Data provided: Tables of critical values

MAS6003

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2019–2020**

Linear Modelling **1.5 hours**

This is an open book exam.
Answer ***all*** *questions.*
The submission deadline is 10 am (BST), twenty-four hours after it is released. Late submission will not be considered without extenuating circumstances. It is expected that you will be able to complete this exam in approximately one and a half hours and it is recommended that you submit the work within four and a half hours. You will not be penalised for taking longer, however.
Unless it is explicitly stated otherwise, it is intended that calculations are performed by hand (possibly with the aid of a calculator). To gain full marks, you will need to show your working. You will not get full marks if you simply write down output from a computer package.
By uploading your solutions you declare that your submission consists entirely of your own work, that any use of sources or tools other than material provided for this module is cited and acknowledged and that no unfair means have been used.

Total marks 60.

1 Data were collected on the times taken for three different brands of computer to start up. Thirty computers were tested in total. For each computer, the information available is its age in months, the brand of the computer (labelled as A, B or C) and the start up time in seconds. An extract, showing the format of the data, is given in the following table, and the complete data set is stored as `times.data` in `R`.

age	brand	time
23	A	48
49	B	73
11	C	32

Consider the model fitted in `R` by the code

```
time.lm <- lm(time ~ age+I(age^2)+factor(brand),data=times.data)
```

(a) Some of the output of the command `summary(time.lm)` appears below.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   31.669670   4.252524   7.447 8.45e-08 ***
age            0.424501   0.304430   1.394 0.175462
I(age^2)       0.011982   0.005596   2.141 0.042208 *
factor(brand)B -7.260091   1.791156  -4.053 0.000432 ***
factor(brand)C -7.907905   2.193142  -3.606 0.001353 **

Residual standard error: 4.025 on 25 degrees of freedom
Multiple R-squared:  0.8808,Adjusted R-squared:  0.8617
F-statistic: 46.17 on 4 and 25 DF,  p-value: 3.421e-11
```

(i) Explain why the number of degrees of freedom for the residual standard error is 25. ***(1 mark)***

(ii) Give an interpretation of the estimated coefficients which are shown in the `R` output as -7.260091 for `factor(brand)B` and -7.907905 for `factor(brand)C`. ***(4 marks)***

(iii) Using the parameter estimates in the `R` output, what is the predicted start up time for the computer in the table with an age of 49 months and of brand B? ***(3 marks)***

1 (continued)

(b) Some of the output of the command `anova(time.lm)` appears below.

```
Analysis of Variance Table

Response: time
              Df  Sum Sq Mean Sq  F value    Pr(>F)
age            1 2662.24 2662.24 164.3338 1.721e-12 ***
I(age^2)       1   16.68   16.68   1.0297 0.3199510
factor(brand)  2  313.04  156.52   9.6616 0.0007787 ***
Residuals     25  405.01   16.20
```

(i) Using this table, assess the evidence for the inclusion of brand in the model given that both age and age squared are already included. ***(3 marks)***

(ii) How would you use the `anova` command to assess the evidence for the inclusion of brand in the model if age has been included but age squared has not been? ***(3 marks)***

(iii) What can you say about choice of model based on this table? ***(1 mark)***

2 A model for non-negative data $\boldsymbol{y}$ is that each observation y_i is an observation from a distribution with probability density function

$$f(y_i;\theta_i) = \exp\left\{y_i\theta_i + \sqrt{-2\theta_i} - \frac{1}{2y_i} + \frac{1}{2}\log(2\pi y_i^3)\right\},$$

for $y_i > 0$, where θ_i is an unknown parameter for observation i satisfying $\theta_i < 0$.

By relating this model to the standard form for a Generalized Linear Model,

$$f(y_i;\theta_i,\phi) = \exp\left\{w_i\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi)\right\},$$

with $w_i = \phi = 1$:

(a) Find the mean of y_i in terms of θ_i. ***(5 marks)***

(b) Show that the canonical link function for this Generalized Linear Model is given by $g(\mu) = \dfrac{-1}{2\mu^2}$. ***(4 marks)***

(c) The link function relates the mean μ_i to a linear predictor η_i via $\eta_i = g(\mu_i)$. Suppose that the linear predictor is of the form $\eta_i = \beta_0 + \beta_1 x_i$, for a known vector of explanatory variables $\boldsymbol{x}$, and that we have estimated values $\hat{\beta}_0 = -5$ and $\hat{\beta}_1 = 1$ using the canonical link. By considering an observation with $x_i = 5.5$, describe a potential problem with using the canonical link for this model. ***(5 marks)***

3 A longitudinal study conducted in the USA by social scientists followed the salary of 85 individuals from 1968 to 1990. Interest lies in the effect of education, sex, and age on salary levels. The data are stored in the data frame `salaries`. The columns are

- `age`: age of the subject in 1968
- `educ`: years of education of the subject
- `sex`: sex of subject - a factor with levels 'F' or 'M'
- `income`: inflation adjusted annual income in dollars
- `year`: calendar year
- `person`: ID number for the subject

The following R code was used to fit a mixed effect model.

```
>
> library(lme4)
> head(salary)
  age educ sex income year person
1  31   12   M   8400   68      1
2  31   12   M   7420   69      1
3  31   12   M   7280   70      1
4  31   12   M   9660   71      1
5  31   12   M  10500   72      1
6  31   12   M  11200   73      1
>
> salary$cyear <- salary$year-78
> fit1 <- lmer(log(income) ~ sex +age+educ+(cyear|person),salary)
>
```

(a) For $i = 1, \ldots, 85$ and $j = 1, \ldots, 23$, let y_{ij} be the income of individual i in observation j, age_i be the age of individual i in 1968, educ_i the years of education for individual i, and year_{ij} the year observation j was made on individual i.

Write down the algebraic specification of the model that has been fitted to the data making sure you give the distribution of any random effects.

(4 marks)

3 (continued)

(b) The summary of the fitted model is given below. Use the R output to give the estimated values for each parameter in your answer to part (a).

```
>
> summary(fit1)
Linear mixed model fit by REML ['lmerMod']
Formula: log(income) ~ sex + age + educ + (cyear | person)
   Data: salary

REML criterion at convergence: 3888.6

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1955  -0.2204   0.0764   0.4028   2.9207

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 person   (Intercept) 0.30072  0.54838
          cyear       0.00764  0.08741  0.28
 Residual             0.46669  0.68315
Number of obs: 1661, groups:  person, 85

Fixed effects:
            Estimate Std. Error t value
(Intercept)  6.82311    0.54850  12.440
sexM         1.18538    0.12093   9.802
age          0.01163    0.01365   0.852
educ         0.10635    0.02165   4.911

Correlation of Fixed Effects:
     (Intr) sexM   age
sexM -0.104
age  -0.874 -0.026
educ -0.598  0.009  0.167
>
```

(4 marks)

(c) Explain, given the aim of the study, why the designation of explanatory variables as fixed and random is reasonable.

(2 marks)

(d) What R command would you use to fit a model in which the two random effect terms are uncorrelated? *(2 marks)*

3 (continued)

(e) Below is some R code and output. Describe the procedure performed and give the conclusions of the analysis.

```
>
> fit.reduced <-lmer(log(income) ~  sex+age+(cyear|person), salary, REML=F)
> fit.full <-lmer(log(income) ~  sex +age+educ+(cyear|person), salary, REML=F)
> obs.test.stat<- - 2*(logLik(fit.reduced)-logLik(fit.full))
>
> N<-1000
> boot.test.stats<-rep(0,N)
>
> for(i in 1:N){
+   new.y<-unlist(simulate(fit.full))
+   fm.reduced.new<-lmer(new.y~  sex+age+(cyear|person), salary, REML=F)
+   fm.full.new <-lmer(new.y~  sex +age+educ+(cyear|person), salary, REML=F)
+
+   boot.test.stats[i]<- -2*(logLik(fm.reduced.new)-logLik(fm.full.new))
+ }
>
> sum(boot.test.stats> obs.test.stat)
[1] 25
>
```

(3 marks)

4 (a) Consider the following dataset, where `NA` is used to denote missing observations:

X	Y
NA	5.2
2.0	1.6
2.3	NA
15.0	7.1
2.0	4.0

Interest lies in the expected value of Y, which we will denote by μ. Using the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ as an estimator, compute an estimate of μ using

(i) Complete-case analysis

(ii) Available-case analysis.

(2 marks)

(b) Stochastic imputation with random sampling has been used to create 3 imputed datasets. These are shown below with the imputed values in **bold**. The bottom row gives the sample mean and the sample variance of each column.

	Imputed 1		Imputed 2		Imputed 3	
	X	Y	X	Y	X	Y
	2.0	5.2	**2.0**	5.2	**15.0**	5.2
	2.0	1.6	2.0	1.6	2.0	1.6
	2.3	**5.2**	2.3	**4.0**	2.3	**7.1**
	15.0	7.1	15.0	7.1	15.0	7.1
	2.0	4.0	2.0	4.0	2.0	4.0
sample mean	4.62	4.62	4.66	4.38	7.26	5.00
sample variance	33.43	4.08	33.43	4.02	49.94	5.36

(i) Use these values to estimate the expected value $\mathbb{E}(\mu|Y_{obs})$.

(ii) Estimate the standard error in your estimate of μ.

Hint: The variance of the sample mean $\hat{\mu} = \frac{1}{n}\sum Y_i$ is

$$\mathbb{V}\text{ar}(\hat{\mu}) = \frac{\mathbb{V}\text{ar}(Y)}{n} \approx \frac{s_Y}{n}$$

where s_Y is the sample variance of Y.

(6 marks)

4 (continued)

(c) Consider the mixed effect model

$$y_{ij} = \alpha + \beta_i + b_j + e_{ij}$$

where $i = 1, 2$ and $j = 1, \ldots, J$. The e_{ij} are IID $N(0, \sigma^2)$ and the random effects have the distributions

$$b_j \sim N(0, \sigma_1^2)$$

with b_j independent of e_{ij} for all i, j. Sum to zero constraints are used for the fixed effects:

$$\beta_1 + \beta_2 = 0.$$

(i) The model can be written in matrix form as

$$\mathbf{Y} = X\boldsymbol{\beta} + Z\mathbf{b} + \boldsymbol{e}$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{21} \\ Y_{12} \\ Y_{22} \\ \vdots \\ Y_{1J} \\ Y_{2J} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_J \end{pmatrix}, \quad \text{and } \boldsymbol{e} = \begin{pmatrix} e_{11} \\ e_{12} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{J2} \end{pmatrix}.$$

Give the matrices X and Z, and the vector $\boldsymbol{\beta}$.

(ii) Consider the following estimators of α and β_1:

$$\hat{\alpha} = \bar{Y} = \frac{1}{2J} \sum_{i,j} Y_{ij}, \qquad \hat{\beta}_1 = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{2}$$

where $\bar{Y}_{i\cdot} = \dfrac{1}{J} \displaystyle\sum_{j=1}^{J} Y_{ij}$. Calculate the correlation between $\hat{\alpha}$ and $\hat{\beta}_1$, i.e., $\text{cor}(\hat{\alpha}, \hat{\beta}_1)$.

(7 marks)

End of Question Paper

Tables of Percentage Points (also known as Quantiles or Critical Values) for Three Standard Distributions

The tables contain values of quantiles q such that $P[X \leq q] = p$ for various probabilities p when X has the specified distribution (which may depend on particular degrees of freedom ν). In these tables, p has been expressed as a percentage rather than a decimal. The relevant R commands for generating the q are also shown. For the $N(0,1)$ distribution, the tabulated function is also known as the Φ^{-1} function.

STANDARD NORMAL DISTRIBUTION PERCENTAGE POINTS

`qnorm(`p`)` where p is percentage, e.g. for 95%, $p = 0.95$

	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
`qnorm`	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

CHI-SQUARED PERCENTAGE POINTS

`qchisq(`p`,` ν`)` where p is percentage, e.g. for 95%, $p = 0.95$

ν	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.708	0.936	1.323	1.642	2.354	2.706	3.841	5.024	6.635	7.879	10.828
2	1.833	2.197	2.773	3.219	4.159	4.605	5.991	7.378	9.210	10.597	13.816
3	2.946	3.405	4.108	4.642	5.739	6.251	7.815	9.348	11.345	12.838	16.266
4	4.045	4.579	5.385	5.989	7.214	7.779	9.488	11.143	13.277	14.860	18.467
5	5.132	5.730	6.626	7.289	8.625	9.236	11.070	12.833	15.086	16.750	20.515
6	6.211	6.867	7.841	8.558	9.992	10.645	12.592	14.449	16.812	18.548	22.458
7	7.283	7.992	9.037	9.803	11.326	12.017	14.067	16.013	18.475	20.278	24.322
8	8.351	9.107	10.219	11.030	12.636	13.362	15.507	17.535	20.090	21.955	26.125
9	9.414	10.215	11.389	12.242	13.926	14.684	16.919	19.023	21.666	23.589	27.877
10	10.473	11.317	12.549	13.442	15.198	15.987	18.307	20.483	23.209	25.188	29.588

STUDENT'S t PERCENTAGE POINTS

`qt(`p, ν`)` where p is percentage, e.g. for 95%, $p = 0.95$

ν	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090