Queen Mary
University of London

MTH5120: Statistical Modelling I

Duration: 2 hours

Date and time: 31 May 2016, 10:00–12:00

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 awarded are shown next to the questions.

Calculators may be used in this examination, but any programming, graph plotting or algebraic facility may not be used. 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 provide.

Complete all rough workings 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.

Examiner(s): I. Goldsheid

Question 1. Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \;\; i = 1, 2, ..., n.$$

(a) List the standard assumptions about the random errors ε_i. **[2]**

(b) Write down the formula for the sum of squares of errors $S(\beta_0, \beta_1)$ and explain the method for obtaining the Least Squares Estimators of the unknown parameters $\beta_0,\ \beta_1$. Also, derive the normal equations for $\widehat{\beta}_0$ and $\widehat{\beta}_1$. **[14]**

(c) You are reminded that the Least Squares estimator for β_1 is given by

$$\widehat{\beta}_1 = \frac{S_{xY}}{S_{xx}}.$$

(i) Write down the formulae for S_{xY} and S_{xx}. **[3]**

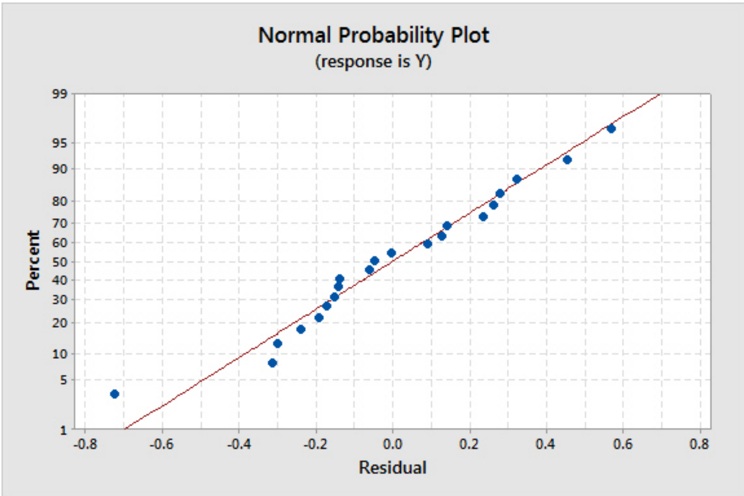
(ii) Prove that $S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})Y_i$. **[4]**

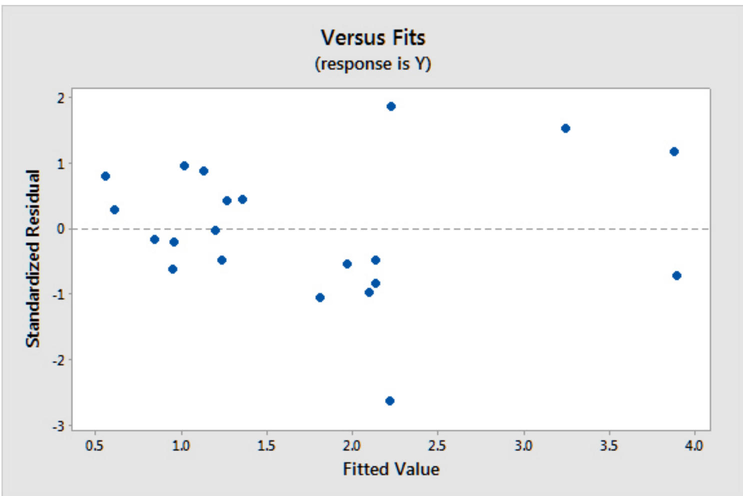
(iii) Now prove that $\widehat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i$ where $c_i = \frac{x_i - \bar{x}}{S_{xx}}$. **[4]**

(d) Prove that $\mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$.

Hint. Use the formula for $\widehat{\beta}_1$ stated in (c)(iii). **[5]**

Question 2. A production process is characterized by a response variable Y which is believed to depend on three explanatory variables X_1, X_2, X_3. Data were collected in order to check whether a multiple linear regression model would provide a good description of the dependence of Y on the explanatory variables. The set of data consists of $n = 21$ measurements.
The normal probability plot, a plot of residuals, and a summary analysis of the obtained data are presented below.





The summary obtained from MINITAB.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	18.9041	6.30136	59.90	0.000
X1	1	2.9623	2.96228	28.16	0.000
X2	1	1.3031	1.30308	12.39	0.003
X3	1	0.0997	0.09965	0.95	0.344
Error	17	1.7883	0.10519		
Lack-of-Fit	16	1.7833	0.11146	22.29	0.165
Pure Error	1	0.0050	0.00500		
Total	20	20.6924			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.324336	91.36%	89.83%	85.89%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.61	8.90	0.41	0.690	
X1	0.0716	0.0135	5.31	0.000	2.91
X2	0.1295	0.0368	3.52	0.003	2.57
X3	-0.152	0.156	-0.97	0.344	1.33

Regression Equation

Y = 3.61 +0.0716X1 +0.1295X2 -0.152X3

Fits and Diagnostics for Unusual Observations

Obs	Y	Fit	Resid	Std Resid	
21	1.500	2.224	-0.724	-2.64	R

R Large residual

(a) Give the definition of the linear model with 3 explanatory variables in terms of $Y_i,\ x_{1,i},\ x_{2,i}\ x_{3,i},\ \varepsilon_i$. Now, write down this model in terms of $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\varepsilon}$ and explain what is $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$. **[5]**

(b) Comment on whether the standard model assumptions are approximately satisfied. **[3]**

(c) Consider the following null hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{versus} \quad H_1 : \text{ at least one of } \beta_1,\ \beta_2,\ \beta_3 \text{ is not } 0.$$

Define what is SS_R and SS_E and state their distributions. Explain how these distributions are used for conducting the standard F-test for H_0. **[12]**

(d) Do the data presented above provide evidence for rejecting H_0? Explain your answer. **[3]**

(e) Define what is SS_{LoF} and SS_{PE}. What can you say about the Lack of Fit for this model? **[6]**

(f) In the summary analysis of the data presented above, consider the part concerning the coefficient of X_3. Do the corresponding p-values suggest that the explanatory variable X_3 is not important in the presence of X_1 and X_2? **[2]**

(g) What can you say about observation 21? **[1]**

Consider now a model for Y with just two explanatory variables, X_1 and X_2. Moreover, in this model observation 21 has been removed. Here is an extract from the summary analysis for the new model.

```
Analysis of Variance

Source          DF   Adj SS   Adj MS  F-Value  P-Value
Regression       2  19.5205  9.76026   150.16    0.000
  X1             1   3.7232  3.72319    57.28    0.000
  X2             1   0.4041  0.40407     6.22    0.023
Error           17   1.1050  0.06500
  Lack-of-Fit   10   0.4283  0.04283     0.44    0.882
  Pure Error     7   0.6767  0.09667
Total           19  20.6255


Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.254949  94.64%     94.01%      92.44%

Regression Equation

Y = -5.108 +0.0863X1 +0.0803X2
```

(h) What is the definition of R^2 and $R^2(adj)$? **[4]**

(i) Judging by the values of S^2, R^2, $R^2(adj)$, and $R^2(pred)$, state with reason which of the two models is better. **[3]**

Question 3. Consider a multiple linear regression model with p unknown regression parameters $\beta_0,\ \beta_1, ..., \beta_{p-1}$:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

(a) Suppose that $\boldsymbol{X}^T\boldsymbol{X}$ is an invertible matrix.

(i) State the formula for the least squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. **[2]**

(ii) Prove that $\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. **[6]**

(iii) State (do not prove) the formula for $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$. **[3]**

(iv) State the joint distribution of $\widehat{\boldsymbol{\beta}}$. **[2]**

(b) The model

$$\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$$

was fitted to a set of $n = 15$ observations and the following least squares estimates of the parameters were obtained:

$$\widehat{\beta}_0 = 10,\ \ \widehat{\beta}_1 = 12,\ \ \widehat{\beta}_2 = 15 \text{ and } s^2 = 2.$$

We also obtained

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} 1 & 0.25 & 0.20 \\ 0.25 & 2 & -0.22 \\ 0.20 & -0.22 & 0.5 \end{pmatrix}.$$

(i) Estimate $\mathrm{Var}(\widehat{\beta}_1)$ and $\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_2)$. **[2]**

(ii) Find the 95% confidence interval for β_1. State explicitly the number of degrees of freedom for the t-distribution which should be used in this particular case. **[6]**

(iii) Suppose now that $SS_T = 92$. Test at the 0.1% significance level (that is, $\alpha = 0.001$) the hypothesis that $\beta_1 = \beta_2 = 0$ against the hypothesis that at least one of these parameters is not 0.

Hint. Recall the relation between SS_E and s^2 and thus find SS_E and SS_R. **[8]**

End of Paper.