The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2014–2015**

Linear Models **3 hours**

Marks will be awarded for your best ***five*** *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 100 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

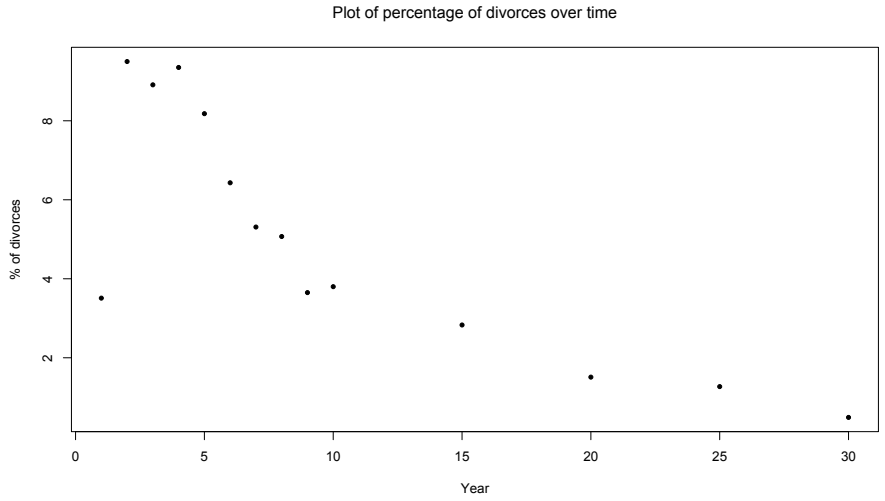


Figure 1: Percentage of divorces over time.

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 1 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 standardised 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 standardised 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)***

(ii) Figure 2 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.

(a) Assess whether the model assumptions are supported. ***(3 marks)***

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

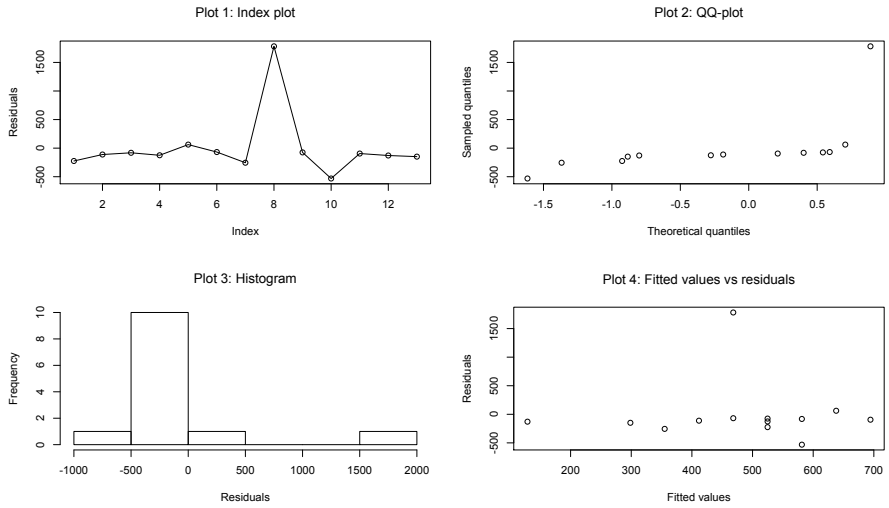


Figure 2: Diagnostic plots.

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 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 standardised 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

$$\frac{\boldsymbol{e}_{-i}^T\boldsymbol{e}_{-i}}{\hat{\sigma}^2 m_{ii}} = n - p - s_i^2, \tag{1}$$

where $\boldsymbol{e}_{-i}$ denotes the residual vector if we remove from the data the i-th observation. HINT: Assume (1) is true and show $\hat{\sigma}^2 = \boldsymbol{e}^T\boldsymbol{e}/(n-p)$. ***(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)

4 A study has been conducted to test the effect of a new drug on lowering systolic blood pressure. Forty patients with high blood pressure are recruited to the study, and are randomly assigned to either the new (active) drug, or a placebo, with twenty patients in each group. Following treatment with the active drug or placebo, each patient's blood pressure is recorded on five occasions. In R, each blood pressure measurement is stored in a variable `pressure`, the drug corresponding to each measurement is stored in a factor variable `treatment`, taking levels `A` for placebo and `B` for the active drug, and the patient corresponding to each measurement is stored in a factor variable `patient`. Below is some edited output from an R session.

```
> fm1<-lmer(pressure~treatment+(1|patient))
> summary(fm1)
Linear mixed model fit by REML ['lmerMod']
Formula: pressure ~ treatment + (1 | patient)

Random effects:
 Groups   Name        Variance Std.Dev.
 patient  (Intercept) 10.575   3.252
 Residual              3.753   1.937
Number of obs: 200, groups:  patient, 40

Fixed effects:
            Estimate
(Intercept) 140.8649
treatmentB   -1.9624
```

(i) Write down the equation of the model that has been fitted and assigned to the name `fm1`, defining your notation carefully. Specify appropriate distributions for the terms in the model, and write down any necessary parameter constraints. ***(4 marks)***

(ii) Briefly justify the modelling choice of random effects in the fitted model. ***(1 mark)***

(iii) Give the estimated values for all four parameters in your model in (i). ***(1 mark)***

(iv) Calculate the estimated variance for any observation. ***(2 marks)***

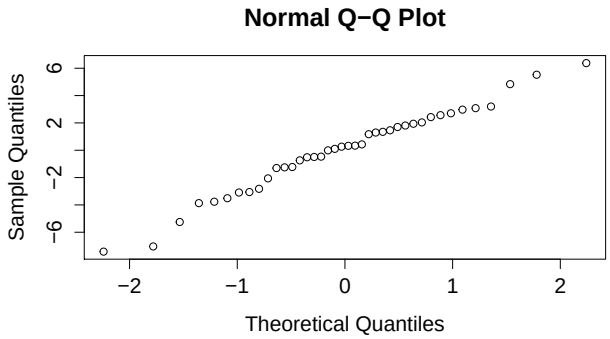
 Question 4 continued on next page

4 (continued)

(v) The R command

```
qqnorm(unlist(ranef(fm1)$patient))
```

is used to produce the following plot.



Explain the reason for producing this plot. If a reference line were to be added to the plot to aid interpretation, what should the gradient of the line be? *(2 marks)*

(vi) The estimator for the `(Intercept)` term is the mean of all the placebo observations, and the estimator for the `treatmentB` term is the difference between the mean of all the active drug observations, and the mean of all the placebo observations. Calculate the estimated standard error for the `treatmentB` term. *(4 marks)*

(vii) The session is continued below.

```
> fm2<-lmer(pressure~1+(1|patient))
> x <- - 2*(logLik(fm2)-logLik(fm1))
> N<-100
> y<-rep(0,N)
> for(i in 1:N){
+   z<-unlist(simulate(fm2))
+   fm2b<-lmer(z~1+(1|patient))
+   fm1b<-lmer(z~drug+(1|patient))
+   y[i]<- -2*(logLik(fm2b)-logLik(fm1b))
+ }
> mean(y>=x)
[1] 0.12
```

Give the name of the procedure that has been used, state what is being tested, and interpret the output.

(3 marks)

4 (continued)

(viii) In the variable `pressure`, elements 1 to 100 correspond to the placebo observations, and elements 101 to 200 correspond to the active drug observations. Some further R output is given below.

```
> t.test(pressure[1:100],pressure[101:200])

Welch Two Sample t-test

data:  pressure[1:100] and pressure[101:200]
t = 3.7218, df = 197.966, p-value = 0.0002577
alternative hypothesis: true difference in means is not equal to 0
```

Compare the result in this analysis with the result with the analysis in part (vii), and explain the difference. ***(3 marks)***

5 Data are collected on 80 individuals in a study assessing the effect of age and smoking status on lung cancer risk. The variables recorded are:

- X_1 - smoking status ($X_1 = 1$ for current smokers and $X_1 = 0$ for ex or non-smokers) abbreviated to `smoke` in the `R` analysis.
- X_2 - age.
- Y - lung cancer status ($Y = 1$ for people with lung cancer and $Y = 0$ for people without diagnosed lung cancer) abbreviated to `cancer` in the `R` analysis.

Various generalized linear models, with a logit link, are fitted where the binary variable Y is the response and X_1 and X_2 are the explanatory variables. Let η_i be the linear predictor for the i-th person. The four fitted models are:

- Model 1: $\eta_i = \beta_0 + \beta_1 X_{1i}$
- Model 2: $\eta_i = \beta_0 + \beta_2 X_{2i}$
- Model 3: $\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$
- Model 4: $\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i}$

In answering this question, you may find the following quantiles useful: $\chi^2_{77,0.95} = 98.48$, $\chi^2_{76,0.95} = 97.35$, $\chi^2_{3,0.95} = 7.81$, $\chi^2_{2,0.95} = 5.99$, $\chi^2_{1,0.95} = 3.84$

(i) What is the relationship between $E(Y_i)$ and η_i for the logit link? What would be the relationship if a probit link were used instead? ***(2 marks)***

(ii) The residual deviances for the four models are given in Table 1. By considering the changes in residual deviance determine the relationship between lung cancer risk, age and smoking status. For any hypothesis tests that you do, state clearly the null hypothesis and the degrees of freedom of the relevant χ^2 distribution used to perform the test. ***(6 marks)***

(iii) Comment on the fit of the model you selected in part (ii) based on an appropriate χ^2 distribution. ***(2 marks)***

Model	Residual deviance
Model 1	59.36
Model 2	80.41
Model 3	38.58
Model 4	32.26

Table 1: Residual deviances for Models 1 to 4.

5 (continued)

(iv) Using `R`, the following output is obtained for Model 4:

```
Call: glm(formula = cancer ~ smoke * age, family = binomial)
Coefficients:
(Intercept)        smoke          age    smoke:age
   -121.456      118.886        1.854       -1.777

Degrees of Freedom: 79 Total (i.e. Null);  76 Residual
Null Deviance:     104.8
Residual Deviance: 32.26  AIC: 40.26
AIC: 20.36
```

Using this output, calculate the odds of lung cancer for a 70 year old current smoker. ***(2 marks)***

(v) Based on the output from (iv), at what age would the estimated odds of lung cancer for smokers and non-smokers be the same? ***(2 marks)***

(vi) For Model 1 write down an expression for the log-likelihood (l) in terms of β_0, β_1 and x_{1i} and hence calculate $\dfrac{\partial^2 l}{\partial {\beta_0}^2}$ in terms of η_i. How might $\dfrac{\partial^2 l}{\partial {\beta_0}^2}$ be useful when making inferences in a generalized linear model? ***(6 marks)***

6 Data were collected relating to the opinions of local residents to wind turbines being erected in North Yorkshire in 2014. The interest was in the factors affecting opinions about the erection of wind turbines (indicated by the factor variable `opinion`). The two main factors of interest were whether the resident was born in the area (indicated by the factor variable `birthplace`) and who owned the turbine (indicated by the factor variable `owner`). `owner` is either `multinational`, `local_company` or `community`, `birthplace` is either `local` or `incomer`, `opinion` is either `object` or `approve` and `count` records the number of people in each category. The data is stored in the data frame `turbine` in `R`. The first few lines of the data frame are shown below.

```
> head(turbine)
   count           owner   birthplace   opinion
1     21   multinational        local    object
2     18   multinational        local   approve
3     19   multinational      incomer    object
4     15   multinational      incomer   approve
5     10   local_company        local    object
6     51   local_company        local   approve
```

The full data are shown in Table 2.

(i) By calculating relevant proportions from Table 2, make some brief observations about how the probability of objecting to the erection of wind turbines relates to birthplace and turbine ownership. ***(3 marks)***

(ii) Various Poisson log-linear models are fitted to the data. The residual deviances for the models are given below along with some relevant quantiles.

Model Number	Model	Residual Deviance	Residual df
1	owner*birthplace+opinion	26.994	5
2	owner*birthplace+opinion*owner	16.533	3
3	owner*birthplace+opinion*birthplace	15.619	4

$\chi^2_{5,0.95} = 11.07$, $\chi^2_{4,0.95} = 9.49$, $\chi^2_{3,0.95} = 7.81$, $\chi^2_{2,0.95} = 5.99$, $\chi^2_{1,0.95} = 3.84$

Describe the most suitable model for the data by referring to the residual deviances. For any hypothesis tests that you do, state clearly the degrees of freedom of the relevant χ^2 distribution used to perform the test.

(5 marks)

local			incomer		
owner	object	approve	owner	object	approve
multinational	21	18	multinational	19	15
local_company	10	51	local_company	19	15
community	2	4	community	9	8

Table 2: The `turbine` data.

6 (continued)

(iii) Does the selected model fit well and is it consistent with your observations from part (i)? ***(2 marks)***

(iv) What does the model `owner*birthplace` say about the probability of objecting to the erection of turbines? ***(1 mark)***

(v) Calculate the Pearson residual for the model `owner*birthplace+opinion` for residents born in the local area who objected to community-owned turbines. ***(4 marks)***

(vi) Consider the model `owner*birthplace+opinion*owner`. Let i represent `opinion`, j represent `birthplace` and k represent `owner`. Assuming that $\mu_{ijk} = n_{jk}\pi_{ijk}$ and that $\pi_{ijk} = \pi_{ik}$, show that the maximum likelihood estimator of π_{ijk} is $\dfrac{\sum_j y_{ijk}}{\sum_j n_{jk}}$. ***(5 marks)***

End of Question Paper