Topic 8: Simple Linear Regression Solutions

Q1

c) r = 1

Q2

c) Neither of the above is necessarily true

Q3

a) X

Q4

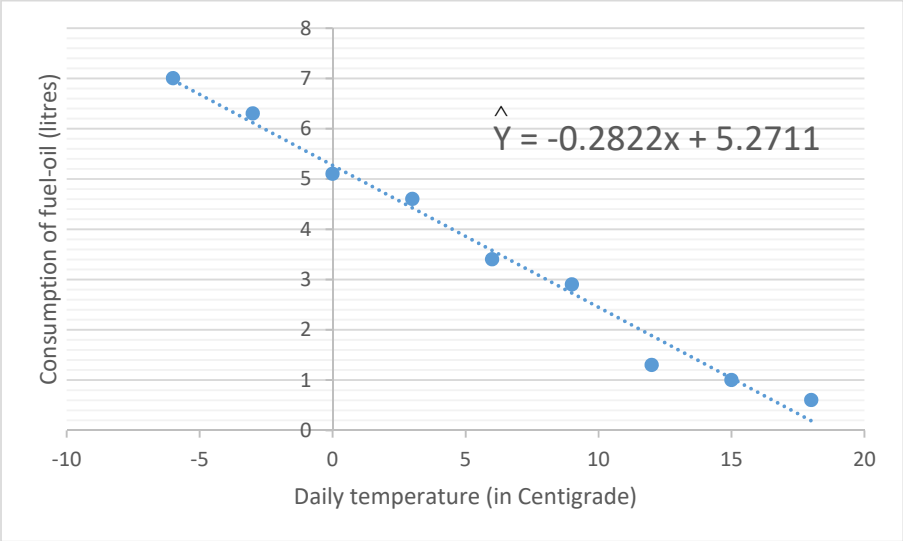
b) Z

Q5

c) Neither of the above

Q6

a) & b) & c)



d) Predicted consumpotion level, $\hat{Y}$ = -0.2822(7.5)+5.2711 = 3.1546 litres

e) r = -0.992. the linear relationshiop between daily temperature and daily fuel consumption is negative and very strong.

Topic 8: Simple Linear Regression Solutions

Q7

a) b_0 = 51.915, if we apply the regression equation at X=0, the average assessed value will be 51.915 hundred thousands dollars. However, it is obviously impossible and meaningless.

b_1 = 16.633, for each additional one thousand square feet in gross area, the estimated assessed value increases by 16.633 hundred thousands dollars.

b) $\hat{Y}$ (when X = 1.75) = 51.915 + 16.633(1.75) = 81.02275
Thus, predicted assessed value is 81.02275 hundred thousands dollars

c) Coefficient of determination, $R^2 = (0.812)^2 = 0.6593$
Interpretation: 65.93% of total variation in assessed value can be explained by variation in gross area.

Q8

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9494729
R Square	0.901498788
Adjusted R Square	0.889186136
Standard Error	42.18594454
Observations	10

ANOVA

	df	SS	MS	F	Significance F
Regression	1	130301.4097	130301.4097	73.2172747	2.68223E-05
Residual	8	14237.23133	1779.653916		
Total	9	144538.641			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	26.72397637	26.54250428	1.006837037	0.343491818	-34.48314821	87.93110095	-34.48314821	87.9311
Reported (X)	0.571887175	0.066834942	8.556709339	2.68223E-05	0.417765521	0.726008828	0.417765521	0.726009

a) Slope = b_1 =0.5719. For each additional thousand units increase in reported newsstand sales, the mean audited sales will increase by an estimated 0.5719 thousand units.

b) $\hat{y} = 26.7240 + 0.5719(300)$ =198.294 thousands. The predicted audited newsstand sales for a magazine that reports newsstand sales of 300,000 is 198.294 thousands

c) Let β_1 be the population slope coefficient

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

$\alpha = 0.05$ Critical Value $= \pm t_{\alpha/2,n-2} = \pm t_{0.05/2,8} = \pm 2.3060$

Reject H_0 if $t < -2.3060$ or $t > 2.3060$ or Reject H_0 if p-value < 0.05

$$t = \frac{b_1 - 0}{S_{b_1}} = \frac{0.5719 - 0}{0.0668}$$

$t = 8.5613$ or $p\text{-value} = P(t \leq -8.5613) \times 2 = 2.68223E-05$

Since t =8.5613 > 2.3060 or p-value $= 2.68223E-05 < 0.05$,

We reject H_0 at $\alpha = 0.05$

There is sufficient evidence to point out that the slope of the population regression line is not 0.

d) 95% confidence interval for β_1:

$b_1 \pm t_{\alpha/2,n-2} s_{b_1}$

$= 0.5719 \pm (2.306)(0.0668)$

$= [0.4179, 0.7259]$

We are 95% confidence that population slope β_1 is between 0.4179 and 0.7259.

Q9

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.850060796
R Square	0.722603356
Adjusted F	0.710542633
Standard E	194.5953946
Observatio	25

ANOVA

	df	SS	MS	F	Significance F
Regressior	1	2268776.545	2268777	59.91376452	7.51833E-08
Residual	23	870949.4547	37867.37		
Total	24	3139726			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	177.1208202	161.0042766	1.1001	0.282669853	-155.9419	510.1835404	-155.9419	510.1835404
X Variable	1.065143906	0.137608412	7.740398	7.51833E-08	0.780479219	1.349808593	0.780479219	1.349808593

a) From the Excel output,

$b_0 = 177.1208$; $b_1 = 1.0651$

$\hat{y} = 177.1208 + 1.0651x$ (where $\hat{y}$ is predicted monthly rental cost for apartments and x is the size of the apartment)

b) b_0 = 177.1208, if we apply the regression equation at X=0, the expected monthly rental will be \$177.1208. However, it is obviously impossible and meaningless. Therefore, 177.1208 has no practical interpretation.
b_1 = 1.0651, for each increase of 1 square foot in space, the expected monthly rental is estimated to increase by \$1.065.

c) $\hat{y} = 177.1208 + 1.0651x = 177.1208 + (1.0651)(1000) = \$1242.2208$

d) An apartment with 500 square feet is outside the relevant range for the independent variable. So, it is not appropriate to use the model to predict the monthly rent for apartments that have 500 square feet

e) From the Excel Output, $b_1 = 1.065, S_{b_1} = 0.1376$
Let β_1 be the population slope in the population linear regression equation
$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$ (exists a linear relationship)
$\alpha = 0.05$, Critical Value $= \pm t_{\alpha/2,n-2} = \pm t_{0.025,23} = \pm 2.0687$
Reject H_0 if $t < -2.0687$ or $t > 2.0687$
$t = \dfrac{b_1 - 0}{S_{b_1}} = \dfrac{1.065 - 0}{0.1376} = 7.7398$
Since t = 7.7398 > 2.0687, we reject H_0 at α =0.05.
There is sufficient evidence that there exists a linear relationship between the size of the apartment and the monthly rent.

f) 95% confidence interval for β_1:
$b_1 \pm t_{\alpha/2,n-2} s_{b_1} = 1.065 \pm (2.0687)(0.1376) = [0.7803, 1.3497]$
We are 95% confidence that population slope β_1 is between 0.7803 and 1.3497.

Q10

a) From the Excel output,
a = 1.646 b = 0.104
$\hat{Y}$ = 1.646 + 0.104 X (where Y is earnings and X is sales)
a = if we apply the regression equation at X =0, the average earnings will be 1.646 million dollars. However, it is obviously impossible and meaningless.
In fact, the regression equation can only be applied for sales between 10.4 to 71.7 million dollars
b = for each increase of 1 million dollars in sales, earnings will have an average increase of 0.104 million dollars.

b) From the Excel output, sample coefficient of correlation between sales & Earnings = r = 0.888

Let β_1 be the population slope coefficient

$H_0 : \beta_1 \leq 0$

$H_1 : \beta_1 > 0$

$\alpha = 0.05$

Reject H_0 if p-value < 0.05

From the output, t-Stat=6.091, and p-value = 0.000117/2 = 0.0000585 <0.05

$\therefore$ We reject H_0. There is sufficient evidence that there exists a positive linear relationship between the sales and earnings.

c) Coefficient of determination = R^2= 0.788

78.8% of total variation in earnings can be explained by variation in sales.

d) X =35, $\hat{Y}$ = 1.646+0.104(35) = 5.286 million dollars

e) From the Excel Output, b_1= 0.1039, S_{b_1} =0.0171

$H_0 : \beta_1 \geq 0.15$

$H_1 : \beta_1 < 0.15$

$\alpha = 0.05$ Critical Value= $= -t_{\alpha,n-2} = -t_{0.05,10} = -1.8125$

Reject H_0 if $t < -1.8125$

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$t = \frac{0.1039 - 0.15}{0.0171}$$

$t = -2.6959 < -1.8125$

We reject H_0. There is sufficient evidence that the population slope in the regression model is not 0.15.

Q11

a) $\hat{Y}$ = 2.0289+0.04643X

b) Slope= 0.04643

For each increase of purchasing frequency, the average purchase quantity per buyer will have an average increase of 0.04643kg

c) When X=20, $\hat{Y}$ = 2.0289+0.04643(20) = 2.9575

d) Let β_1 be the slope of the population regression line

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$\alpha = 0.05$

Reject H_0 if p-value < 0.05

$t = \frac{0.04643 - 0}{0.022302} = 2.082$

p-value= $P(t \leq -2.082) + P(t \geq 2.082)$ = (0.05, 0.1) > 0.05

(or directly, from the Excel output p-value = 0.0825)

Therefore we do not reject H_0. There is insufficient evidence to point out that the slope of the population regression line is not 0.

e) $R^2 = 0.419$

41.9% of variation in y (the average quantity per buyer) can be explained by the variability in x (purchasing frequency).

As the is R^2 is not very large, it shows that this model is not very good to predict y (the average quantity per buyer) or explain by x (purchasing frequency).

Q12

a) $\hat{Y}$ = 15-2X

Y(X=2) = 15-2(2) = 11

Y(X=4) = 15-2(4) = 7

Y(X=6) = 15-2(6) = 3

b) SSE = $\Sigma(Y_i - \hat{Y}_i)^2$ = $(10-11)^2 + (12-11)^2 + (6-7)^2 + (8-7)^2 + (1-3)^2 + (5-3)^2 = 12$

c) $\bar{Y}$ = (10+12+6+8+1+5)/6=7

SST = $\Sigma(Y_i - \bar{Y})^2 = (10-7)^2 + (12-7)^2 + (6-7)^2 + (8-7)^2 + (1-7)^2 + (5-7)^2 = 76$

d) $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{12}{76} = 0.8421$

Interpretation: 84.21% of the variation in Y can be explained by the variability in X

e) ∵ slope = -2, the relationship between X & Y is negative

coefficient of correlation, r = $-\sqrt{r^2}$ = -0.9177

f) Let β_1 be the population slope coefficient in the regression model

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Let $\alpha = 0.05$

∵ p-value = 0.00989 < 0.05

Thus, H_0 is rejected. There is sufficient evidence that there is a linear relationship between X & Y. i.e. X is an important factor affecting Y

Q13

a) $\hat{Cost}$ = 16593.65 + 2600.68 NUMSTORES

b) Point estimate: $b_1 = 21.76$

95% CI: $b_1 \pm t_{0.025,12} s_{b_1}$ =21.76 ± 2.1788*7.37=21.76 ± 16.0578 = [5.7022, 37.8178]

We are 95% confident that true value of β_1 is between 5.7022 and 37.8178.

c) Suppose the population linear regression equation relating COST and NUMSTORE is

$\hat{Cost} = \beta_0 + \beta_1$ NUMSTORE+ ε

$Ho: \beta_1 \geq 3000$

$H_1: \beta_1 < 3000$

α =0.05, n=14, d.f. =n-2=12, reject H_0 if t < -1.7823

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{2600.68 - 3000}{267.66} = \frac{-399.32}{267.66} = -1.49$$

Since t = -1.49 (> -1.7823), Do not reject Ho. There is insufficient evidence that each new store adds is less than 3000 USD of cost.

d) According to R^2 of the two regression models, NUMSTORE explained 88.72% of variations in COST, STORESIZE explained 42.09% of variations in COST. Therefore, NUMSTORE is more useful to explain the variations in store setup cost.

e) r = 0.9419, there is a strong positive relationship between number of stores in an area and store setup cost

OR b_1=2600.68, for each additional store added, the store setup cost is estimated to be increased by 2600.68USD on average.

f) $\hat{Cost}$ =10961.14+21.76 STORESIZE=10961.14+21.76*1000=32721.14USD.

The average total setup cost for a business area that needs 1000 square feet of store area is estimated to be 32721.14 USD.

Q14

a) $\hat{SALARY}$ = 4.5538+2.6708*GPA

Intercept: If applying the regression equation at GPA=0, the average starting salary is $4553.8.

However, the regression equation can only be applied for GPA within $[2.23,3.85]$

Slope: For each increase of one unit in GPA, the starting salary will have an average increase of $ 2670.8

b) It is inappropriate for predicting the starting salary by this regression equation since GPA=4.0 is not within $[2.23,3.85]$

c) Let β_1 be the population slope in the population linear regression equation:

H_0: $\beta_1 \leq 0$

H_1: $\beta_1 > 0$

α =0.01, Reject H_0 if p-value <0.01

From the excel output, p-value =P($t \geq 8.475337535$) = 0.000375688/2

≈ 0.0001878

Since p-value=0.0001878<0.01

We reject H_0 at α=0.01

There is sufficient evidence that there is a positive linear relationship between SALARY and GPA

d) From the relationship between SALARY and GPA, R^2= 0.9349

From the relationship between SALARY and IELTS, $R^2=(0.92)^2$ =0.8464

GPA explained 93.49% of variations in SALARY while IELTS explained 84.64%

=> GPA is more useful to explain the variations in starting salary.

Q15

a) $\hat{Y} = 250 - 2X$

Intercept: If applying the regression equation at line speed, $X = 0$, the average number of defective parts is 250. However, the regression equation can only be applied for X within $[20,60]$

Slope: For each increase of one foot in line speed, the number of defective parts will have any average decrease of 2

b) When $X = 55$,

$\hat{Y} = 250 - 2(55) = 140$ defective parts

c) Let β_1 be the population slope in the population linear regression equation

$H_0 : \beta_1 \geq 0$ $H_1 : \beta_1 < 0$

At $\alpha = 0.05$,

Reject H_0 if t < -2.3534 *or* Reject H_0 if *p*-value < 0.05

From the Excel output,

t = -5.7966 *or* *p*-value = P(t≤-5.7966) = 0.0103/2 = 0.00515

Since t < -2.353 *or* Since *p*-value = 0.00515 < 0.05,

we reject H_0 at $\alpha = 0.05$

There is sufficient evidence that line speed and number of defective parts found are negatively related.

d)

X_i	Y_i	$\hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$	$(Y_i - \bar{Y})^2$
20	220	210	100	1764
20	200	210	100	484
40	160	170	100	324
60	130	130	0	2304
40	180	170	100	4

$SSE = 100 + 100 + 100 + 0 + 100 = 400$

$\bar{Y} = \frac{220 + 200 + 160 + 130 + 180}{5} = 178$

$SST = 1764 + 484 + 324 + 2304 + 4 = 4880$

e) $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{400}{4880} = 0.9180$

$1 - R^2$ = 1-0.9180 = 0.0820 = 8.2% of total variation is unexplained by the estimated regression equation

f) Since the slope is negative, the correlation coefficient is also negative.

r $= -\sqrt{R^2} = -\sqrt{0.9180} = -0.9581$

which indicates there is a strong negative linear relationship between Y and X.

Q16

a) There is a strong positive correlation between DAMAGE and DISTANCE.

b) From Excel output, we can find that b_0 = 10.2779 and b_1 = 12.5254*0.3927=4.9187

$\therefore \quad \hat{Y} = 10.2779 + 4.9187\,X$

c) For each increase of one additional mile in DISTANCE, the estimated average amount of DAMAGE will increase by 4918.7 dollars.

d) H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Use t- test for population slope β_1

Reject H_0 if p-value < α= 0.05

$t = \frac{b_1 - 0}{S_{b_1}} = \frac{4.9187 - 0}{0.3927} = 12.5254$ (from Excel output)

p-value = $2 \times P(t \geq 12.5254) = 1.2478\text{E-}08$. (from Excel output)

Since p-value = 1.2478E-08 < $\alpha = 0.05$, we reject H_0.

There is sufficient evidence that there is linear relationship between DISTANCE and DAMAGE

e) Since

$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{69.750975}{911.51733} = 0.9235$

92.35% of the variation in amount of DAMAGE can be explained by the variability in the DISTANCE.

f) Yes. Since the coefficient of determination is very high. It indicates that the regression line fits the data set very well.
Or: DISTANCE is a good predictor for the amount of DAMAGE.

g) Correlation coefficient r=0.9610

h) The 95% confidence interval of β_1 is [4.0708, 5.7678] from Excel output.