119.78409$ (using $x_0 = 40$ in $\mu_0 = \hat{\alpha} + \hat{\beta} x_0$)

$$\therefore se.(\hat{\mu}_0) = \sqrt{\left(1/n + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)\hat{\sigma}^2} \quad \text{(Table book)}$$

$$= 10.59894$$

95 % CI of μ_0

$$\hat{\mu}_0 \pm t_{10, 2.5\%} \cdot se(\hat{\mu}_0)$$

$$= (96.17866\ ,\ 143.40852\ )$$

Q3] Comments on the plot

- The mean of all 4 distributions seem to be different with distribution D having lowest looking at the ~~spread~~ variation of the data. D is most spread out compared to others but all seem relatively concentrated and symmetric. Hence normality can be assumed but consideration to small sample size should be given.

ii] Assumptions for ANNOVA :

1. population assumed to be distributed normally
2. there is common variance
3. all observations are independent

iii] H_0 : mean difference is same for each city

H_1 : mean of difference is not same

$$SS_{TOT} = S_{YY}$$
$$= \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$
$$= 1116.11$$

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

$$SS_{RES} = 1116.11 - 516.11 = 600$$

Table

Source	DOF	SS	MSS
Regression	3	516.11	172.04
Residual	24	600	25
Total	27	1116.11	

Under Ho,

$$F = \frac{MSS_{REG}}{MSS_{RES}} = \frac{172.04}{25} = 6.88$$

T.S = 6.88

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

Hence there exists sufficient evidence to reject Ho at 5% level.

Hence there are underlying difference between the city means.

iv) Analysis of mean difference

Since, $\bar{y}_{1*} = 9.14$ $\bar{y}_{3*} = 1.14$ $\bar{y}_{1*} > \bar{y}_{2*} >$

$\bar{y}_{2*} = 6.29$ $\bar{y}_{4*} = -1.86$ $\bar{y}_{3*} > \bar{y}_{4*}$

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

The least singnificance difference b/w any pair of mean is $t_{24, 0.025} \times \hat{\sigma}\sqrt{\frac{1}{7}+\frac{1}{7}} = 5.52$

Now we can examine the difference b/w each of the pairs of means. If difference is less than the least significance diff.

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 diff are less than 5.52, we there pairs to show that they have no sign. diff

4] a) Mathematical Model of one way ANOVA is given by

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

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 mean response

e_{ij} is the error term

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

b) H_0: mean claim amt of 5 companies are equal
H_1: mean claim amt of 5 companies are not equal

$n_1 = 7$ $n_4 = 7$
$n_2 = 8$ $n_5 = 5$
$n_3 = 6$ $n_6 = 33$

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

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

$$CF = \left(\sum_{i=1}^{5} \sum_{j=1}^{n_i} y_{ij}\right)^2 / n = 104385.94$$

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

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

$$SS_{REG} = \sum_{i=1}^{5} \left(\sum_{j=1}^{n_i} y_{ij}\right)^2 / n_i - CF$$

$$= 22325.69$$

Source of Var	d.f	SS	MSS
Regression	4	22326	5581.42
Residual	28	47604	1700.16
	32	69930	

$$T.S = \frac{MSS_{REG}}{MSS_{RES}} = 3.283$$

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

We have sufficient evidence to reject H_0.

c) O_i

Paper Clear	Salary per annum (in Rs Lakhs) 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

E_i

Paper Cleared	Salary per annum (in Rs Lakhs) 3-5	5-8	8-10	10-12
0-3	27.42	23.08	14.43	11.06
4-6	15.15	12.758	7.97	6.113
7-9	14.43	12.152	7.59	5.882

H_0: Salaries are independ of papers cleared

H_1: Salaries aren't independent of papers cleared

$$T.S = \sum \frac{(O_i - E_i)^2}{E_i} = 48.81146$$

$DOF = (r-1)(c-1)$

$= 2 \times 3 = 6$

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

Hence we have sufficient evidence to reject Ho.

They are dependant on each other.

Q5] i) Assumptions underlying ANOVA

- Populations are normal
- There is common variance
- Observations are independent

The 4 sample var are different from each other. Hence common var assumption won't hold.

ii) Mean at rate 1 = 4.52443
for $\sqrt{x}$

Mean at rate 2 = 0.12127
for $\log(x)$

iii) The scientist is correct with this argument as log transformation helps hold the equal variance assumption required for ANOVA analysis.

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

$j = 1, 2, 3$

b) For ANOVA,

$H_0: \tau_i = 0$

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

ln (x

Rate	Mean	Var	$y_{i\cdot}^2$
1	2.882	0.163	80.582
2	4.827	0.121	208.678
3	5.716	0.186	284.053
4	6.575	0.148	388.06
Total	20.110	0.618	973.373

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

$$= \frac{973.373}{3} - \frac{(3 \times 20 \times 110)^2}{12}$$

$$= 21.153$$

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

$$= 2 \times 0.618$$

$$= 1.236$$

Variation	DOF	SS	MSSS
Rate	3	21.153	7.051
Residual	8	1.236	0.155
Total	11	22.38	

$T.S = 45.638$

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

c) Given the observed F statistic value as much larger than this, hence p value is almost 0.
Or we can say there is more than enough evidence to reject H_0

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

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

$S_{xx} = \Sigma x^2 - n\bar{x}^2$
$= 54.8$

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

$r = 0.94466$

ii) Fitted regression

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 9.2285$

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

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

iii) $SS_{TOT} = S_{yy} = 8923.6$

$SS_{REG} = \dfrac{S_{xy}^2}{S_{xx}} = 7963.30482$

$SS_{RES} = SS_{TOT} - SS_{REG} = 960.29808$

$R^2 = \dfrac{S_{xy}^2}{S_{xx} S_{yy}} = 89.2387\%$

Comment :

Model is a good fit as explains almost 89% of variation. Any model over $> 75\%$ is considered a good fit. Further hypothesis test could be carried out.

Q7] i) Linear regression eq

$S_{xx} = 4788.4221$ $\qquad S_{yy} = 1189.20767$

$S_{xy} = 2176.84$

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

$$= 0.45451$$

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

$$= 10.78056$$

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

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

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

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

$$= 22.20136$$

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

$$t_{8,\,2.5\%} = 2.262$$

$$= 6.67568$$

Hence we have sufficient evidence to reject H_0.

$\beta \neq 0$. Hence linear relationship exists.

Assumptions:

1. Population have common variance
2. Population follows the normal distribution
3. the observation are independent.

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

$= 0.81213$

iv) For mean response (μ_0)

$x_0 = 25$

$\hat{\mu}_0 = 22.15331$

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

$= 1.55181$

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

95% CI for $\mu_0 = \hat{\mu}_0 \pm t_{9, 2.5\%} \cdot se(\hat{\mu}_0)$

$= 22.15331 \pm 3.51018$

$(18.64313, 25.66348)$

For individual response

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

$= 22.15331$

$se(\hat{y}_0) = 4.96079$

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

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

$= (10.932, 33.37462)$

iv) There is clearly some relationship b/w the residuals and the explanatory variable values. Hence our assumption of normality might be wrong. The model is not a good fit.

Q8] i)

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 number of variables in regression model. PCA is a technique for reducing the dimensionality of such datasets.

While newly indentified PC are chose to be
- Uncorrelated
- linear combination of original variables of data
- Which maximizes variance

ii)

PC	Diagnal entry (PC_i)	PC_i / (sum PC_i over 1 to 5)
PC 1	0.456	65%
PC 2	0.137	18.5%
PC 3	0.08	11.4%
PC 4	0.0165	2.4%
PC 5	0.012	1.7%
	0.7015	100%

(iii) As 1st 3 ~~few~~ PC explain over 95% of total variance, dimensionality can be reduced to 3 for this dataset.

The 1st 3 PC can then be used for building further classified or regression modelling purpose

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

ii) $S_{xx} = 75$

$S_{yy} = 2482.4$

$S_{xy} = -296$

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

Hence there is negative correlation

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

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

Test statistic $= -2.07893$

$Z_{5\%} = 1.6448$

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

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

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) $R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$

$= 47.05983\%$

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

vi) $x_0 = 1$

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

$= 78.2133$

Q10]

x	0.27	0.36	0.3
y	0.05	0.1	0.08

$S_{xx} = 0.0042$

$S_{yy} = 0.00126$

$S_{xy} = 0.0022$

$SS_{TOT} = 0.00126$ $\quad SS_{REG} = \frac{S_{xy}^2}{S_{xx}} = 0.00115$

$SS_{RES} = 0.00011$

ANOVA Table

Source	DOF	SS	MSS
Reg	1	0.00115	0.00115
Res	1	0.00011	0.00011
Total			

$H_0 : \beta = 0$

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

$$T.S = \frac{0.00115}{0.00011} = 10.45455$$

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

$$F_{cal} < F_{tab}$$

We have insufficient evidence to reject H_0.