**Linear regression**

**QUESTION 9: 2009**

(a) List the key assumptions that are made when a statistician fits a linear regression model to a data set. (5 marks)

(b) Figure 1 below shows global mean temperatures (in ° C) for the period from 1955 to 2007. Figure 2 shows a graph of residuals versus fitted values from a linear regression of Temperature versus Year. Discuss whether these data violate any of the assumptions you listed in your answer to part (a). (5 marks)

(c) Discuss whether the data show evidence of the following problems in this data set: (i) non-linearity, (ii) non-constant error variance, (iii) outliers, and (iv) influential points. Discuss how you would deal with any problem you see relating to (i), (ii), (iii) and (iv). (5 marks)

(d) Some Australian journalists argue that global climate change stopped in the year 1998. The basis of their claim is that 1998 is the warmest year on record, and every year since then has been colder than 1998. Discuss whether you agree or disagree with their argument. Base your discussion on the data shown in Figures 1 and 2. (The point for 1998 is represented by the box symbol in Figures 1 and 2.) (5 marks)

**Figure 1: Global mean temperature for the period, 1955 – 2007**



**Figure 2: Residuals versus Fitted Values for the linear regression of Global Mean Temperature versus Year**

**QUESTION 10: 2009**

Figure 3 below shows Global Mean Temperature (in ° C) versus Year for the 25-year period from 1958 to 1982, and Figure 4 shows Global Mean Temperature from 1983 to 2007. The X-axis in both figures ranges from 1 to 25.

**Figure 3: Global mean temperature (°C) versus Year for the first period (1958–1982).**



**Figure 4: Global mean temperature (°C) versus Year for the second period (1983–2007).**



The Regression Analysis below shows results of a multiple linear regression of TEMPERATURE versus the 3 variables: YEAR, PERIOD and YEAR \* PERIOD.

The variable PERIOD has values of 0 for the first period (1958 – 1982) and 1 for the second period (1983 – 2007).

Use these results to answer the following questions:

1. Discuss whether Temperature increased during the first period (1958 – 1982).

(4 marks)

1. What were the annual rates of temperature increase in the first and second periods? According to the fitted lines, by how much did global mean temperature increase over the 50-year period from 1957 to 2007? (4 marks)
2. Discuss whether the fitted linear relationship between Temperature and Year for the second period is significantly different from the relationship for the first period. Discuss whether temperature increased more quickly in the second 25-year period than in the first period. What is the p-value for statistical significance of differences in the rates of increase in the two periods? (8 marks)
3. Based on the relationship fitted to data for the second period, in what year do you predict that global mean temperature will reach 15.6 ° C, which is 2 °C higher than in the mid 19th Century? Briefly discuss whether your estimate is likely to be an accurate prediction of what will happen in the real world. (4 marks)

**Regression Analysis**

The regression equation is

**TEMPERATURE = 13.9 + 0.00614 YEAR + 0.149 PERIOD + 0.0141 PERIOD \*YEAR**

**Predictor Coeff SE Coeff T P**

**Constant 13.8618 0.0418 331.63 0.000**

**YEAR 0.006138 0.002812 2.18 0.034**

**PERIOD 0.14935 0.05911 2.53 0.015**

**PERIOD \*YEAR 0.014096 0.003976 3.54 0.001**

S = 0.101378 R-Sq = 80.6% R-Sq(adj) = 79.3%

Analysis of Variance

Source DF SS MS F P

Regression 3 1.96404 0.65468 63.70 0.000

Residual Error 46 0.47276 0.01028

Total 49 2.43680

Source DF Seq SS

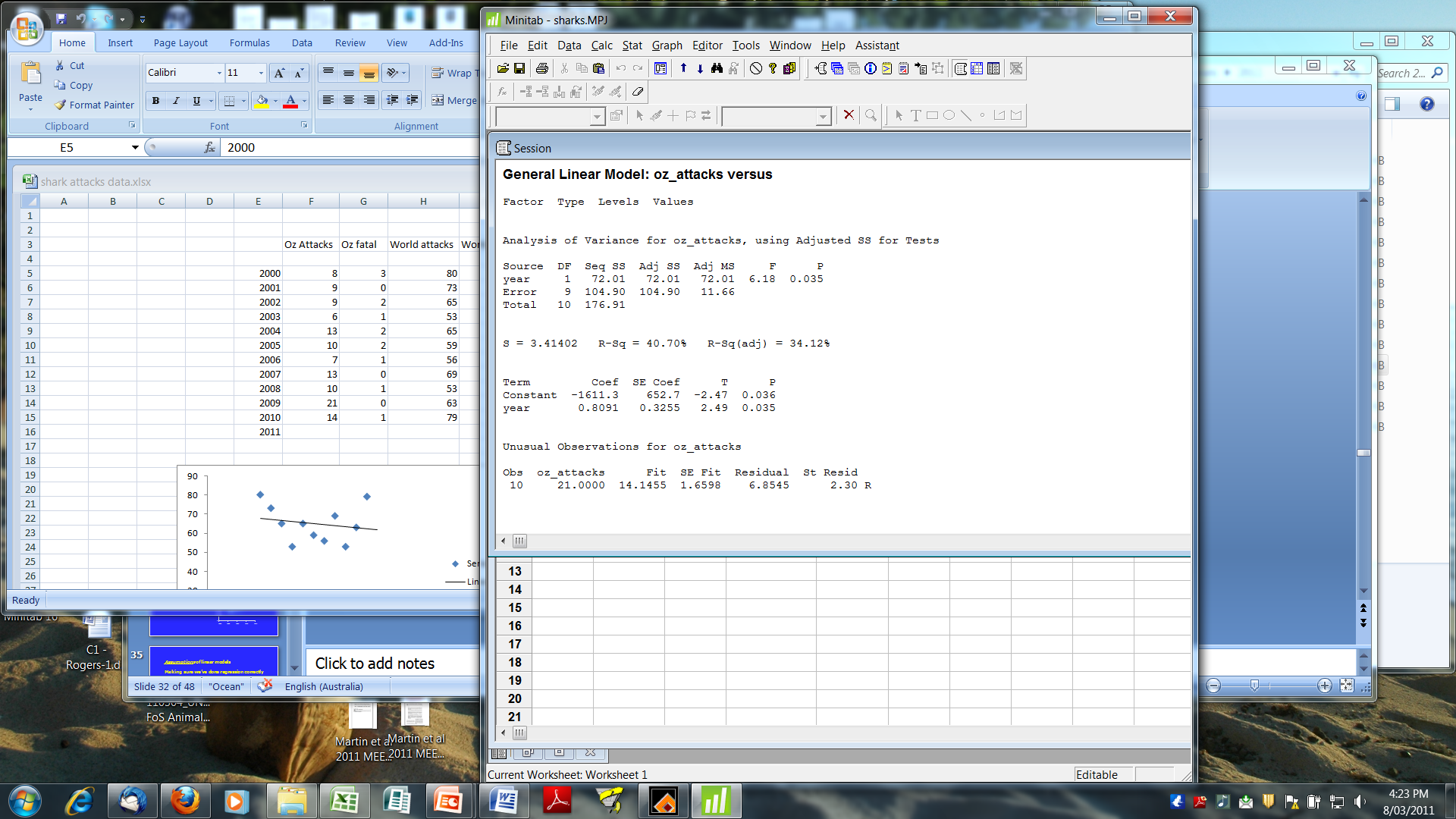
YEAR 1 0.45210

PERIOD 1 1.38278

PERIOD \*YEAR 1 0.12916

**QUESTION 6. 2011**

Below are the results from an analysis of data on the number of scuba diving accidents reported over a period of a number of years. The diving organization PADI is interested in knowing if there any changes over time.



Answer all parts 6A-6I.

**6A)** Based on the GLM results above, is there evidence for any trend over time? What number(s) support this conclusion? (2 marks)

**6B)** Is the trend increasing or decreasing? How do you know? (3 marks)

**6C)** What does the R2 represent? What is its value here? (2 marks)

**6D)** Show how you can calculate it from information in the table. (2 marks)

**6E)** What is the GLM being tested here (in words)? (3 marks)

**6F)** What is the full GLM being tested? (2 marks)

**6G)** What is the null hypothesis in words? And written as a GLM? (2 marks)

**6H)** Write out the full linear model including parameter values. (3 marks)

**6I)** What is the traditional name for this kind of analysis? (1 mark)

**QUESTION 6. 2012**

Below are the results from an analysis of data on the basal area of coarse woody debris from red gum forest with varying tree density due to thinning practices. The fire managers want to know if there is any difference in the build-up of coarse woody debris under forest thinning practices.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | SS | df | MS | F | P |
| Regression | 3.205 x 104 | 1 | 3.205 x 104 | 24.303 | <0.001 |
| Residual | 18465.57 | 14 | 1318.969 |  |  |

R2=0.634

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coefficient | Standard error | T | P |
| Intercept | -77.099 | 30.608 | -2.519 | 0.025 |
| Slope | 0.116 | 0.023 | 4.929 | <0.001 |

Answer all parts 6A-6H

**6A)** Based on the linear regression results above is there evidence for any trend in coarse woody debris with increasing tree density? (2 marks)

**6B)** Is the trend increasing or decreasing? How do you know? (3 marks)

**6C)** What does the R2 represent? What is its value here? (2 marks)

**6D)** What is the null hypothesis in words? (2 marks)

**6E)** Write out the full linear model including parameter values. (3 marks)

**6F)** What source of variation do the sums of squares for the residual measure? (3 marks)

**6G)** What source of variation do the sums of squares for the regression measure? (3 marks)

**6H)** Comment on the relative size of the sums of squares for the regression versus the residual. (2 marks)

**QUESTION 6. 2013**

Below are the results from an analysis of data on the density of larval fish with varying volumes of environmental flow. The river managers want to know if there is any difference in the numbers of fish recruiting under different environmental flow practices.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | SS | df | MS | F | P |
| Regression | 56240 | 1 | 56240 | 369.37 | <0.001 |
| Residual | 1827 | 12 | 152 |  |  |

R2=0.634

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coefficient | Standard error | T | P |
| Intercept | -35.781 | 6.060 | -5.90 | <0.001 |
| Slope | 8.8827 | 0.4622 | 19.22 | <0.001 |

Answer all parts 6A-6H.

**6A)** Based on the linear regression results above, is there evidence for any trend in larval fish density with environmental flow volume? (2 marks)

**6B)** Is the trend increasing or decreasing? How do you know? (3 marks)

**6C)** What does the R2 represent? What is its value here? (2 marks)

**6D)** What is the null hypothesis in words? (2 marks)

**6E)** Write out the full linear model including parameter values. (3 marks)

**6F)** What source of variation do the sums of squares for the residual measure? (3 marks)

**6G)** What source of variation do the sums of squares for the regression measure? (3 marks)

**6H)** Comment on the relative size of the sums of squares for the regression versus the residual. (2 marks)

**QUESTION 6. 2004**

Below are the results from an analysis of data on the numbers of koalas from mixed *Eucalyptus* forest with varying species composition due to fire history. The forest managers want to know if there is any difference in the numbers of koalas under control fire practices.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | SS | df | MS | F | P |
| Regression | 26082 | 1 | 26082 | 9.351 | <0.005 |
| Residual | 69728 | 25 | 2789 |  |  |

R2=0.522

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coefficient | Standard error | T | P |
| Intercept | 258.5 | 15.409 | 16.779 | <0.001 |
| Slope | -5.592 | 1.829 | -3.058 | <0.005 |

Answer all parts 6A to 6H.

**6A)** Based on the linear regression results above is there evidence for any trend in koalas with increasing fire frequency? (2 marks)

**6B)** Is the trend increasing or decreasing? How do you know? (3 marks)

**6C)** What does the R2 represent? What is its value here? (2 marks)

**6D)** What is the null hypothesis in words? (2 marks)

**6E)** Write out the full linear model including parameter values. (3 marks)

**6F)** What source of variation do the sums of squares for the residual measure? (3 marks)

**6G)** What source of variation do the sums of squares for the regression measure? (3 marks)

**6H)** Comment on the relative size of the sums of squares for the regression versus the residual. (2 marks)

**QUESTION 10: 2006**

Figure 1 below shows measurements of the weight of trees of two species that were grown for one year at a range of CO2 concentrations. Log-transformed data are shown in Fig. 2. Graphs of Residuals versus Fitted Values are shown for the original data in Fig. 3, and for the log-transformed data in Fig. 4. In all Figures, Species 1 is represented by diamonds and Species 2 by triangles

1. Discuss whether there is evidence of non-constant error variance, non-linearity, or clumping, or outliers in the data for Species 1. 6 marks
2. Discuss whether there is evidence of non-constant error variance, non-linearity, or clumping, or outliers in the data for Species 2. 6 marks
3. Discuss which of the lines shown in Figures 1 and 2 is the best model for weight of Species 1. 4 marks
4. Discuss which of the lines shown in Figures 1 and 2 is the best model for weight of Species 2. 4 marks

**Figure 1:** Tree weight versus CO2 concentration for Species 1 (diamonds) and 2 (triangles).

**Figure 2:** Log-transformed Weight and CO2 concentrations for Species 1 (diamonds) and for Species 2 (triangles).

**Figure 3:** Graphs of Residuals versus Fitted Values for regression of Tree Weight versus CO2 concentration for Species 1 (diamonds) and for Species 2 (triangles).

**Figure 4:** Graphs of Residuals versus Fitted Values for regression of log(Tree Weight) versus log(CO2 concentration) for Species 1 (diamonds) and for Species 2 (triangles).

# QUESTION 12: 2006

The Figures below show measurements of weekly relative growth rate (DW/W) of two plants species as a function of plant weight (W). Figure 1 shows results for Species 1, and Figure 2 shows results for Species 2. Figures show both the data and lines of best fit. The regression lines shown are:

Species 1: y = 0.0460 - 0.0112 \* x .

Species 2: y = 0.0370 - 0.0080 \* x .

where Y represents relative growth rate (DW/W) and X represents plant weight (W).

1. Give the equation of either a differential or difference-equation model for the rate of change in tree weight (W) that fits these data. Explain the meaning of each of the parameters in your equation. 10 marks
2. Use either the Figures or the equations to estimate approximate values of the initial relative growth rate, and asymptotic, maximum tree weight for each species. 10 marks

**Figure 1:** Weekly relative growth rate (DW/W) of Species 1 over a 52-week period.

**Figure 2:** Weekly relative growth rate (DW/W) of Species 2 over a 52-week period.

**QUESTION 11:2007**

# The data below give the Systolic blood pressures (BP) of 18 males of various ages. Data are shown in the Table and in the Figure below for 9 males who are on a low-salt diet (triangles) and for 9 males who are not on a low-salt diet (boxes).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MEN ON LOW-SALT DIET** | | |  | **MEN NOT ON LOW-SALT DIET** | | |
| Subject | **AGE**  **(years)** | **Blood Pressure (mm Hg)** |  | Subject | **AGE**  **(years)** | **Blood Pressure (mm Hg)** |
| 1 | 19 | 122 |  | 10 | 18 | 124 |
| 2 | 25 | 126 |  | 11 | 26 | 128 |
| 3 | 30 | 128 |  | 12 | 30 | 130 |
| 4 | 42 | 129 |  | 13 | 39 | 135 |
| 5 | 46 | 133 |  | 14 | 45 | 135 |
| 6 | 52 | 138 |  | 15 | 53 | 144 |
| 7 | 57 | 134 |  | 16 | 56 | 146 |
| 8 | 62 | 145 |  | 17 | 64 | 155 |
| 9 | 70 | 148 |  | 18 | 72 | 165 |

The data have been analysed using Minitab by setting up a dummy variable called DIET with its value equal to 0 for men NOT on a diet and 1 for men on a Low-Salt diet. Data were fitted to the model:

BP = **a** + **b** \* DIET + **c** \* AGE + **d**  \* DIET \* AGE.

The MINITAB analysis is shown below.

1. What are estimated values of **a**, **b**, **c** and **d**? 4 marks
2. Calculate the predicted blood pressure at age 60 years for each group.

6 marks

1. What are the y-intercepts and slopes of the relationships between BP and AGE for the two diet groups ? 4 marks
2. Use the Minitab analysis below to discuss whether the slopes and y-intercepts of the relationship between BP and AGE are significantly different for the two diet groups. If they are different, what are the probability levels for differences?

4 marks

1. Use the Minitab analysis to discuss how the low-salt diet affects blood pressure in young versus older men.

2 marks

**MINITAB Regression Analysis of**

**Blood pressure (BP) versus AGE, DIET, and DIET \* AGE**

The regression equation is

BP = 108 + 4.70 DIET + 0.723 AGE - 0.251 DIETxAGE

Predictor Coef SE Coef T P

Constant 107.845 2.989 36.08 0.000

DIET 4.700 4.306 1.09 0.293

AGE 0.72306 0.06237 11.59 0.000

DIETxAGE -0.25138 0.09008 -2.79 0.014

S = 3.196 R-Sq = 93.6% R-Sq(adj) = 92.3%

Analysis of Variance

Source DF SS MS F P

Regression 3 2103.97 701.32 68.67 0.000

Residual Error 14 142.98 10.21

Total 17 2246.94