UNIVERSITY OF STIRLING
Computing Science & Mathematics

MATU9D2 : PRACTICAL STATISTICS

EXAMINATION : SPRING SEMESTER 2016

MONDAY 2 MAY 2016 : 0900 - 1200 HOURS

CREDIT WILL BE GIVEN FOR AT MOST FOUR QUESTIONS.

NON-PROGRAMMABLE CALCULATORS MAY BE USED, BUT FULL WORKING SHOULD BE GIVEN.

THE NUMBER OF MARKS ALLOCATED TO EACH PART OF A QUESTION IS SHOWN IN BRACKETS.

STATISTICAL TABLES, A FORMULA SHEET AND GRAPH PAPER ARE PROVIDED.

: PAST EXAM PAPER : SPRING 2017

1. A company is exploring ways of improving productivity. It is trialling a new ergonomic keyboard amongst administrative staff. A random sample of 15 administrative staff have had the number of words per minute (wpm) they could achieve with the current keyboard and with the new keyboard recorded. The data is presented below.

Person,i	Current(x_i)	New(y_i)
1	25.50	43.60
2	59.20	69.90
3	38.40	39.80
4	66.80	73.40
5	44.90	50.20
6	47.40	53.90
7	41.60	40.30
8	48.90	58.00
9	60.70	66.90
10	41.00	66.50
11	36.10	27.40
12	34.40	33.70
13	39.50	40.50
14	41.90	44.80
15	61.60	65.60

Let $z_i = x_i - y_i$ be the difference for the ith person (i =1,2,….15) in the wpm x_i using the current keyboard and the wpm y_i using the new keyboard.

(i) Calculate z_i (i = 1,2,..... 15) and obtain the five number summary. Construct a box and whisker plot of the differences and check for the presence of outliers. **[12]**

(ii) Name a formal test that could be used to investigate whether there is a change in the number of words per minute achieved. What are the assumptions? **[3]**

(iii) Calculate a 95% confidence interval for the average difference in the number of words per minute. Is there a significant change? Justify your answer. **[5, 2]**

(iv) Calculate a 95% confidence interval for the standard deviation of the number of words per minute with the new keyboard.

Based on a very large study, the standard deviation of the number of words per minute in a company that are already using this keyboard is 22 wpm. Test, at the 5% significance level, whether the standard deviation in this group differs significantly from 22 wpm. **[6, 2]**

(Question 2 overleaf)

2. A poll was conducted to determine the relationship between gender and attitudes towards the EU in the anticipated referendum. The results are given in the following table:

	Stay	Leave	Undecided
Male	198	326	410
Female	232	273	425

(i) By calculating appropriate percentages, describe the observed relationship (if any) between gender and the polled opinion. **[6]**

(ii) By performing an appropriate test, test the assumption that the way a person will vote in the referendum is associated with their gender. **[10]**

(iii) Calculate the 95% confidence interval estimate for the proportion of people that wish to remain in the EU. **[4]**

(iv) Assess formally the assumption that the proportions of those undecided is independent of gender by comparing the proportions of the undecided at the 5% significance level. **[10]**

3. The manager of a large electronics company is investigating any possible differences in dexterity scores of workers on the day shift and the night shift. The dexterity scores of a random sample of workers from each shift are given below.

Day Shift Score (x)	80.3	71.1	67.9	74.7	75.9	67.5
	62.5	70.9	70.9	62.3	76.9	70.9
	71.4	72.4	71.9	67.9	62.3	72.3
Night Shift Score (y)	87.9	94.2	78.1	78.2	78.8	67.8
	85.4	90.5	68.8	70.8	77.0	86.9
	94.2	76.0	90.7	81.2	76.1	80.0
	82.4	68.8	69.7			

Summary Statistics:

$$\sum x_i = 1270.00 \qquad \sum y_i = 1683.50$$
$$\sum x_i^2 = 90020.76 \qquad \sum y_i^2 = 136373.35$$

(i) Construct a back-to-back stem and leaf plot of this data. Describe the distributions of the two sets of data. **[6, 2]**

(Question 3 continued overleaf…)

(ii) Verify that the sample means and sample standard deviations for dexterity scores on both shifts are as shown in the Tables overleaf and comment (briefly) on the results. **[6]**

(iii) Carry out a formal test of the equality of the population variance, using the 1% significance level. State clearly your conclusion. **[6]**

(iv) (a) Provide full details (including hypotheses, test statistic and assumptions) of a formal test to investigate if whether the mean dexterity scores are different on the day and night shifts. For a significance level of 1%, use the given Minitab output to perform this test. **State which Table you are interpreting** and which **Figures you are using** to validate the assumptions.

(b) Calculate a 95% confidence interval for the mean difference in the dexterity scores. State clearly your conclusion for the test and interpret the confidence interval. Further, compare your conclusions with (iv) (a). **[10]**

Figure 1 : Normal Probability Plot - Day Shift Dexterity Scores

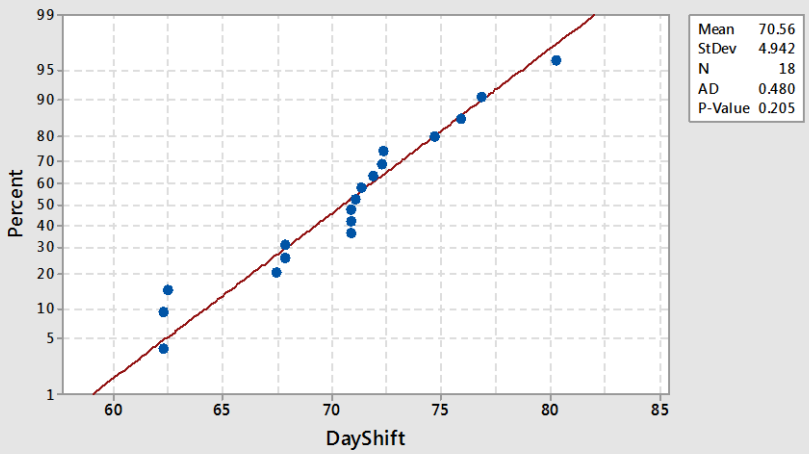
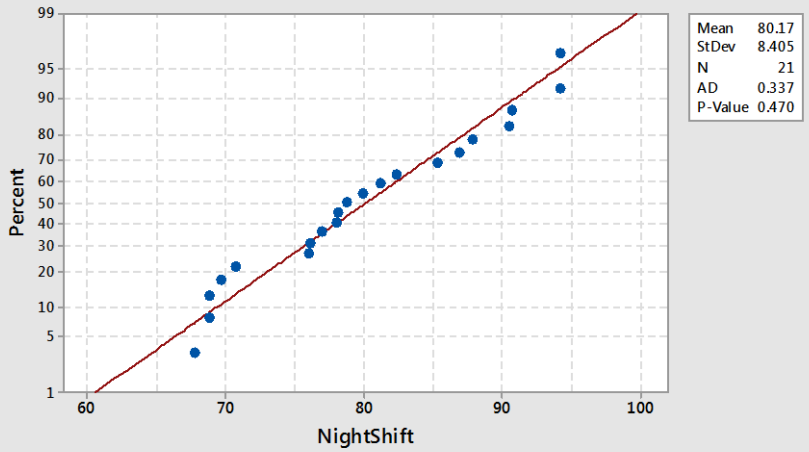


Figure 2 : Normal Probability Plot - Night Shift Dexterity Scores



(Question 3 continued overleaf…)

Table 1. Two-Sample T-Test and CI: DayShift, NightShift

```
Two-sample T for DayShift vs NightShift

             N   Mean  StDev  SE Mean
DayShift    18  70.56   4.94      1.2
NightShift  21  80.17   8.40      1.8

T-Test of difference = 0 (vs ≠): T-Value = -4.42
P-Value = 0.000  DF = 33
```

Table 2. Two-Sample T-Test and CI: DayShift, NightShift

```
Two-sample T for DayShift vs NightShift

             N   Mean  StDev  SE Mean
DayShift    18  70.56   4.94      1.2
NightShift  21  80.17   8.40      1.8

T-Test of difference = 0 (vs ≠): T-Value = -4.26
P-Value = 0.000  DF = 37
Both use Pooled StDev = 7.0288
```

Table 3. Two-Sample T-Test and CI: DayShift, NightShift

```
Two-sample T for DayShift vs NightShift

             N   Mean  StDev  SE Mean
DayShift    18  70.56   4.94      1.2
NightShift  21  80.17   8.40      1.8

T-Test of difference = 0 (vs <): T-Value = -4.42
P-Value = 0.000  DF = 33
```

Table 4. Two-Sample T-Test and CI: DayShift, NightShift

```
Two-sample T for DayShift vs NightShift

             N   Mean  StDev  SE Mean
DayShift    18  70.56   4.94      1.2
NightShift  21  80.17   8.40      1.8


T-Test of difference = 0 (vs <): T-Value = -4.26
P-Value = 0.000  DF = 37
```

(Question 4 overleaf)

4. The mean levels of CO_2 levels in a region are shown below (in some arbitrary units) for a range of years.

Year	Quarter	CO_2 Level	Trend
1	1	1.05	
	2	1.15	
	3	1.60	2.106
	4	4.00	2.456
2	1	2.30	**A**
	2	2.70	3.691
	3	4.59	4.275
	4	6.35	4.665
3	1	4.62	4.920
	2	3.50	5.400
	3	5.83	5.939
	4	8.95	6.215
4	1	6.33	**B**
	2	4.00	7.279
	3	8.16	
	4	12.30	

(i) Working with 3 decimal places use the 4 point centred moving average to calculate the trend for the CO_2 levels of the first quarters of year 2 and year 4 (i.e. those cells marked **A** and **B** in the table). **[4]**

(ii) Plot this time series and the trend on the graph paper supplied and describe (briefly) the time series graph. **[5]**

(iii) Estimate the CO_2 levels for the first quarter of year 5 using the trend. **[5]**

(iv) Calculate the quarterly adjusted seasonal indices during this time period, stating (with justification) the model you are using. **[6]**

(v) Hence calculate the seasonal adjusted values. Plot the seasonal adjusted values on the plot in (ii) and comment on the change in the CO_2 level calculated (if any). **[5,3,2]**

(Question 5 overleaf)

5. (i) (a) Briefly discuss how you would formally and informally decide whether data is normally distributed.

(b) Give details of four different types of data; giving examples of each.

[5, 5]

(ii) A car manufacturer has collected data on fuel consumption on its bestselling family car. A random sample of 20 drivers were randomly allocated to drive around a standard route using one of four strategies. Data for the fuel consumption (in 0.01 gallons) for the drivers for each driving strategy (y_{ij} ; i = 1, ... 4 ; j = 1,...5) where i identifies the driving strategy and j the observation within the group, is presented below.

	Driving Strategy			
	1	2	3	4
	21	34	25	38
	23	29	20	32
	28	33	26	37
	25	28	18	24
	19	26	23	26
Mean	23.20	30.00	22.40	31.40

$$\sum\sum y_{ij} = 535 \qquad \sum\sum y_{ij}^2 = 14929$$

(a) Using the above summary statistics and the plot given in Figure 3, describe the sample results. **[3]**

(b) What formal test could be used to compare the mean fuel consumption under the four different driving strategies? What are the assumptions of this test?

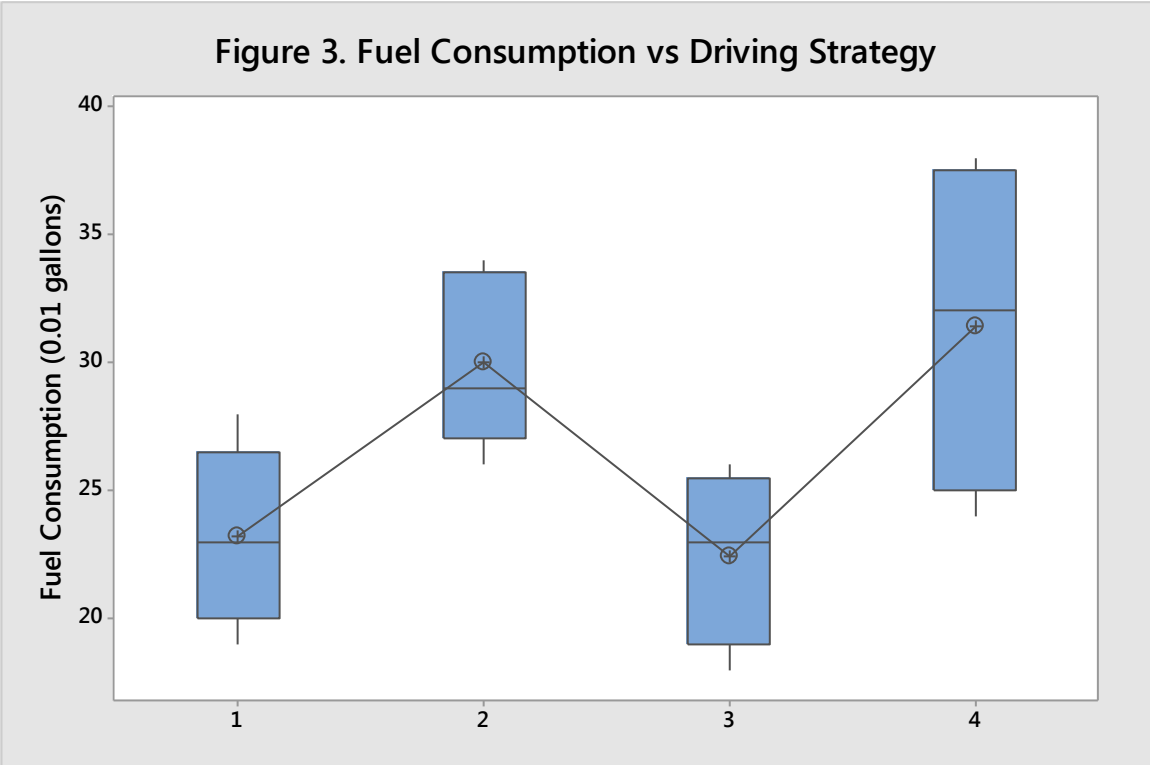
[4]

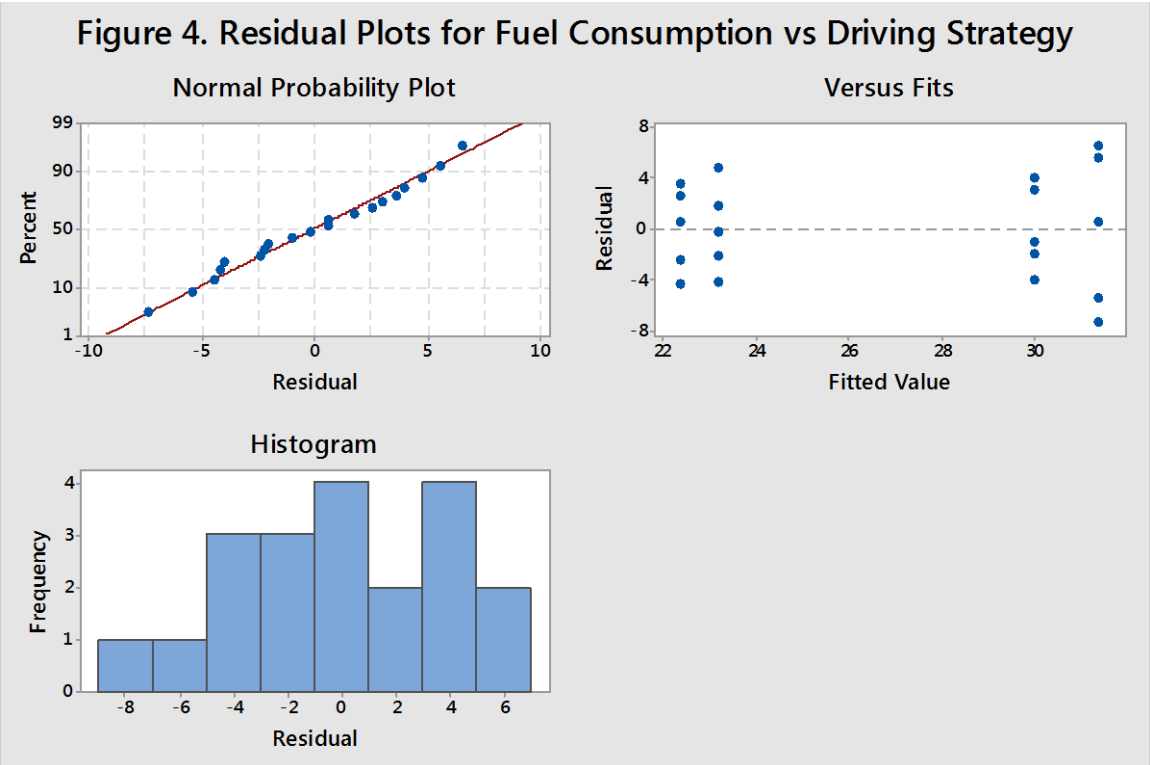
(c) Perform the formal test that you gave in (b). Clearly state your conclusions.

[8]

(d) Discuss whether there are any differences in the mean fuel consumption in the light of (a) and (c) and using the Minitab output (Figures 3 and 4 and Table 5) overleaf. **[5]**

(Question 5 continued overleaf…)





(Question 5 continued overleaf…)

Table 5. Minitab Output

Means

Factor	N	Mean	StDev	95% CI
1	5	23.20	3.49	(19.10, 27.30)
2	5	30.00	3.39	(25.90, 34.10)
3	5	22.40	3.36	(18.30, 26.50)
4	5	31.40	6.31	(27.30, 35.50)

Pooled StDev = 4.32435

Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

Factor	N	Mean	Grouping
4	5	31.40	A
2	5	30.00	A B
1	5	23.20	B
3	5	22.40	B

Means that do not share a letter are significantly different.

Tukey Simultaneous Tests for Differences of Means

Difference of Levels	Difference of Means	SE of Difference	95% CI	T-Value	Adjusted P-Value
2 - 1	6.80	2.73	(-1.03, 14.63)	2.49	0.101
3 - 1	-0.80	2.73	(-8.63, 7.03)	-0.29	0.991
4 - 1	8.20	2.73	(0.37, 16.03)	3.00	0.038
3 - 2	-7.60	2.73	(-15.43, 0.23)	-2.78	0.058
4 - 2	1.40	2.73	(-6.43, 9.23)	0.51	0.955
4 - 3	9.00	2.73	(1.17, 16.83)	3.29	0.022

Individual confidence level = 98.87%

(Question 6 overleaf)

6. The management of a Fast Food chain is examining the profitability of its outlets so has collected data from a random sample of outlets. The net profit and the drive-through sales (in £millions) for 10 outlets are presented below:

Outlet (i)	1	2	3	4	5	6	7	8	9	10
Net Profit (y_i)	1.5	0.8	1.2	1.4	0.2	0.8	0.6	0.9	0.4	0.6
Drive-Through Sales (x_i)	7.7	4.5	8.4	7.8	2.4	4.8	2.5	3.4	2.0	4.1

Summary Statistics :

$$\sum x_i = 47.60 \qquad \sum y_i = 8.40$$

$$\sum x_i^2 = 278.36 \qquad \sum y_i^2 = 8.66$$

$$\sum x_i y_i = 48.29$$

Use the Minitab output (Figures 6 and 7 and Table 6) on the next two pages, where appropriate, to answer the questions below.

(i) Calculate the correlation coefficient between the net profit and the drive-through sales. Give full details of how you could use the Minitab output to check your answer. **[4, 2]**

(ii) Write down a simple linear regression model for y in terms of x, stating all assumptions and explaining all parameters. From the Minitab output, discuss whether the assumptions are valid for this data and provide estimates of all the parameters in the model including, in particular, an estimate of the error variance σ^2. **[3, 3, 3]**

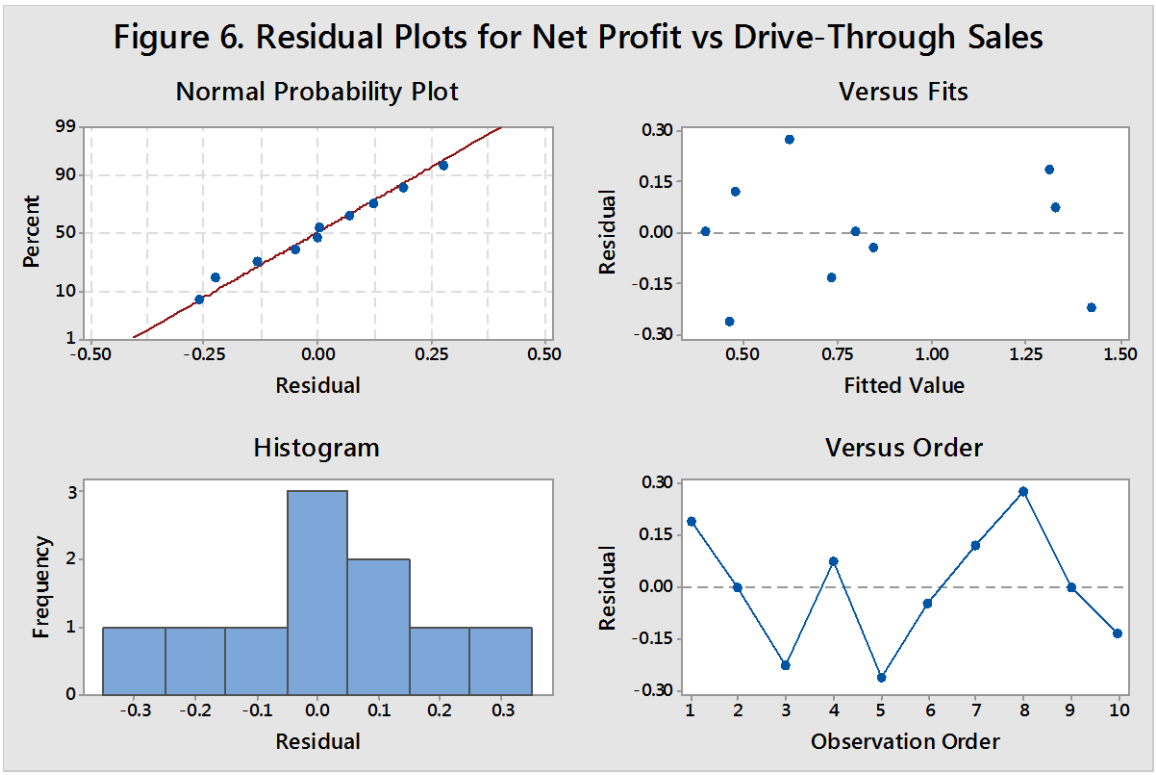
(iii) Using the Minitab output, carry out an appropriate test to investigate whether there is a linear relationship. Give full details of the hypothesis test and calculate a 95% confidence interval that could also be used to test the hypothesis. **[5]**

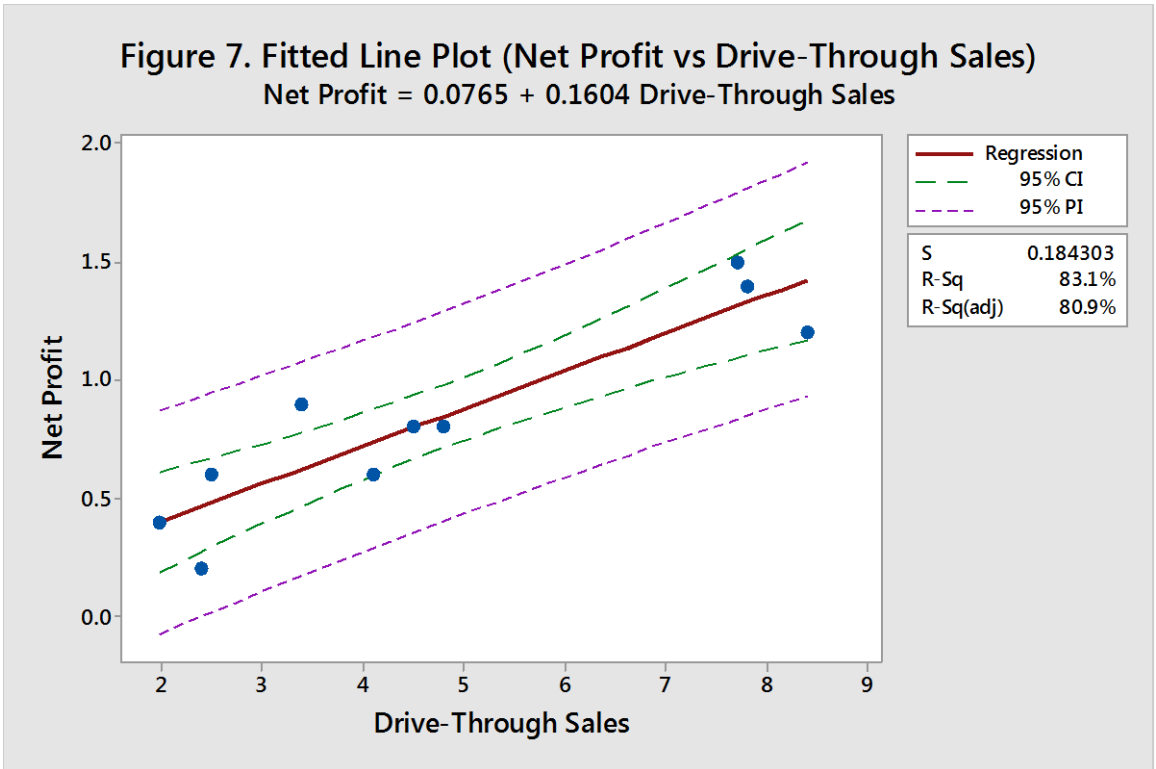
(iv) Calculate a 95% confidence interval for the mean net profit for outlets with drive-through sales of £6,000,000. **[3]**

(v) Calculate a 95% prediction interval for the net profit for an outlet with drive-through sales of £6,000,000. **[3]**

(vi) Discuss how good the model is in relation to your answers in (i), (iii), (iv) and (v) and the Minitab output. **[4]**

(Question 6 continued overleaf…)





(Question 6 continued overleaf…)

Table 6. Regression Analysis: Net Profit versus Drive-Through Sales

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1.3323	1.33226	39.22	0.000
Drive-Through Sales	1	1.3323	1.33226	39.22	0.000
Error	8	0.2717	0.03397		
Total	9	1.6040			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.184303	83.06%	80.94%	71.45%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.077	0.135	0.57	0.587	
Drive-Through Sales	0.1604	0.0256	6.26	0.000	1.00

Regression Equation

Net Profit = 0.077 + 0.1604 Drive-Through Sales

[END OF EXAMINATION]