STAT 3A03 Applied Regression With SAS

Fall 2017

Assignment 1 Solution Set

Q. 1 **a)** From class notes we have that the standard error of $\hat{\beta}_1$ is

$$\text{se}(\hat{\beta}_1) \;=\; \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \;=\; \sqrt{\frac{\text{SSE}}{(n-2)S_{xx}}}$$

Hence to complete this question we need to show that $\text{SSE} \;=\; S_{yy}\left(1-r^2\right)$.

$$\begin{aligned}
\text{SSE} &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \\
&= \sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)^2 \qquad \left(\text{since } \hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i\right) \\
&= \sum_{i=1}^{n}\left(y_i-\overline{y}+\hat{\beta}_1\overline{x}-\hat{\beta}_1 x_i\right)^2 \qquad \left(\text{since } \hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}\right) \\
&= \sum_{i=1}^{n}\left[(y_i-\overline{y})-\hat{\beta}_1\left(x_i-\overline{x}\right)\right]^2 \\
&= \sum_{i=1}^{n}(y_i-\overline{y})^2-2\hat{\beta}_1\sum_{i=1}^{n}\left(x_i-\overline{x}\right)(y_i-\overline{y})+\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\overline{x})^2 \\
&= S_{yy}-2\left(\frac{S_{xy}}{S_{xx}}\right)S_{xy}+\left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} \qquad \left(\text{since } \hat{\beta}_1 \;=\; \frac{S_{xy}}{S_{xx}}\right) \\
&= S_{yy}-\frac{S_{xy}^2}{S_{xx}} \\
&= S_{yy}\left(1-\frac{S_{xy}^2}{S_{xx}S_{yy}}\right) \\
&= S_{yy}\left(1-r^2\right)
\end{aligned}$$

[10 marks]

b) Using the result above and the definition of $\hat{\beta}_1$ we have

$$\begin{aligned}\frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} &= \frac{S_{xy}}{S_{xx}}\sqrt{\frac{(n-2)S_{xx}}{S_{yy}\,(1-r^2)}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}\sqrt{\dfrac{n-2}{1-r^2}}} \\ &= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\end{aligned}$$

[5 marks]

c) As suggested in the hint, let us first find $\overline{\hat{y}}$, $S_{y,\hat{y}}$ and $S_{\hat{y},\hat{y}}$

$$\begin{aligned}\overline{\hat{y}} &= \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= \hat{\beta}_0 + \hat{\beta}_1\overline{x} \\ &= (\overline{y} - \hat{\beta}_1\overline{x}) + \hat{\beta}_1\overline{x} \\ &= \overline{y}\end{aligned}$$

$$\begin{aligned}S_{y,\hat{y}} &= \sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}}) \\ &= \sum_{i=1}^{n}(y_i - \overline{y})(\hat{y}_i - \overline{y}) \\ &= \sum_{i=1}^{n}(y_i - \overline{y})(\hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{y}) \\ &= \sum_{i=1}^{n}(y_i - \overline{y})(\overline{y} - \hat{\beta}_1\overline{x} + \hat{\beta}_1 x_i - \overline{y}) \\ &= \sum_{i=1}^{n}(y_i - \overline{y})\hat{\beta}_1(x_i - \overline{x}) \\ &= \hat{\beta}_1 S_{xy}\end{aligned}$$

$$
\begin{aligned}
S_{\hat{y},\hat{y}} &= \sum_{i=1}^{n} (\hat{y}_i - \overline{\hat{y}})^2 \\
&= \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \\
&= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{y})^2 \\
&= \sum_{i=1}^{n} (\overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i - \overline{y})^2 \\
&= \sum_{i=1}^{n} \hat{\beta}_1^2 (x_i - \overline{x})^2 \\
&= \hat{\beta}_1^2 S_{xx}
\end{aligned}
$$

Now we simply need to use Theorem 3 with $\hat{y}$ in place of x everywhere to get

$$
\begin{aligned}
\hat{\alpha}_1 &= \frac{S_{y\hat{y}}}{S_{\hat{y},\hat{y}}} &&= \frac{\hat{\beta}_1 S_{xy}}{\hat{\beta}_1^2 S_{xx}} &&= 1 \\
\hat{\alpha}_0 &= \overline{y} - \hat{\alpha}_1 \overline{\hat{y}} &&= \overline{y} - 1 \times \overline{y} &&= 0
\end{aligned}
$$

[10 marks]

Q. 2 **a)** First we find the sum of squares

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 612 - \frac{1}{24} \times 108^2 = 126$$

$$S_{xx} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 10.47 - \frac{1}{24} \times 108 \times 1.84 = 2.19$$

[2 marks]

Hence we get the estimated slope

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{2.19}{126} = 0.01738$$

[2 marks]

and the estimated intercept

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{1.84}{24} - \frac{2.19}{126} \times \frac{108}{24} = -0.00155$$

[2 marks]

b)

$$\begin{aligned}
\text{SSE} &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \\
&= \sum_{i=1}^{n} \left(y_i - \left(\overline{y} - \hat{\beta}_1 \overline{x}\right) - \hat{\beta}_1 x_i\right)^2 \\
&= \sum_{i=1}^{n} \left((y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x})\right)^2 \\
&= \sum_{i=1}^{n} (y_i - \overline{y})^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) + \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
&= S_{yy} - 2\left(\frac{S_{xy}}{S_{xx}}\right) S_{xy} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} \\
&= S_{yy} - \frac{(S_{xy})^2}{S_{xx}}
\end{aligned}$$

[4 marks]

To calculate this we need

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 0.1839 - \frac{1}{24} \times 1.84^2 = 0.04283$$

[1 mark]

Hence an unbiased estimate of σ^2 is

$$\begin{aligned}\hat{\sigma}^2 = \frac{\text{SSE}}{n-2} &= \frac{0.04283 - \frac{2.19^2}{126}}{24-2} \\ &= \frac{0.04283 - 0.03806}{22} \\ &= \frac{0.00477}{22} \\ &= 0.00022\end{aligned}$$

[2 marks]

c) The standard errors are:

$$\begin{aligned}\text{se}(\hat{\beta}_0) &= \sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\hat{\sigma}^2} = \sqrt{\left(\frac{1}{24} + \frac{(108/26)^2}{126}\right) \times 0.00022} \\ &= \sqrt{0.00004387} \\ &= 0.006624\end{aligned}$$

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{0.00022}{126}} = 0.001312$$

[6 marks]

d) The correct critical value to use is $t_{(22;0.025)} = 2.074$. Since this is not contained in the table in your textbook, you should use $t_{(20;0.025)} = 2.09$ but either is acceptable.
For the intercept we have

$$\hat{\beta}_0 \pm t_{(20;0.025)}\text{se}(\hat{\beta}_0) = -0.00155 \pm 2.09 \times 0.006624 = (-0.0154,\ 0.0123).$$

[2 marks]

For the slope we have

$$\hat{\beta}_1 \pm t_{(20;0.025)}\text{se}(\hat{\beta}_1) = 0.01738 \pm 2.09 \times 0.001312 = (0.0146,\ 0.0201).$$

[2 marks]

e) We would conclude that each beer increases your blood alcohol count by about 0.017 units. Plausible values of the increase are between about 0.015 and 0.020 units. The estimated intercept is small and negative but we note that 0 is a plausible value of the true intercept since it is included in the confidence interval. **[2 marks]**

Q. 3 **a)** The fitted regression line is

$$\text{Late} \;=\; 28.6358 + 0.20349\ \text{Early}$$

[3 marks]

b) The mean square for error gives the unbiased estimator of σ^2 and so we have $\hat{\sigma}^2 \;=\; 188.119$.
[3 marks]

c) In the case we have $n = 93$ so we should use $t_{(91;0.025)} = 1.986$ but since this value is rarely in a table (and not in the one in your text) I will use $t_{(60;0.025)} = 2.000$ instead.
For the intercept we have

$$\hat{\beta}_0 \pm t_{(60;0.025)}\text{se}(\hat{\beta}_0) \;=\; 28.6358 \pm 2.000 \times 2.6149 \;=\; (23.406,\ 33.866).$$

[2 marks]

For the slope we have

$$\hat{\beta}_1 \pm t_{(60;0.025)}\text{se}(\hat{\beta}_1) \;=\; 0.2035 \pm 2.000 \times 0.1311 \;=\; (-0.0586,\ 0.4656).$$

[2 marks]

d) The proportion of the variability in the amount of late snow which can be explained by the amount of early snow is given by the coefficient of determination, R^2, which is 0.0258 or 2.58%.
[5 marks]

e) The correlation coefficient between the two variables is $r = \sqrt{0.0258} = 0.16$. Note that it is positive since the sign of r must agree with the sign of $\hat{\beta}_1$. **[5 marks]**

f) It does not appear the early season snow is a reliable predictor of the amount of late season snow. Even though there is a trend that heavier early season snow means heavier late season snow (positive $\hat{\beta}_1$ and r), this effect is not statistically significant. The p-value for testing if $\beta_1 = 0$ is 0.1239 meaning that there is no evidence in the data that the true slope is different from 0. Also early snow only explains 2.58% of the variability in late season snow leaving 97.42% of the variability unexplained. With so little explained variability we would be unwise to use early snowfall as a reliable predictor of late snowfall. **[5 marks]**

Q. 4 First we need to read the data into SAS. Here is the code I used.

```
Libname S3A3 "D:\STAT 3A03\Fall 2017\Data";

PROC IMPORT Out=S3A3.Anscombe
Datafile="D:\STAT 3A03\Fall 2017\Data\Anscombe.txt"
DBMS=DLM REPLACE;
Getnames=YES;
Datarow=2;
run;
```

a) **(i)** The correlation coefficient between $Y1$ and $X1$ is $r = 0.81642$.
The SAS code that I used is

```
PROC CORR Data=S3A3.Anscombe;
VAR Y1;
With X1;
run;
```

and the output produced is

The SAS System

The CORR Procedure

1 With Variables:	X1
1 Variables:	Y1

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
X1	11	9.00000	3.31662	99.00000	4.00000	14.00000
Y1	11	7.50091	2.03157	82.51000	4.26000	10.84000

Pearson Correlation Coefficients, N = 11

Prob > |r| under H0: Rho=0

	Y1
X1	0.81642 0.0022

[1 mark]

(ii) The correlation coefficient between $Y2$ and $X2$ is $r = 0.81624$.
The same SAS code can be used with the appropriate changes of variable names. The output is

The SAS System

The CORR Procedure

1 With Variables:	X2
1 Variables:	Y2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
X2	11	9.00000	3.31662	99.00000	4.00000	14.00000
Y2	11	7.50091	2.03166	82.51000	3.10000	9.26000

Pearson Correlation Coefficients, N = 11
Prob > |r| under H0: Rho=0

	Y2
X2	0.81624 0.0022

[1 mark]

(iii) The correlation coefficient between $Y3$ and $X3$ is $r = 0.81629$.

The SAS System

The CORR Procedure

1 With Variables:	X3
1 Variables:	Y3

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
X3	11	9.00000	3.31662	99.00000	4.00000	14.00000
Y3	11	7.50000	2.03042	82.50000	5.39000	12.74000

Pearson Correlation Coefficients, N = 11
Prob > |r| under H0: Rho=0

	Y3
X3	0.81629 0.0022

[1 mark]

(iv) The correlation coefficient between $Y4$ and $X4$ is $r = 0.81652$.

The SAS System

The CORR Procedure

1 With Variables:	X4
1 Variables:	Y4

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
X4	11	9.00000	3.31662	99.00000	8.00000	19.00000
Y4	11	7.50091	2.03058	82.51000	5.25000	12.50000

Pearson Correlation Coefficients, N = 11

Prob > |r| under H0: Rho=0

	Y4
X4	0.81652 0.0022

[1 mark]

b) The following code will do the all of the regressions as well as constructing the plots which I will show in part (c).

```
PROC REG Data=S3A3.Anscombe;
Model Y1=X1;
Plot Y1*X1;
Model Y2=X2;
Plot Y2*X2;
Model Y3=X3;
Plot Y3*X3;
Model Y4=X4;
Plot Y4*X4;
run;
quit;
```

The SAS output for the four regressions is on the following pages. We note that all of these outputs are essentially identical except for rounding.

(i)

The REG Procedure
Model: MODEL1
Dependent Variable: Y1

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27.51000	27.51000	17.99	0.0022
Error	9	13.76269	1.52919		
Corrected Total	10	41.27269			

Root MSE	1.23660	R-Square	0.6665
Dependent Mean	7.50091	Adj R-Sq	0.6295
Coeff Var	16.48605		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.00009	1.12475	2.67	0.0257
X1	1	0.50009	0.11791	4.24	0.0022

[2 marks]

(ii)

The REG Procedure
Model: MODEL2
Dependent Variable: Y2

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27.50000	27.50000	17.97	0.0022
Error	9	13.77629	1.53070		
Corrected Total	10	41.27629			

Root MSE	1.23721	R-Square	0.6662
Dependent Mean	7.50091	Adj R-Sq	0.6292
Coeff Var	16.49419		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.00091	1.12530	2.67	0.0258
X2	1	0.50000	0.11796	4.24	0.0022

[2 marks]

(iii)

The REG Procedure
Model: MODEL3
Dependent Variable: Y3

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27.47001	27.47001	17.97	0.0022
Error	9	13.75619	1.52847		
Corrected Total	10	41.22620			

Root MSE	1.23631	R-Square	0.6663
Dependent Mean	7.50000	Adj R-Sq	0.6292
Coeff Var	16.48415		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.00245	1.12448	2.67	0.0256
X3	1	0.49973	0.11788	4.24	0.0022

[2 marks]

(iv)

The REG Procedure
Model: MODEL4
Dependent Variable: Y4

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27.49000	27.49000	18.00	0.0022
Error	9	13.74249	1.52694		
Corrected Total	10	41.23249			

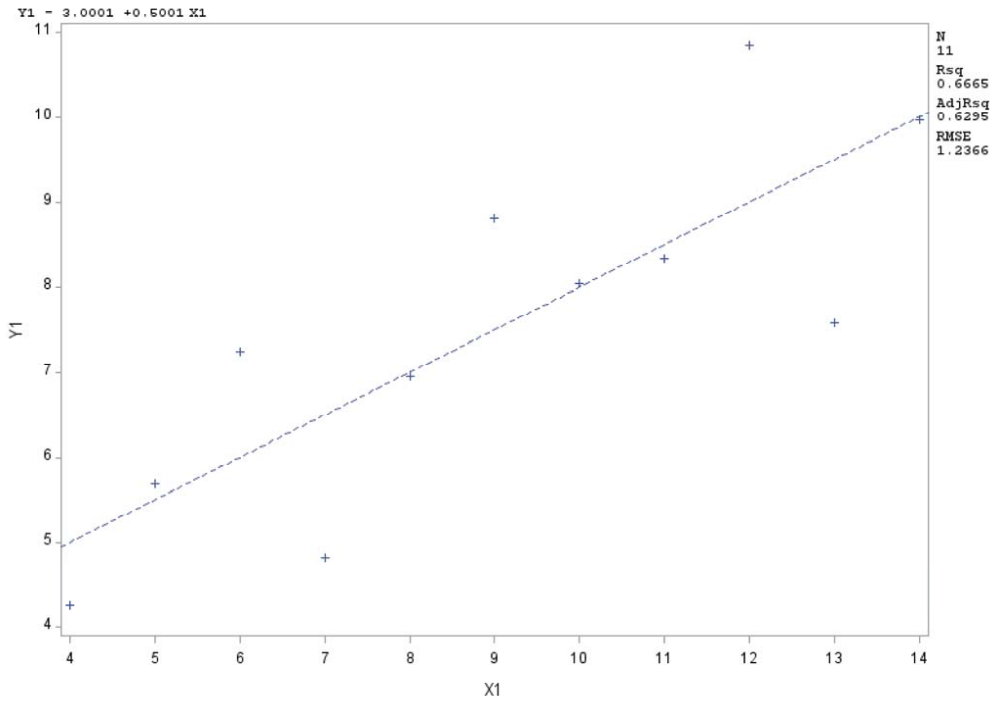
Root MSE	1.23570	R-Square	0.6667
Dependent Mean	7.50091	Adj R-Sq	0.6297
Coeff Var	16.47394		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.00173	1.12392	2.67	0.0256
X4	1	0.49991	0.11782	4.24	0.0022

[2 marks]

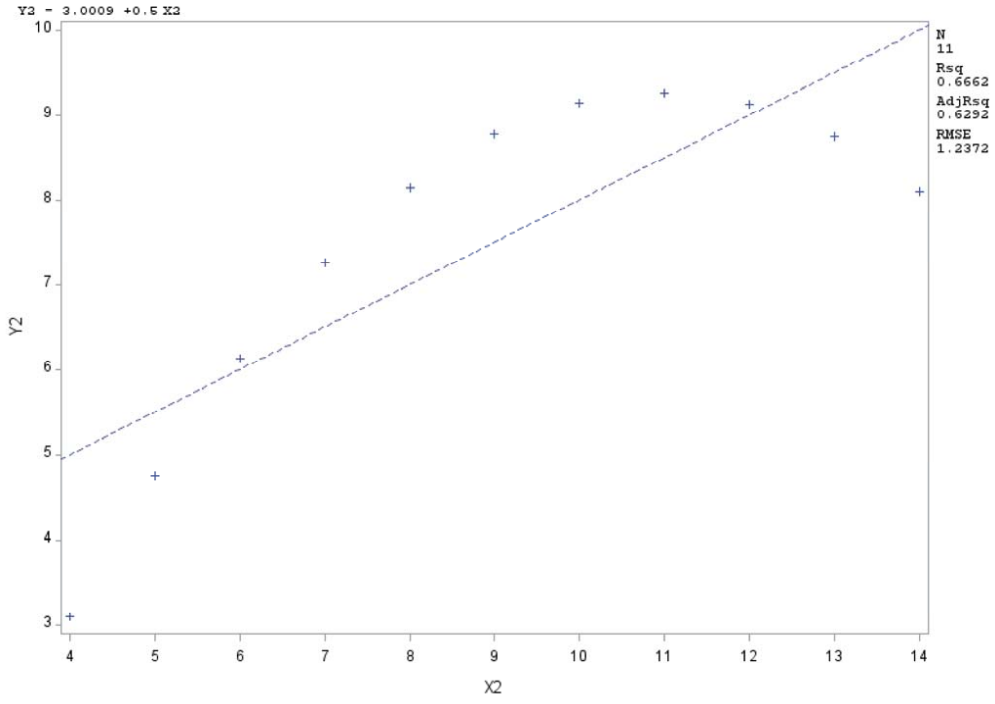
c) The four plots with fitted lines are as follows

(i)



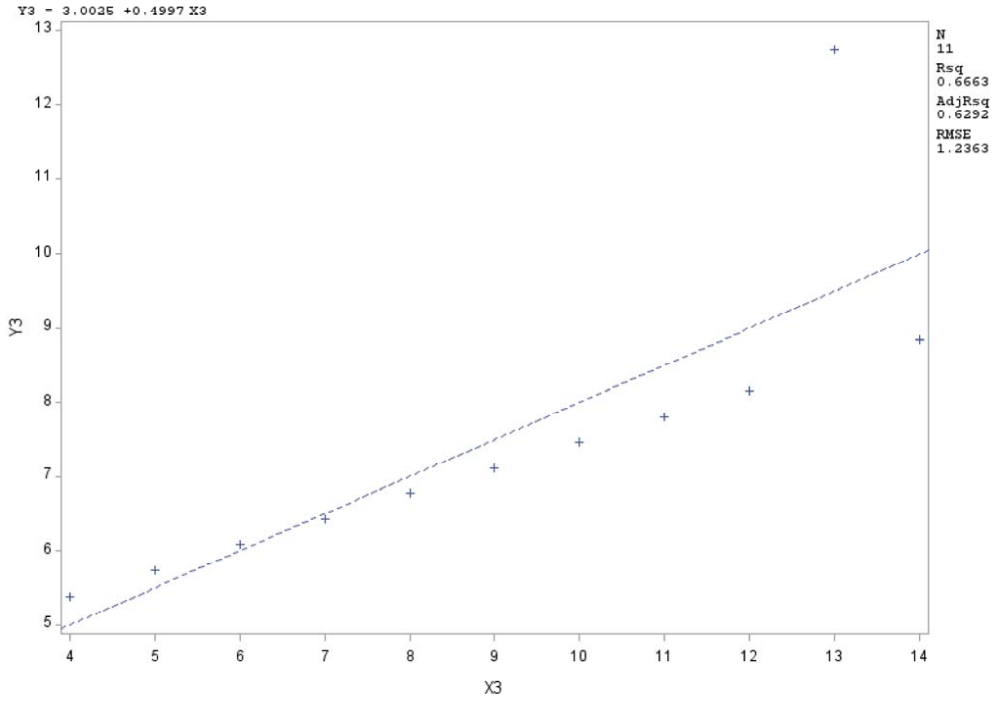
[2 marks]

(ii)



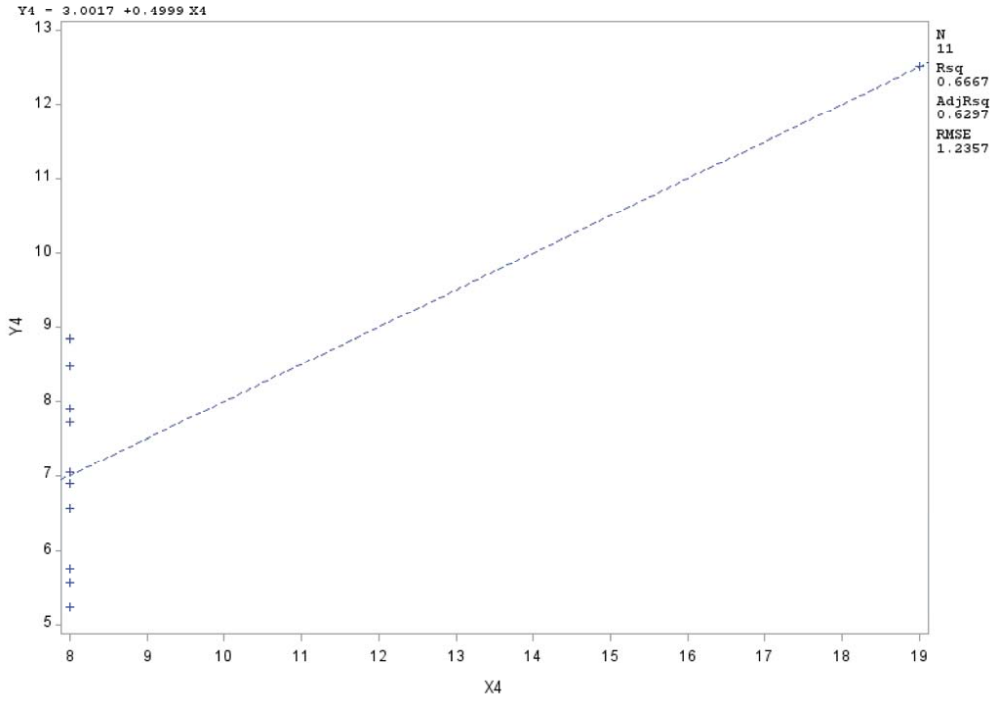
[2 marks]

(iii)



[2 marks]

(iv)



[2 marks]

d) The pair Y1 & X1 seem to show a linear relationship and so linear regression is appropriate here.

For the pair Y2 & X2, the relationship seems quite curved so linear regression would not be appropriate on the raw scale. Possibly a transformation of X2 could be used to make the relationship more linear or a quadratic term (the square of X2) could be included with the linear term.

Most of the points for the pair Y3 & X3 lie very close to a straight line but that is not the fitted line. The fitted line is very heavily influenced by one single point. This point may be an outlier and so we should investigate why that point is different from the others and consider whether to include the point in our analysis or not.

The pair Y4 & X4 have only two x values and all but one of the y values occurs at the same x value. We cannot really say if linear regression is appropriate here or not since there is insufficient data about a range of x values.

What is interesting about this dataset is that the correlation and fitted lines are almost exactly the same for all four pairs but the characteristics are very different. The point Anscombe was trying to make was that one cannot rely only on the numerical output, plots are absolutely essential in a proper regression analysis. **[5 marks]**