Assessed Coursework 2

This coursework is **assessed** and is worth 10% of the module mark. You should submit your solution using the coursework collection box on A Floor of the Mathematical Sciences Building by 5pm on Tuesday 3rd May 2016.

1. The table below shows the number of hospital admissions related to drug misuse in each English region, together with the region's population (in millions).

Region	Admissions	Population
North East	246	2.6
North West	1392	7.1
Yorkshire	618	5.3
East Midlands	452	4.5
West Midlands	397	5.6
East of England	334	5.9
London	1432	8.2
South East	413	4.5
South Central	279	4.2
South West	414	5.3

   (a) Produce a scatterplot of hospital admissions versus population. Explain why a non-parametric hypothesis test is appropriate to investigate correlation between hospital admissions and population.

   (b) Conduct a non-parametric hypothesis test at the 5% significance level to assess whether there is a non-zero correlation between hospital admissions and population.

   [6 marks]

2. The test statistic for testing for non-zero linear correlation is

$$t_1 = r\sqrt{\frac{n-2}{1-r^2}}$$

   and the test statistic for testing for the existence of regression is

$$t_2 = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{(n-1)s_x^2}}.$$

   Use the formulae given in the notes for $\hat{\sigma}$, $\hat{\beta}_1$ and r to prove that $t_1 = t_2$. [5 marks]

3. The table below shows the number of deaths related to drug misuse in each year from 1997 to 2011.

Year	Deaths	Year	Deaths	Year	Deaths
1997	1312	2002	1613	2007	1727
1998	1458	2003	1432	2008	1939
1999	1628	2004	1497	2009	1876
2000	1604	2005	1608	2010	1784
2001	1808	2006	1560	2011	1605

Fit a least squares linear regression model that has the number of deaths as the response (Y) variable and the year as the predictor (X) variable.

(a) Estimate the model parameters.

(b) Calculate a 95% confidence interval for the gradient of the fitted regression line. Is there evidence for the existence of regression?

(c) Produce suitable residual plot(s) and comment on whether the residuals satisfy the assumptions of (i) the model and (ii) a test for the existence of regression.

(d) Use your fitted model to calculate a 95% prediction interval for the number of deaths in each year from 2012 to 2020.

(e) Produce a single scatterplot showing:

i. the data,

ii. the fitted regression line,

iii. the limits of the prediction intervals calculated in (d).

[15 marks]

4. In 2011 there were 1605 deaths related to drug misuse. Let π be the proportion of deaths where the subject is male. A researcher samples 40 cases at random and records that 30 are male and 10 are female.

(a) Test $H_0 : \pi = 0.5$ vs. $H_1 : \pi > 0.5$ at the 5% significance level.

(b) For the test in (a), calculate the probability of a type II error for $H_1 : \pi = \pi_1$ when (i) $\pi_1 = 0.6$, (ii) $\pi_1 = 0.7$ and (iii) $\pi_1 = 0.8$.

(c) Plot the power of the test in (a) against π for $0.4 \leq \pi \leq 0.9$.

(d) The researcher would like to know if the proportion of deaths where the subject is male is different for under 40 year olds and over 40 year olds. The number of deaths by gender and age group in the sample are given below.

	Under 40	Over 40
Male	16	14
Female	4	6

Perform a hypothesis test, at the 5% significance level, to answer the researcher's question.

[14 marks]