Queen Mary University of London

Main Examination period 2018

MTH5120: Statistical Modelling I

Duration: 2 hours

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

You should attempt ALL questions. Marks available are shown next to the questions.

Only non-programmable calculators that have been approved from the college list of non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

Statistical functions provided by the calculator may be used provided that you state clearly where you have used them.
The New Cambridge Statistical Tables are provided.

Complete all rough work in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it shall be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

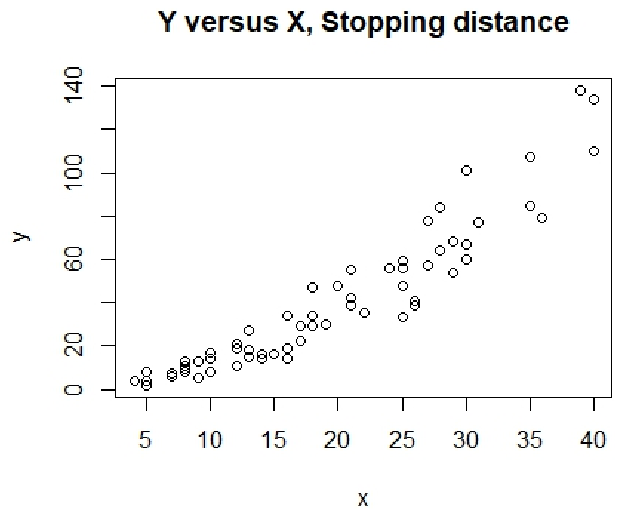
Exam papers must not be removed from the examination room.

Examiners: L I Pettit, I Goldsheid

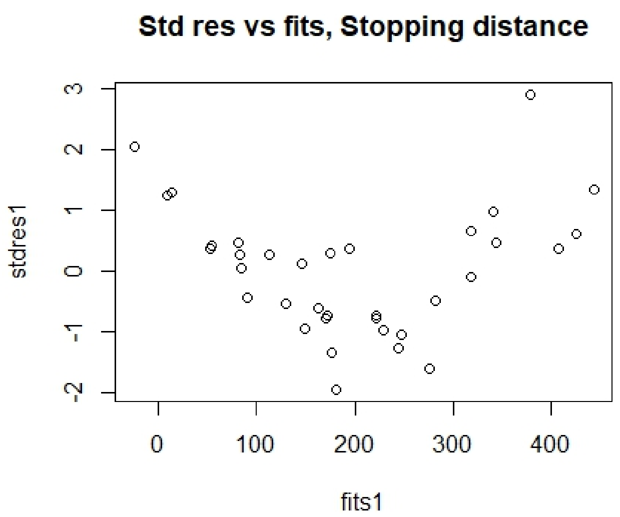
Question 1. **[24 marks]** A company manufacturing cars was testing one of their models to test the stopping distance y (in feet) as a function of speed x (in miles per hour). Several cars were tested by the same driver. Sixty three observations were obtained for a number of different speeds.

(a) A simple linear regression model of Y on X was fitted to the data. Write down the model and state the assumptions made. [**4**]

(b) A scatterplot is given below. Comment on whether the simple linear regression model seems appropriate. [**2**]



(c) The plot of standardized residuals versus fitted values is given below. Comment on the linearity of the model and the constant variance assumption. [**2**]



(d) It was decided to transform the response variable using a square root transformation. The following commands and output were produced using R.

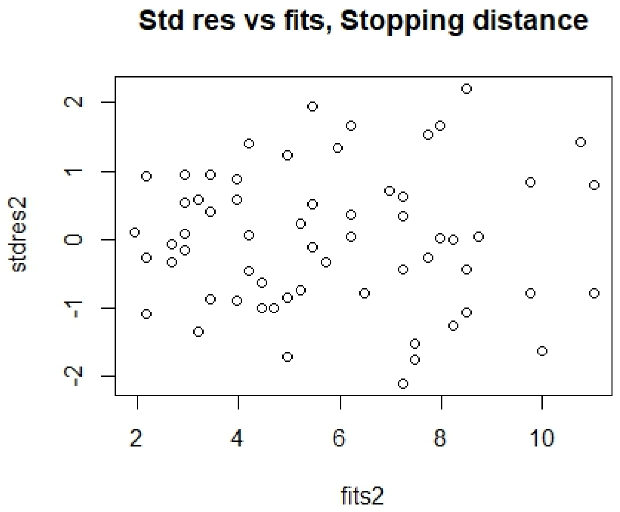
```
> sy <- (y^0.5)
> modsy <- lm(sy ~ x)
> summary(modsy)

Call:
lm(formula = sy ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4879 -0.5487  0.0098  0.5291  1.5545

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.918283   0.197406   4.652 1.82e-05 ***
x           0.252568   0.009246  27.317  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7193 on 61 degrees of freedom
Multiple R-squared:  0.9244,Adjusted R-squared:  0.9232
F-statistic: 746.2 on 1 and 61 DF,  p-value: < 2.2e-16
```

(i) Write down the fitted model. [**2**]

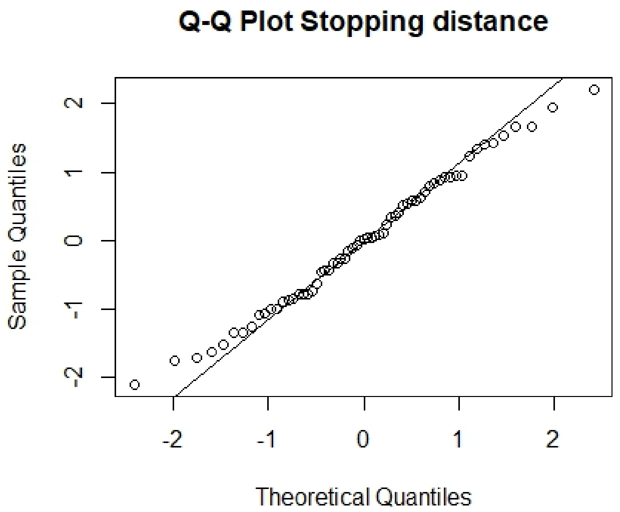
(ii) Writing the slope parameter as β_1 what is the conclusion of a test of the null hypothesis $H_0 : \beta_1 = 0$ against a two sided alternative? [**2**]

(iii) Find a 95% confidence interval for the intercept parameter β_0. [**4**]

(e) The plot of standardized residuals versus fitted values is given below. Comment on whether the assumptions of linearity and constant variance seem to be satisfied by this transformed model. [**3**]



(f) A Q-Q plot of the standardized residuals is shown below. What assumption is this plot examining? What is your conclusion? What test could be carried out to check this assumption? [**5**]



Question 2. [**27 marks**]

(a) Write the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \ldots, n$$

where $\varepsilon_i \sim N(0, \sigma^2)$ as a general linear model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

by identifying $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$. [**5**]

(b) State the distribution of $\boldsymbol{\varepsilon}$. [**3**]

(c) Hence find the least squares estimators of β_0 and β_1. [**4**]

(d) Using the result that

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

find $\mathrm{Var}(\hat{\beta}_0)$, $\mathrm{Var}(\hat{\beta}_1)$ and $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$. [**6**]

(e) The hat matrix is defined as

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

Show that it is symmetric and idempotent. [**4**]

(f) Show that the vector of fitted values $\hat{\boldsymbol{Y}}$ is given by

$$\hat{\boldsymbol{Y}} = \boldsymbol{HY}.$$

[**2**]

(g) Hence find the vector of fitted values for the simple linear regression model. [**3**]

Question 3. [**25 marks**] A group of bears were studied in an American National Park. In order to find their weight it was necessary to anaesthetise them and it was hoped an adequate estimate of their weight could be determined from various body measurements and records of their age. Accordingly 35 bears were captured and their weight (in pounds) determined. The possible regressor variables were age $x1$ (in months) and the following body measurements in inches, head length $x2$, neck girth $x3$ and chest girth $x4$. A preliminary analysis suggested that a log transform of the weight, denoted by ly was necessary.
To find the best fitting model the method of backwards fitting was to be employed.

(a) Describe this method of fitting a multiple regression model as implemented in R, including a definition of the AIC. [**6**]

(b) The following R output was obtained. Which variables are dropped at each step and which retained in the final chosen model? [**2**]

```
> ly<-log(y)
> modly <- lm(ly ~ x1+x2+x3+x4)
> reduced.model <- step(modly,direction="backward")
Start:  AIC=-119.87
ly ~ x1 + x2 + x3 + x4

       Df Sum of Sq     RSS     AIC
- x1    1  0.006313 0.86252 -121.61
- x2    1  0.029508 0.88572 -120.69
<none>              0.85621 -119.87
- x3    1  0.233978 1.09019 -113.42
- x4    1  0.261001 1.11721 -112.56

Step:  AIC=-121.61
ly ~ x2 + x3 + x4

       Df Sum of Sq     RSS     AIC
- x2    1  0.043757 0.90628 -121.88
<none>              0.86252 -121.61
- x4    1  0.254722 1.11725 -114.56
- x3    1  0.301786 1.16431 -113.11

Step:  AIC=-121.88
ly ~ x3 + x4

       Df Sum of Sq     RSS     AIC
<none>              0.90628 -121.88
- x4    1   0.33089 1.23717 -112.99
- x3    1   0.33491 1.24119 -112.88
```

(c) For the chosen model you wish to check the assumptions. Say what plots you would look at and why. **[4]**

(d) The following plot (on the next page) shows the standardised residuals versus $x4$. Justify the decision to add a quadratic term in $x4$ to the model. **[2]**

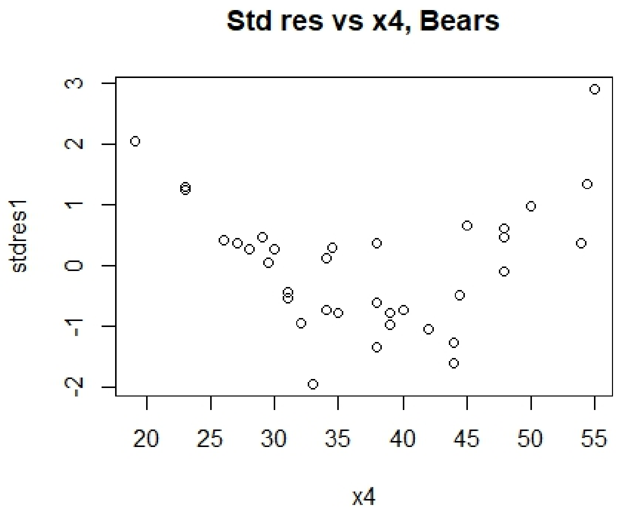
(e) This part refers to the output on page 7 and the plots on page 8.

(i) Write down the final fitted model from the output below. **[2]**

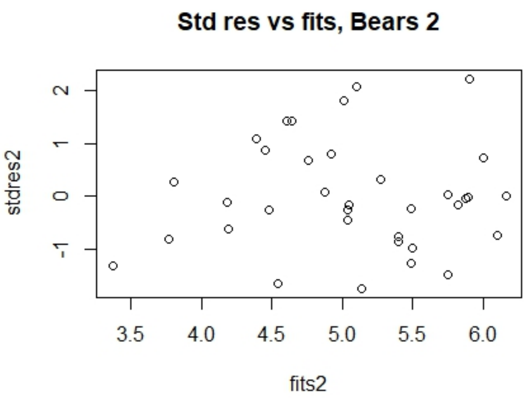
(ii) Comment on the overall fit of this model. **[6]**

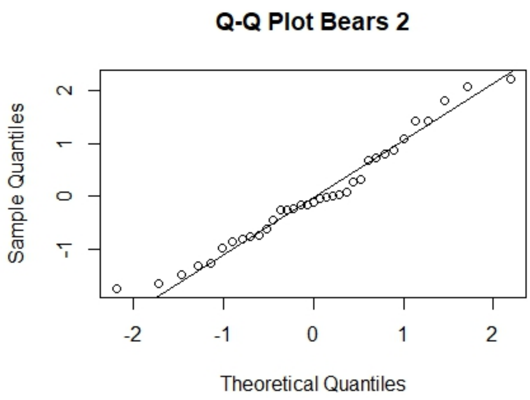
(iii) A statistician objected to the way that backwards fitting was used for deciding to drop some variables but a quadratic term was added based on the residual plots. What would be your reply? **[3]**

```
> modlyrq<-lm(ly~x3+poly(x4,2,raw=TRUE))
> summary(modlyrq)

Call:
lm(formula = ly ~ x3 + poly(x4, 2, raw = TRUE))
```




```
Residuals:
     Min       1Q   Median       3Q      Max
-0.18106 -0.07537 -0.01130  0.07220  0.21296

Coefficients:
                           Estimate Std. Error t value
(Intercept)               0.4304006  0.2739114   1.571
x3                        0.0634505  0.0126381   5.021
poly(x4, 2, raw = TRUE)1  0.1395401  0.0165976   8.407
poly(x4, 2, raw = TRUE)2 -0.0013136  0.0001932  -6.798
                         Pr(>|t|)
(Intercept)                 0.126
x3                       2.02e-05 ***
poly(x4, 2, raw = TRUE)1 1.70e-09 ***
poly(x4, 2, raw = TRUE)2 1.30e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1083 on 31 degrees of freedom
Multiple R-squared:  0.9794,Adjusted R-squared:  0.9774
F-statistic: 490.2 on 3 and 31 DF,  p-value: < 2.2e-16
```





Question 4. [24 marks] A transport company is interested in the relationship between the time Y required to handle shipments of chemicals in drums, the number of drums X_1 in the shipment and the total weight of the shipment X_2. Data on $n = 20$ shipments were collected and the following calculations for a multiple regression analysis of the model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

were obtained:

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} 0.307 & -0.033 & 0.015 \\ -0.033 & 0.012 & -0.012 \\ 0.015 & -0.012 & 0.014 \end{pmatrix}, \boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} 1889.0 \\ 27246.0 \\ 21648.8 \end{pmatrix}.$$

Also $\boldsymbol{Y}^T\boldsymbol{Y} = 242449.0$ and $\bar{Y} = 94.45$.

(a) Find the least squares estimates $\hat{\boldsymbol{\beta}}$ and hence write down the fitted model. **[4]**

(b) Use the results to construct the Analysis of Variance Table. **[12]**

(c) Test the null hypothesis that the overall regression is non-significant using a significance level of 5%. **[4]**

(d) Find a 95% confidence interval for β_1. **[4]**

End of Paper.