BRAC UNIVERSITY
Inspiring Excellence

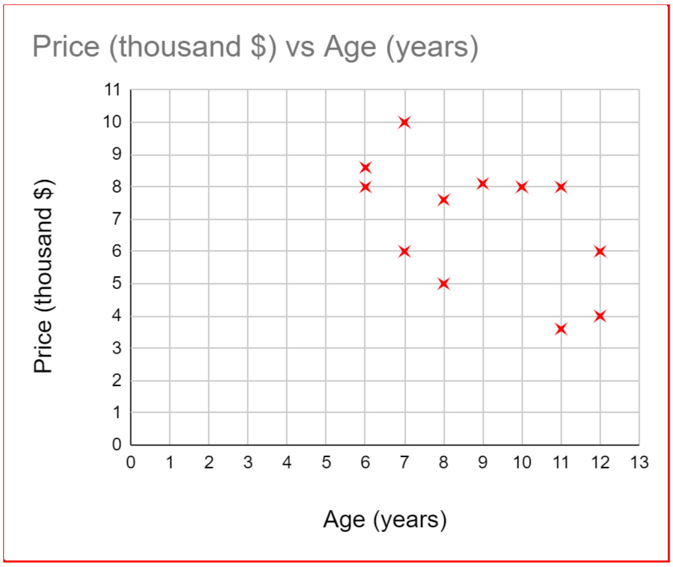
STA201 Assignment 3 Solution

Problem 1

The owner of a second-hand car dealership wants to study the relationship between the age of a car and its selling price. Listed below is a random sample of 12 used cars sold at the dealership during the last year.

Age (years)	9	7	11	12	8	7	8	11	10	12	6	6
Price (thousand $)	8.1	6	3.6	4	5	10	7.6	8	8	6	8.6	8

a. Draw a scatter diagram and comment on the relation between the age of the car and its selling price.



From the scatter diagram, a negative correlation between the age of a car and its selling price can be observed.

b. Determine the Pearson correlation coefficient and the coefficient of determination and interpret it.

	Age (years) (X)	Price (thousand $) (Y)	X^2	Y^2	XY
	9	8.1	81	65.61	72.9
	7	6	49	36	42
	11	3.6	121	12.96	39.6
	12	4	144	16	48
	8	5	64	25	40
	7	10	49	100	70
	8	7.6	64	57.76	60.8
	11	8	121	64	88
	10	8	100	64	80
	12	6	144	36	72
	6	8.6	36	73.96	51.6
	6	8	36	64	48
Sum:	107	82.9	1009	615.29	712.9

$$\bar{x} = \frac{\sum_{i=1}^{12} x_i}{n} = \frac{107}{12} = 8.9167;\ \bar{y} = \frac{\sum_{i=1}^{12} y_i}{n} = \frac{82.9}{12} = 6.9083$$

$$r = \frac{\sum_{i=1}^{12}(x_i y_i) - n \bullet \bar{x} \bullet \bar{y}}{\sqrt{\left(\sum_{i=1}^{12} x_i^2 - n\bar{x}^2\right) \times \left(\sum_{i=1}^{12} y_i^2 - n\bar{y}^2\right)}} = \frac{712.9 - (12 \times 8.9167 \times 6.9083)}{\sqrt{\left(1009 - (12 \times 8.9167^2)\right) \times \left(615.29 - 12 \times 6.9083^2\right)}} = -\ 0.5436$$

∴ There is a moderate negative correlation between the age of a used car and its selling price

$$r^2 = 0.2956 = 29.56\%$$

∴ 29.56% of the variation in the price of a used car can be explained by the age of the car.

Problem 2

At a business case competition, two judges were tasked with scoring (out of 1000) the case reports submitted by 12 teams. Using the following data, we would like to examine the association between the scores of the two judges.

Judge 1	650	760	740	700	590	620	700	690	900	500	610	700
Judge 2	900	720	690	850	920	800	890	920	1000	690	700	760

a. Compute the Spearman rank correlation between the scores of the two judges.

Judge 1 (X)	Judge 2 (Y)	Rank X	Rank Y	d = RX - RY	d^2
650	900	8	4	4	16
760	720	2	9	-7	49
740	690	3	11.5	-8.5	72.25
700	850	5	6	-1	1
590	920	11	2.5	8.5	72.25
620	800	9	7	2	4
700	890	5	5	0	0
690	920	7	2.5	4.5	20.25
900	1000	1	1	0	0
500	690	12	11.5	0.5	0.25
610	700	10	10	0	0
700	760	5	8	-3	9
				Sum:	244

$$r_s = 1 - \frac{6\sum_{i-1}^{12} d_i^2}{n\left(n^2 - 1\right)} = 1 - \frac{6 \times 244}{12 \times 143} = 0.1469$$

b. Comment on the association between the scores of the two judges.

The scores of the two judges have a very weak positive correlation.

Problem 3

An electric company is studying the relationship between the energy consumption (in thousand kilowatt-hours) and the number of rooms in a private single-family residence. A random sample of 10 homes yielded the following.

Number of Rooms	12	9	14	6	10	8	10	10	5	7
Energy Consumption (thousand kWh)	9	7	10	5	8	6	8	10	4	7

a. Determine the regression equation of energy consumption on the number of rooms.

	Number of Rooms	Energy Consumption (thousand kWh)	X^2	Y^2	XY
	12	9	144	81	108
	9	7	81	49	63
	14	10	196	100	140
	6	5	36	25	30
	10	8	100	64	80
	8	6	64	36	48
	10	8	100	64	80
	10	10	100	100	100
	5	4	25	16	20
	7	7	49	49	49
Sum	91	74	895	584	718

$$b_1 = \frac{n\left(\sum_{i=1}^{10} x_i y_i\right) - \left(\sum_{i=1}^{10} x_i\right)\left(\sum_{i=1}^{10} y_i\right)}{n\left(\sum_{i=1}^{10} x_i^2\right) - \left(\sum_{i=1}^{10} x_i\right)^2} = \frac{(10*718)-(91*74)}{(10*895)-(91)^2} = 0.66667$$

$$b_0 = \overline{y} - b_1\overline{x} = 7.4-(0.66667*9.1) = 1.3333$$

$$\therefore \hat{y} = 1.333 + 0.667x$$

b. Interpret the model.

b_0= 1.333 means overall energy consumption will be 1.333 (thousand kWh), when the number of rooms is 0.

b_1= 0.667 means overall energy consumption will increase by 0.667 (thousand kWh) for increasing 1 room.

c. What is the predicted energy consumption, in thousand kWh, for a six-room house.

Six-room means x=6

$\hat{y} = 1.333 + (0.667 * 6)$= 5.335 (thousand kWh)

d. Comment on the goodness of fit of the model.

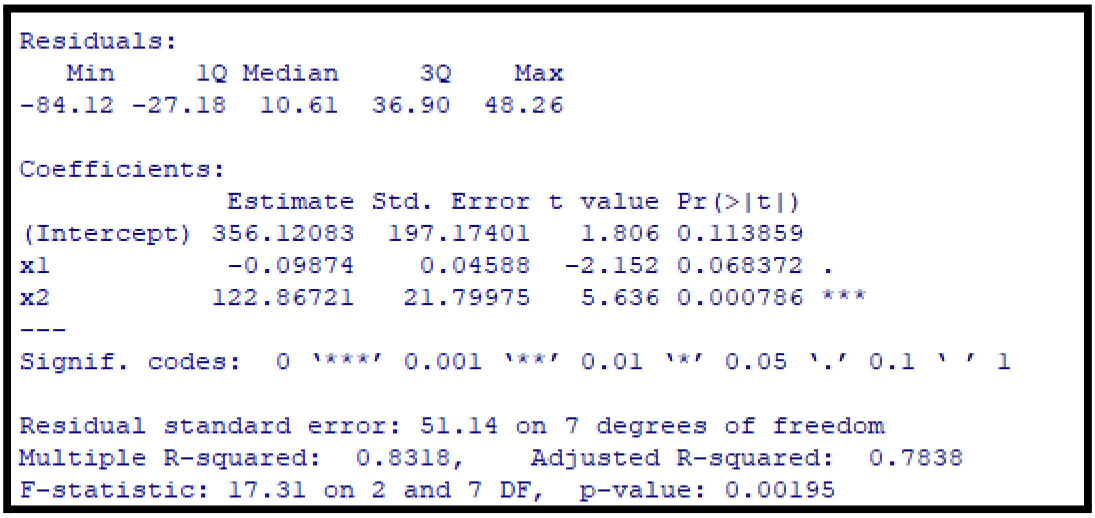
$SST = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$= 36.4; $SSE = \Sigma y_i^2 - b_0\ \Sigma y_i - b_1 \Sigma x_i y_i$ = 6.667

$r^2 = 1 - \frac{SSE}{SST}$ = 1 - $\frac{6.667}{36.4}$= 0.8168

r^2= 0.8168, so 81.68% of the variation in energy consumption can be explained by the number of rooms.

Problem 4

Designers of backpacks use exotic material to make packs that fit comfortably and distribute weight to eliminate pressure points. For fitting a regression model of price of backpack on the capacity (cubic inches) and comfort rating of backpacks, data for 10 backpacks are used. Comfort was measured using a rating from 1 to 5, with a rating of 1 denoting average comfort and a rating of 5 denoting excellent comfort. The output of the regression model is as follows on the next page:



```
Residuals:
   Min      1Q Median     3Q    Max
-84.12 -27.18  10.61  36.90  48.26

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 356.12083  197.17401   1.806 0.113859
x1           -0.09874    0.04588  -2.152 0.068372 .
x2          122.86721   21.79975   5.636 0.000786 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.14 on 7 degrees of freedom
Multiple R-squared:  0.8318,    Adjusted R-squared:  0.7838
F-statistic: 17.31 on 2 and 7 DF,  p-value: 0.00195
```

a. Determine the estimated regression equation that can be used to predict the price of a backpack given the capacity and the comfort rating.

   $\hat{y} = b_0 + b_1x_1 + b_2x_2 = 356.12083 - 0.09874x_1 + 122.86721x_2$

b. Interpret the model.

   $b_0 = 356.12083$means that the price of a backpack will be 356.12083, when both the capacity and comfort rating are zero.

   $b_1 = -0.09874$ means that the price of a backpack will decrease by 0.09874 units when the capacity will increase by one cubic inch keeping the comfort rating fixed.

   $b_2 = 122.86721$ means that the price of a backpack will increase by122.86721unit when the comfort rating will increase by one unit keeping the capacity fixed.

c. Predict the price for a backpack with a capacity of 4500 cubic inches and a comfort rating of 4.

   $\hat{y} = 356.12083 - (0.09874 * 4500) + (122.86721 * 4)$= 403.25967

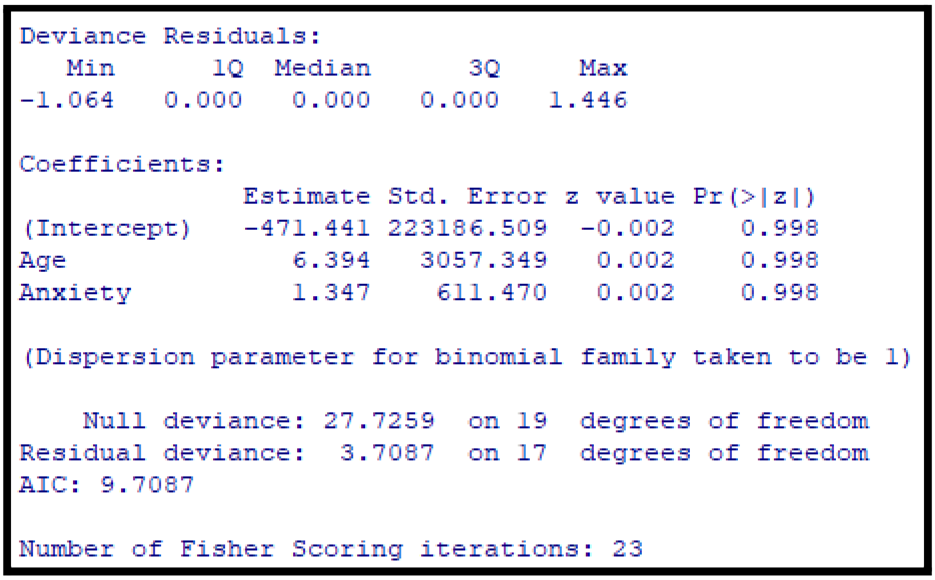
d. Comment on the goodness of fit of the model.

   Adjusted R-squared = 0.7838 = 78.38%

   ∴ 78.38% of the variation in the price of a backpack can be explained by taking into account the effects of capacity and comfort rating.

Problem 5

Suppose that we are working with some doctors on heart attack patients. The dependent variable is whether the patient has had a second heart attack within 1 year (yes=1). We have two independent variables, one is age of the patient and the other is a score on the anxiety scale (a higher score means more anxious). After applying logistic regression model, we have the following output:



```
Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.064    0.000    0.000    0.000    1.446

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -471.441 223186.509  -0.002    0.998
Age              6.394   3057.349   0.002    0.998
Anxiety          1.347    611.470   0.002    0.998

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.7259  on 19  degrees of freedom
Residual deviance:  3.7087  on 17  degrees of freedom
AIC: 9.7087

Number of Fisher Scoring iterations: 23
```

a. Determine the estimated logistic regression equation.

$$\hat{y} = \frac{e^{\left(\beta_0+\beta_1x_1+\beta_2x_2\right)}}{1+e^{\left(\beta_0+\beta_1x_1+\beta_2x_2\right)}} = \frac{e^{\left(-471.441+6.394x_i+1.347x_2\right)}}{1+e^{\left(-471.441+6.394x_i+1.347x_2\right)}}$$

b. Calculate the odds ratio and interpret.

$b_1 = 6.394$

Odds Ratio = $e^{6.394} \simeq 598.2448$

Interpretation:

The odds of having a second heart attack within 1 year is increased by 598.2448 for every unit increase in age of the patient

$b_2 = 1.347$

Odds Ratio = $e^{1.347} \simeq 3.8459$

Interpretation:

The odds of having a second heart attack within 1 year is increased by 3.8459 for every unit increase in the score of anxiety