Assignment 2

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

$\Sigma x_i = 850$ $\quad \Sigma y_i = 1737$

$\Sigma x_i^2 = 92500$ $\quad \Sigma y_i^2 = 372717$

$\bar{x} = 85$ $\quad \bar{y} = 173.7$ $\quad \Sigma xy = 182560$

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

$$S_{xx} = \Sigma x_i^2 - n\bar{x}^2$$
$$= 92500 - 72250$$
$$= 20250.$$

$$S_{xy} = \Sigma xy - n\bar{x}\bar{y}$$
$$= 182560 - 147645$$
$$= 34915$$

$$\hat{\beta} = \frac{34915}{20250} = \underline{1.7242}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 173.7 - 1.7242(85)$$
$$= 27.143$$

$$Y_i = 27.143 + 1.7242 x_i$$

ii) The gradient or $\hat{\beta} = 1.7242$

A slope gradient of higher than one means that the numbers of hours increases at a faster pace than amount withdrawn

Q4) $\Sigma x = 385.2$, $\Sigma x^2 = 12666.58$

$\Sigma y = 1162.5$ $\Sigma y^2 = 119026.9$

$\Sigma xy = 38191.41$

i $S_{xx} = 12666.58 - 12(32.1)^2$

$= 12666.58 - 12364.92$

$= 301.66$

$S_{xy} = 38191.41 - 12(32.1)(96.875)$

$= 38191.41 - 37316.25$

$= 875.16$

$\hat{\beta} = \frac{875.16}{301.66} = 2.9011$

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

$= 96.875 - (2.9011)(32.1)$

$= 96.875 - 93.12531$

$= 3.7497$

$y_i = 3.75 + (2.9)x_i$

ii) 95% CI for β.

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

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

$$S_{yy} = \frac{(1162.5)^2}{119026.9} - 12(96.875)^2$$

$$= \cancel{125140622} = 112617.21875$$

$$= \cancel{1238}\ 6409.7125$$

$$\sigma^2 = \frac{1}{10}\left(6409.7125 - 2538.5968\right)$$

$$= 887.074$$

$$\sqrt{\frac{\sigma^2}{S_{xx}}} = \sqrt{\frac{387.074}{301.66}} = \sqrt{1.2831}$$

$$= 1.13276$$

$$2.9011 \pm 2.228(1.13276)$$

$$= 2.9011 \pm 2.5238$$

$$= (0.3773, 5.4249)$$

iii) $$y_i = 3.25 + (2.9)40)$$

$$= 119.25.$$

$$s.e(y_i) = 10.5989$$

95% CI for y_i

$$= 119.75 \pm 2.228(10.5989)$$

$$= 119.795 \pm 23.6143.$$

$$= (96.1796, 143.4084)$$

classmate
Date
Page

3i) • The difference in commute time between peak and non peak hours is maximum in City A followed by B, C and D in that order.

• It is clear by the plots that the variation in each of the cities is different having a different mean.

ii) Assumption for ANNOVA test include.
• population must be normal
• population must have a common variance
• Observations are independent

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

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

$$= 1116.11$$

$$SS_B = \frac{1}{7}\left(64^2 + 44^2 + 8^2 + (-13)^2\right) - \frac{103^2}{28} = 51.611$$

he variance of the

classmate
Date
Page

$SS_R = SS_T - SS_B = 600$

ANNOVA -

Source.	df	SS	MSS	F
	3	516.11	172.04	
	24	600	25	6.88.
Total	27	1116.11		

Under H_0 . $F = \frac{172.04}{25} = 6.88.$

as we use the $F_{3,24}$ dist.

The 5% critical value is 3.009 so we have sufficient evidence to reject H_0

(a) Question 3 part 2.

b) $\frac{T^2}{N} = \frac{(\Sigma y_1 + \Sigma y_2 + \Sigma y_3 + \Sigma y_4 + \Sigma y_5)^2}{N}$

$= 104385.939$

$SS_{reg} = \frac{(354)^2}{7} + \frac{(386)^2}{8} + \frac{(87)^2}{6} + \frac{(645)^2}{7} + \frac{(384)^2}{5} - \frac{T^2}{N}$

$= 22325.6892$

$SS_{TOT} = 69930.0606$

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

$= 47604.37143$

ANOVA

	df	SS	MS	F
Reg	4	22325.69	5581.4275	3.28
Res	28	47604.363	1700.154	
Tot	33	69930.066		

$F_{4,28}$ is 2.21

and calc. value = 3.28.

So we have sufficient evidence to reject H_0

c) H_0 : They are independent of each other

H_1 : They are not independent.

E_i

	3-5	5-8	8-10	10-12
0-3	27.42	23.089	14.43	11.06
4-6	15.15	12.76	7.97	6.14
7-9	14.43	12.15	7.59	5.82

classmate
Date
Page

$\chi^2_{cal} = \sum \frac{(E_i - O_i)^2}{E_i}$

$= 89.929$

Since $\chi^2_{cal} > \chi^2_{tab}$, we have sufficient evidence to reject H_0.

Correlation coefficient =

6) $S_{YY} = \cancel{89}\ \cancel{89} = 8923.6$

$S_{XY} = 661.2$

$S_{XX} = 54.9$

$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}} = 0.94466$

ii) $\hat{\beta} = \frac{S_{XY}}{S_{XX}}$ $\alpha = \bar{y} - \hat{\beta}\bar{x}$,

$\hat{\beta} = 12.0437$

$\alpha = 9.22957$

$y_i = 9.22957 + 12.0437 x_i$

iii) $SS_{RES} = \left[S_{YY} - \frac{S^2_{XY}}{S_{XX}} \right]$

$= 960.2951$

$SS_{TOT} = 8923.6$

$SS_{REG} = 7963.305$.

iv) $R^2 = \dfrac{SS_{REG}}{SS_{TOT}} = \dfrac{7963.305}{8923.6} = 0.89242$

Hence 89% variation is showed by our model.

7) $S_{xy} = \Sigma x^2 - n(\bar{x})^2$
$= 4789.4221$

$S_{yy} = \Sigma y^2 - n(\bar{y})^2$
$= 1189.2077$

$S_{xy} = \Sigma(x-\bar{x})(y-\bar{y})$
$= 2176.84$.

$\hat{\beta} = \dfrac{S_{xy}}{S_{xy}} = \dfrac{2176.84}{4789.4221}$

$= 0.45451$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
$= 1079056$

$y = 10.79056 + 0.45451x$.

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

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

$$Z_c = \frac{\hat{\beta} - \beta}{\hat{\sigma}/\sqrt{S_{xx}}}$$

$$= \frac{0.45451}{\frac{\sqrt{22.20136}}{\sqrt{4789.4221}}}$$

$$= \frac{0.45451}{0.06808} = 6.67611$$

$Z_{tab} = 2.262.$

Since $Z_{cal} > Z_{tab}$

we have insufficient evidence to reject H_0.

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

$= 0.91213.$

iv) given $x^3 = 25 \Rightarrow 2.924$

for mean response

$\hat{\mu}_0 = 10.79 + 0.4545(2.924)$

$\hat{\mu} = 12.11773.$

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

$$= 22.20136\left[\frac{1}{11} + \frac{(2.724-15.83)^2}{4789.6221}\right]$$

est var $(\hat{\mu}_0) = 2.790417$

95% CI

$$\hat{\mu}_0 \pm 2.262\sqrt{2.79042}$$

$$= 12.11773 \pm 2.262\sqrt{2.79042}$$

$$\Rightarrow (8.3392,\ 15.89629)$$

For individual response.

$\hat{y}_0 = 12.11773$

$V(\hat{y}_0) = 24.991777$

95% CI

$$\hat{y}_0 \pm 2.262\sqrt{24.991777}$$

$$\Rightarrow 12.11773 \pm 2.262\sqrt{24.9917}$$

$= (0.8096, 23.4259)$

v) We cannot say with certainty that it shows a normal distribution as some pattern is seen.

8) The main purpose of factor analysis is to reduce many individual items into a fewer numbers of dimensions.

Principle component analysis is a method to decrease dimension without data loss.

PC	Diagonal Entries	$Pc_i / Sum(PC_i)$
PC1	0.456	65%
PC2	0.137	18.5%
PC3	0.08	11.4%
PC4	0.0165	2.4%
PC5	0.72	1.7%
	0.1015	100%

iii) As 1st 3 principle components explain over 95% of total variance, dimension can be reduced to 3 for the data set.

i) There appears to be a negative correlation

ii) $S_{xx} = 75$ $\quad S_{xy} = -296$

$S_{yy} = 2482.4$

$$r = \frac{-296}{\sqrt{2482.4(75)}} = -0.686$$

This shows (-ve) coefficient.

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

$w = \tanh^{-1}(r)$

$= -0.84036$

$$w \sim N\left(0, \frac{1}{7}\right)$$

$$z_c = \frac{-0.84036 - 0}{\sqrt{\frac{1}{7}}} \Rightarrow -2.22$$

$z_{tab} = -1.96$

Since $z_c < z_{tab}$, we have sufficient evidence to reject the null hypothesis

(ii) $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = -3.94667$

$\alpha = \bar{y} - \hat{\beta}\bar{x} = 82.16$

$y = 8216 - 3.94667x$

(i) $R = r^2 = 0.470546$

So ~~to~~ 47% of variation is shown by our model.

when $x = 1$

$y = 82.16 - 3.94667$

$= 78.21333$

when $x = 2$.

$y = 74.2666$

$\therefore$ for every additional 1 hr spent, expected change is -3.94667.

(b) $S_{xx} = 0.0042$

$S_{yy} = 0.00126$

$S_{xy} = 0.0022$

$SS_{TOT} = 0.00126$.

$SS_{REG} = \frac{S_{xy}^2}{S_{xx}} = 0.00115$,

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

$= 0.0011$

ANNOVA TABLE

Source	df	SS	MSS	F
Reg	1	0.00115	0.00115	10.454
Res	1	0.00011	0.00011	
Tot	2	0.00126		

$H_0 : \hat{\beta} = 0$

$H_1 : \beta \neq 0$

$F_{1,1,5\%} = 161.42.$

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

Hence we cannot reject H_0