The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2014–15**

Linear Models **2 hours**

*Marks will be awarded for your best **three** answers.*

RESTRICTED OPEN BOOK EXAMINATION

Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator that conforms to University regulations.

There are 60 marks available on the paper.

Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

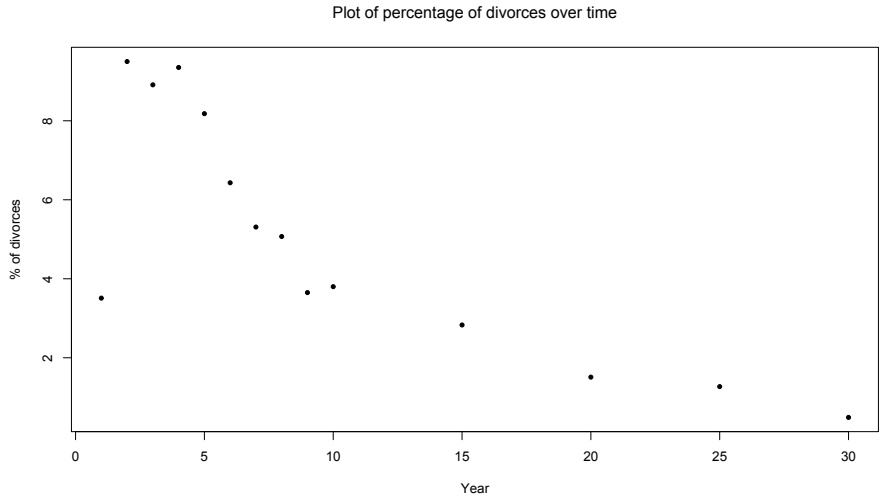

Figure 1: Percentage of divorces over time.

1 The data in the table below give the percentage of divorces caused by adultery per year of marriage[1]

Year	1	2	3	4	5	6	7
%	3.51	9.50	8.91	9.35	8.18	6.43	5.31
Year	8	9	10	15	20	25	30
%	5.07	3.65	3.80	2.83	1.51	1.27	0.49

(i) Figure **??** above plots the percentages of divorces over time (in years). Briefly describe the data based on this plot and suggest whether the percentage is constant over time. ***(2 marks)***

[1]Source: Bingham, N.H. and Fry, J.M. (2010) *Regression: Linear Models in Statistics*, Springer, p. 125.

1 (continued)

(ii) It is decided to fit a polynomial model in order to describe the relationship of the percentage (as response variable y) and Year (as a covariate x). The suggested model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \epsilon_i,$$

where $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are the regressor coefficients and ϵ_i is the error term. The following R output was produced

```
Call:
lm(formula = y ~ x + I(x^2) + I(x^3) + I(x^4))

Residuals:
     Min       1Q   Median       3Q      Max
-2.91296 -0.95566 -0.03176  1.06910  2.21700

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.0366750  2.1037831   2.394   0.0403
x            1.6792834  1.0247317   1.639   0.1357
I(x^2)      -0.3084371  0.1475975  -2.090   0.0662
I(x^3)       0.0156855  0.0075646   2.074   0.0680
I(x^4)      -0.0002484  0.0001245  -1.995   0.0772
---

Residual standard error: 1.702 on 9 degrees of freedom
Multiple R-squared:  0.7887,Adjusted R-squared:  0.6948
F-statistic: 8.398 on 4 and 9 DF,  p-value: 0.004169
```

(a) Write down the estimated relationship of x and y. ***(1 mark)***

(b) Given the additional commands

```
> qt(0.95, 12)        > qnorm(0.95)
[1] 1.782288          [1] 1.644854

> qt(0.95,13)         > qt(0.975,12)
[1] 1.770933          [1] 2.178813
```

Provide 90% confidence intervals for β_1 and β_2. ***(3 marks)***

(c) Comment on the overall adequacy of the model fit. You should comment on the plot (Figure 1), coefficient of determination, the F-statistic and the residuals. You should write down any hypothesis you are using. ***(5 marks)***

1 (continued)

(iii) The following R output gives the analysis of variance (ANOVA) table of the model in (ii) above:

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x          1 80.966  80.966 27.9594 0.000502
I(x^2)     1  2.826   2.826  0.9759 0.349028
I(x^3)     1  1.964   1.964  0.6783 0.431466
I(x^4)     1 11.522  11.522  3.9789 0.077209
Residuals  9 26.063   2.896
```

(a) Use this table to test the hypothesis $H_0 : \beta_3 = \beta_4 = 0$ against the alternative H_1 that $\beta_3 \neq 0$ or $\beta_4 \neq 0$. You can use the additional R output:

```
> pf(2.328, 2, 9)
[1] 0.8468435
```

(3 marks)

(b) Based on the ANOVA table above, perform suitable tests in order to obtain the best model; state clearly the null hypotheses you are testing. Write down what you consider to be the best model.

(3 marks)

(c) Explain why the ANOVA table above cannot be used to help perform the test H_0: $\beta_1 = \beta_4 = 0$ and suggest how this may be undertaken in R (you are not asked to perform such a test). ***(3 marks)***

2 The table below (unknown source) displays data that relate to the number of oil changes per year and the cost of engine repairs.

Oil changes per year	3	5	2	3	1	4	6	4
Cost of repair (US$)	300	300	500	400	700	400	100	2250
Oil changes per year	3	2	0	10	7			
Cost of repair (US$)	450	50	600	0	150			

It is suggested that the Cost of repair (variable `cost`) is linearly correlated with the Oil changes per year (variable `changes`), hence a simple linear model with response the `cost` and covariate the `changes` is proposed to explore this relationship.

(i) The following commands in R are used to obtain the residuals and the standardized residuals of this linear model.

```
> b <- lm(cost~changes)
> b$resid
       1        2        3        4        5        6
-224.775 -111.669  -81.328 -124.7 75   62.118  -68.222
       7        8        9       10       11       12
-255.116 1781.777  -74.775 -531.328  -94.434 -128.904
      13
-148.563
>
> stdres(b)
       1        2        3        4        5        6
-0.40693 -0.20296 -0.14983 -0.22589 0.11816 -0.12296
       7        8        9       10       11       12
-0.47387  3.21159 -0.13537 -0.97888 -0.18875 -0.32177
      13
-0.28635
```

(a) Calculate the deletion residuals. ***(4 marks)***

(b) In order to apply the Sidak correction we use the following R command for the new level of the t distribution

```
qt(0.001968932, 10)
[1] -3.725742
```

Explain how the value of 0.001968932 is obtained. ***(3 marks)***

(c) Using (a) and (b) identify any possible outliers. ***(3 marks)***

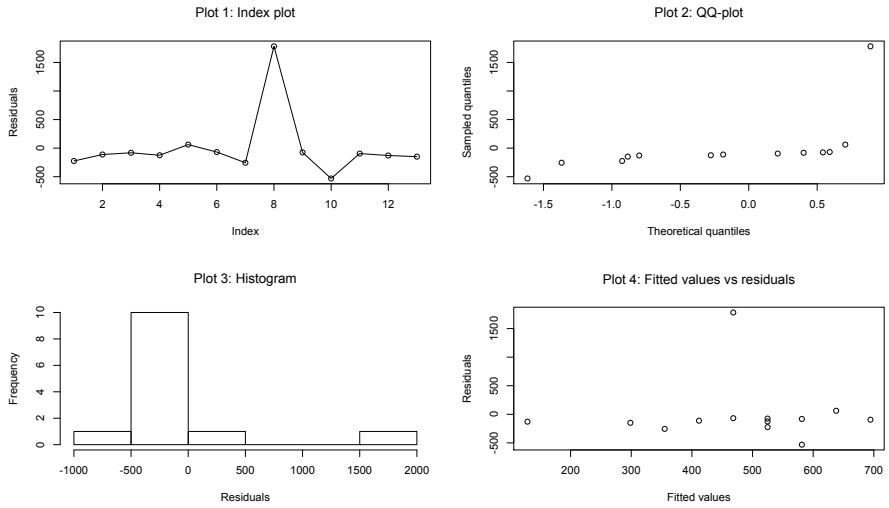

Figure 2: Diagnostic plots.

2 (continued)

(ii) Figure **??** above shows four diagnostic plots for the simple linear regression model that is fitted in part (i) above. Plot 1 shows the index plot of the residuals, Plot 2 shows the QQ-plot of the residuals, Plot 3 shows the histogram of the residuals and Plot 4 shows the plot of the residuals against the fitted values of the model. Comment on these plots, particularly addressing the following points:

(a) is there evidence from the data to support the model assumptions? ***(3 marks)***

(b) is the fit adequate? ***(1 mark)***

2 (continued)

(iii) A further analysis is conducted in order to explore the influence the `change` regressor variable has in the linear model. Below is part of an R output that shows the hat values and the Cook's distance.

```
> lm.influence(b)
$hat
         1          2          3          4          5
0.08527828 0.09245961 0.11669659 0.08527828 0.17145422
         6          7          8          9         10
0.07719928 0.13105925 0.07719928 0.08527828 0.11669659
        11         12         13
0.24955117 0.51885099 0.19299820

> cooks.distance(b)
      1         2         3        4         5         6
0.007719 0.002098 0.001482 0.002378 0.001444 0.000632
      7         8         9       10        11       12
0.016934 0.431437 0.000854 0.063296 0.005923 0.055825
     13
0.009804
```

(a) Use the above output in order to assess which values of the explanatory variable `change` are influential. ***(3 marks)***

(b) Based on your answers in part (ii) and part (iii,a) suggest how the model fit may be improved (give reasoning for your suggestions). ***(3 marks)***

3 (i) In the context of variable selection for linear models show that the Bayesian information criterion (BIC) can be expressed in terms of the coefficient of determination (R^2) as

$$\text{BIC} = n \log[\boldsymbol{y}^T \boldsymbol{y}(1 - R^2)] - (n - p) \log n,$$

where n is the number of observations and p is the number of explanatory variables, for a linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, *not* including an intercept.

(8 marks)

3 (continued)

(ii) The table below shows an extract from data on 35 Scottish hill races. The first column gives the names of the hills considered in the study, the second column shows the overall race distance (in miles), the third column shows the total height climbed (in feet) and the last column shows the record time (in minutes). The analyst wishes to fit a model to the full data set and to predict the record time. To this end the explanatory variables `distance` and `climb` are considered as well as `distance`^{2} and `climb`^{2}.

Location	distance	climb	time
Greenmantle	1	100	16.083
Carnethy	6.0	2500	48.350
Craig Dunain	6.0	900	33.650
Ben Rha	7.5	800	45.600

A full stepwise variable selection, given in the R output below, is considered in order to select the explanatory variables that need to be included in the linear model.

```
> a <- lm(time~1, data=hills)
> step(a,scope=list(upper=time~climb+dist+I(climb^2)+I(dist^2)))
Start:  AIC=274.88
time ~ 1

              Df Sum of Sq   RSS    AIC
+ dist         1     71997 13142 211.49
+ I(dist^2)    1     62545 22594 230.45
+ climb        1     55205 29934 240.30
+ I(climb^2)   1     52958 32181 242.83
<none>                    85138 274.88

Step:  AIC=211.49
time ~ dist

              Df Sum of Sq   RSS    AIC
+ I(climb^2)   1      8439  4703 177.52
+ climb        1      6250  6892 190.90
+ I(dist^2)    1       793 12348 211.31
<none>                    13142 211.49
- dist         1     71997 85138 274.88

Step:  AIC=177.52
time ~ dist + I(climb^2)
```

3 (continued)

```
             Df Sum of Sq   RSS    AIC
+ I(dist^2)   1     346.5  4356 176.84
<none>                     4703 177.52
+ climb       1     188.1  4515 178.09
- I(climb^2)  1    8439.0 13142 211.49
- dist        1   27478.3 32181 242.83

Step:  AIC=176.84
time ~ dist + I(climb^2) + I(dist^2)

             Df Sum of Sq     RSS    AIC
<none>                     4356.1 176.84
- I(dist^2)   1     346.5  4702.6 177.52
+ climb       1      50.5  4305.6 178.43
- dist        1     722.9  5079.0 180.21
- I(climb^2)  1    7992.2 12348.3 211.31

Call:
lm(formula = time ~ dist + I(climb^2) + I(dist^2), data = hills)

Coefficients:
(Intercept)         dist   I(climb^2)    I(dist^2)
  1.046e+01    3.785e+00    1.950e-06    8.716e-02
```

(a) Give the number of steps used in the selection and, for each step, briefly explain what action is being performed. ***(7 marks)***

(b) Based on the above analysis recommend the overall best model including the selected regressor variables and the estimated coefficients. ***(2 marks)***

(c) For the selected model perform a manual calculation to verify the value of the AIC given in the R output above. ***(3 marks)***

4 Consider the linear model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$ and the design matrix X is assumed to have full rank.

(i) Show that the vector covariance of the residuals $\boldsymbol{e} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$ and $\boldsymbol{y}$ is

$$Cov(\boldsymbol{e}, \boldsymbol{y}) = \sigma^2 M,$$

where $M = I_n - X(X^TX)^{-1}X^T$ and $\hat{\boldsymbol{\beta}}$ is the usual least squares estimator $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y}$. ***(2 marks)***

(ii) If m_{ii} is the i-th diagonal element of M find an expression for m_{ii} in terms of the design matrix X and its elements $(i = 1, 2, \ldots, n)$.

Show $0 < m_{ii} < 1$ and suggest whether it is possible to have $m_{ii} \approx 1$. ***(4 marks)***

(iii) Define the i-th standardized residual and the i-th deletion residual respectively as

$$s_i = \frac{e_i}{\hat{\sigma}\sqrt{m_{ii}}} \quad \text{and} \quad s_{-i} = \frac{e_i}{\hat{\sigma}_{-i}\sqrt{m_{ii}}}, \quad i = 1, 2, \ldots, n,$$

where $\hat{\sigma}$ is the residual standard error and $\hat{\sigma}_{-i}$ is the residual standard error if the i-th observation is removed from the data.

(a) If $m_{ii} \approx 1$ show that

$$\frac{\boldsymbol{e}_{-i}^T\boldsymbol{e}_{-i}}{\hat{\sigma}^2 m_{ii}} = n - p - s_i^2,$$

where $\boldsymbol{e}_{-i}$ denotes the residual vector if we remove from the data the i-th observation. ***(4 marks)***

(b) If $m_{ii} \approx 1$ show that

$$\frac{\hat{\sigma}}{\hat{\sigma}_{-i}} = \sqrt{\frac{n-p-1}{n-p-s_i^2}}.$$

(8 marks)

(c) If $m_{ii} \approx 1$ show that

$$s_{-i} = s_i\sqrt{\frac{n-p-1}{n-p-s_i^2}}.$$

(2 marks)

End of Question Paper