Data provided: Tables of critical values

The University Of Sheffield.

MAS367

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2019–20**

Linear and Generalized Linear Models **2 hours**

Attempt all the questions. The allocation of marks is shown in brackets.

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

1 Data were collected on the times taken for thirty computers of three brands to start up. 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) Describe a suitable design matrix X for this linear model. You should illustrate the description by showing the rows of X corresponding to the three data points in the above table. ***(6 marks)***

(b) 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`. ***(3 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? ***(2 marks)***

1 (continued)

(c) 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(make)  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. ***(2 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? ***(2 marks)***

(iii) What choice of model does this table suggest? ***(1 mark)***

(d) Values of Akaike's Information Criterion (AIC) for a number of models are shown in the table below.

Model label	R code	AIC
I	`time ~ age+I(age^2)+factor(brand)`	88.08
II	`time ~ age+factor(brand)`	91.13
III	`time ~ age+I(age^2)`	101.26
IV	`time ~ age`	99.95
V	`time ~ 1`	143.88

(i) Based on this information, which model would you choose? ***(1 mark)***

(ii) Explain why the choices of model here and in (c)(iii) might be different. ***(2 marks)***

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 $\theta_i < 0$.

(a) 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 $w_i = \phi = 1$:

(i) Find the mean of y_i in terms of θ_i. ***(4 marks)***

(ii) Find the variance of y_i in terms of θ_i. ***(2 marks)***

(iii) What relationship would you expect to see between the mean and variance of the observations if this model were a good fit? ***(2 marks)***

(iv) Show that the canonical link function for this Generalized Linear Model is given by $g(\mu) = \dfrac{-1}{2\mu^2}$. ***(3 marks)***

(b) 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.

(i) Using this linear predictor and the canonical link, give the fitted mean of y_i for an observation with $x_i = 4.5$. ***(3 marks)***

(ii) By considering an observation with $x_i = 5.5$, describe a potential problem with using the canonical link for this model. ***(3 marks)***

(iii) Would the link function defined by $g(\mu) = \log(\mu)$ avoid the problem you identified in (ii)? Give a reason for your answer. ***(3 marks)***

3 Data have been collected on the number of fruit produced by 40 fruit trees of two different species in locations with different amounts of sunshine recorded. In the data frame `fruit.data`, for each tree the species is recorded as `species` and is either 0 or 1, the number of hours of sunshine in the growing season is recorded as `sun`, and the total number of fruit produced by the tree is recorded as `fruit`. An extract from the data frame, with data for three of the trees, is presented in the following table.

species	sun	fruit
0	232.2	51
0	227.3	36
1	185.0	35

A Poisson Generalized Linear Model with natural log link is fitted in `R` using

```
model.both=glm(fruit~factor(species)+sun,family=poisson,data=fruit.data)
```

(a) Some of the output from `summary(model.both)` appears below.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.443615   0.287578   8.497  < 2e-16 ***
factor(species)1 0.137780   0.050599   2.723  0.00647 **
sun               0.005883   0.001361   4.323 1.54e-05 ***

    Null deviance: 73.380  on 39  degrees of freedom
Residual deviance: 51.385  on 37  degrees of freedom
```

(i) Using the estimates from the output, what is the fitted mean of the number of fruit for the tree in the table of species 0 and with 232.2 hours of sunshine? What would it be if instead the tree were of species 1, with the same amount of sunshine? ***(4 marks)***

(ii) Calculate the Pearson residual for the tree in the table of species 0 and with 232.2 hours of sunshine. ***(3 marks)***

(iii) Based on the output, what can you say about the fit of the model fitted by `model.both`? You may use the fact that `1-pchisq(51.385,37)` in `R` gives `0.04706`. ***(3 marks)***

(iv) Suppose the fit of the model were judged to be inadequate. Give two ways in which you might find a better model. ***(2 marks)***

3 (continued)

(b) The output from `anova(model.both)` appears below

```
Analysis of Deviance Table

Model: poisson, link: log

Response: fruit

Terms added sequentially (first to last)

                 Df Deviance Resid. Df Resid. Dev
NULL                                39     73.380
factor(species)  1   3.1505        38     70.230
sun              1  18.8447        37     51.385
```

Using this output:

(i) Assess the evidence for the inclusion of species in the model compared with the null hypothesis where neither species nor sunshine is included. *(2 marks)*

(ii) Assess the evidence for the inclusion of sunshine in the model compared with the null hypothesis where species is included but sunshine is not. *(2 marks)*

(c) The tables above refer to a null model, with deviance 73.380. Explain the distribution of the observations under this model, any parameters of the model, and how you could estimate these parameters from the data. *(4 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