Data provided: Tables of critical values

MAS467

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2018–19**

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 A data set, `weight.data` in R, on 30 chicks of a particular species of bird at age 2 months gives `weight`, the weight of the chick in grams, `temp`, the average temperature in Celsius over the 2 month period, and `rain`, the total rainfall in millimetres over the 2 month period.

(a) A linear model was fitted to the data in R using the command

```
weight.lm <- lm(weight~temp+rain,data=weight.data)
```

giving output

```
Coefficients:
(Intercept)         temp         rain
  -1120.782      157.857       -4.698
```

The first observation had average temperature 11.7°C and rainfall 15.5mm, and the observed weight of the chick was 435g. Calculate the fitted value and the residual for this observation. ***(2 marks)***

(b) Following on from (a), the command

```
anova(weight.lm)
```

was entered into R, producing the following output:

```
Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value    Pr(>F)
temp       1 4788381 4788381  19.300 0.0001553 ***
rain       1 1003809 1003809   4.046 0.0543566 .
Residuals 27 6698742  248102
```

Based on this output,

(i) assess the evidence for the inclusion of temperature in the model against the null hypothesis where neither temperature nor rainfall is included; ***(2 marks)***

(ii) assess the evidence for the inclusion of rainfall in the model given that temperature is already included. ***(2 marks)***

(c) Following on from (a) and (b), the code

```
par(mfrow=c(2,2))
plot(weight.lm$resid,xlab="index",ylab="residuals",main="Index plot")
qqnorm(weight.lm$resid,main="QQ - plot")
hist(weight.lm$resid,xlab="Residuals",main="Histogram")
plot(weight.lm$fit,weight.lm$resid,xlab="Fitted values",
 ylab="Residuals",main="Residuals versus fitted values")
```

was entered into R, producing the plots in Figure 1. Using the residual plots, comment on whether there appear to be any problems with the assumptions made for a linear model. ***(4 marks)***

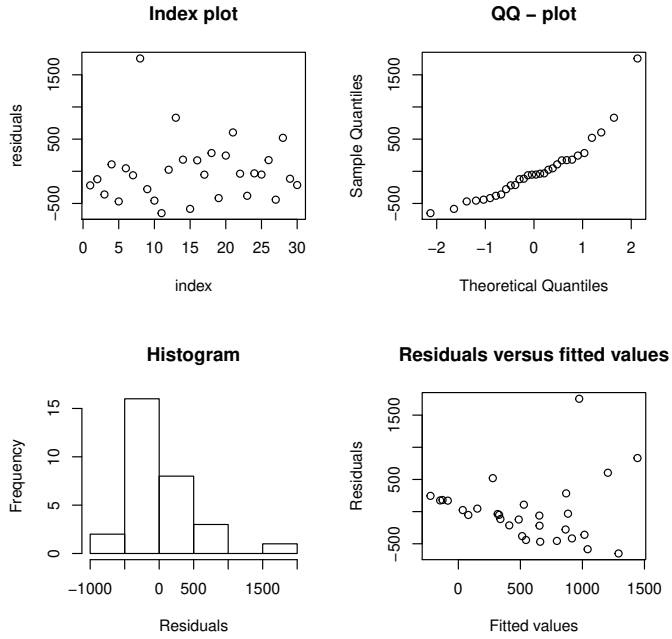

Figure 1: Residual plots for the bird weight data.

1 (continued)

(d) A new model was fitted in R using the command

```
weight.log.lm <- lm(log(weight)~temp+rain,data=weight.data)
```

(i) The output from the new model is

```
Call:
lm(formula = log(weight) ~ temp + rain, data = weight.data)

Coefficients:
(Intercept)         temp         rain
    1.81651      0.36827     -0.01282
```

For the first observation of the data given in (a), calculate the fitted value for the natural log of the weight in the new model and the corresponding residual. ***(3 marks)***

(ii) Residual plots for the new model are shown in Figure 2. Compare these plots with the ones for the original model. Does this model appear to be an improvement? You should give a reason for your answer. ***(4 marks)***

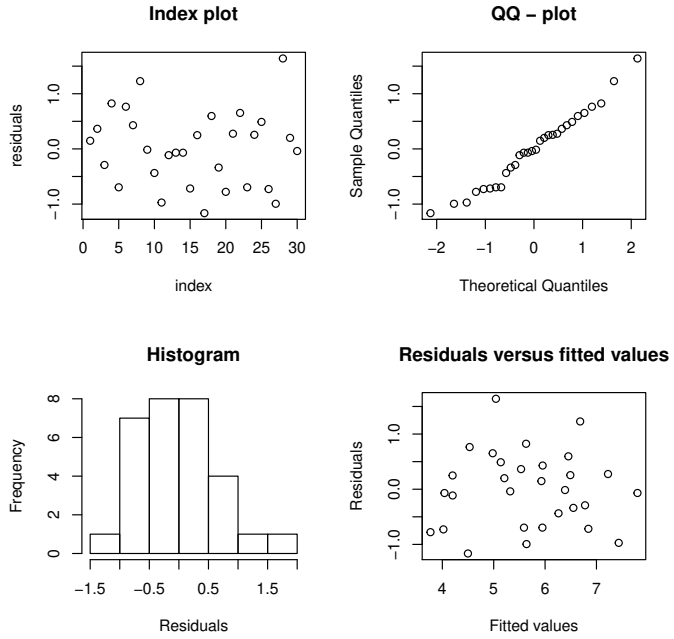

Figure 2: Residual plots for the transformed bird weight data.

1 (continued)

(iii) The code

```
anova(weight.log.lm)
```

gave the output

```
Analysis of Variance Table

Response: log(weight)
          Df  Sum Sq Mean Sq F value    Pr(>F)
temp       1 26.8556 26.8556  54.189 6.427e-08 ***
rain       1  7.4756  7.4756  15.084 0.0006013 ***
Residuals 27 13.3809  0.4956
```

Comparing this to the output from (b), how would your answers for (b)(i) and (b)(ii) change with this output? ***(3 marks)***

2 A model for data $\boldsymbol{y}$ is that each observation y_i is an observation from a Negative Binomial distribution with parameters 2 and p_i so that its probability mass function is

$$f(y_i; p_i) = (y_i - 1)p_i^2(1 - p_i)^{y_i - 2}.$$

(a) Show that the probability mass function can be put into 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 $w_i = \phi = 1$, and identify the functions b and c. You should explain the relationship between the quantities θ_i and p_i. ***(8 marks)***

(b) (i) Use the function b to show that the mean of y_i is $\dfrac{2}{1 - e^{\theta_i}}$. ***(2 marks)***

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

(iii) Hence also give the mean and variance of y_i in terms of p_i. ***(2 marks)***

(c) What would the canonical link function be for this Generalized Linear Model? ***(3 marks)***

(d) Suppose that the linear predictor is of the form $\eta_i = \beta_0 + \beta_1 x_i + \beta_2 z_i$, for known vectors of explanatory variables $\boldsymbol{x}$ and $\boldsymbol{z}$. Assuming that the canonical link is used, find the estimated value of p_i for an observation with $x_i = 1$ and $z_i = 2$ for a model with $\beta_0 = -8$ and $\beta_1 = \beta_2 = 1$. ***(3 marks)***

3 An agricultural experiment was carried out to investigate the success of a type of crop, testing the effects of amount of fertilizer and whether the seeds were fresh or not. For each combination of amount of fertilizer and seed freshness, an area of soil was prepared with that amount of fertilizer applied per square metre, 20 seeds were sown, and the number which produced fruit was recorded.

The data are in the following table, and were stored in `R` as `crop.data` with `fert` being the amount of fertilizer per square metre, `storage` indicating whether the seeds had been stored for a year (coded as `1`) or not (coded as `0`), `prop.fruit` indicating the proportion out of the 20 which produced fruit, and `n` being 20 for all observations.

	Amount of fertilizer applied (g m^{-2})					
	0	20	40	60	80	100
Fresh	12	9	16	17	18	20
Stored for one year	4	8	10	13	15	17

Two Generalized Linear Models with Binomial family and logit link were fitted using the following code.

```
model.fert <- glm(prop.fruit~fert,family=binomial,weights=n,data=crop.data)
model.both <- glm(prop.fruit~fert+factor(storage),
 family=binomial,weights=n,data=crop.data)
```

Some R commands and some of the output they produced follow.

```
> summary(model.fert)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.602470   0.245178  -2.457    0.014 *
fert         0.028487   0.004847   5.878 4.16e-09 ***

    Null deviance: 62.558  on 11  degrees of freedom
Residual deviance: 20.693  on 10  degrees of freedom
AIC: 59.331
> summary(model.both)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.06123    0.28869  -0.212 0.832029
fert              0.03046    0.00510   5.972 2.34e-09 ***
factor(storage)1 -1.16689    0.31888  -3.659 0.000253 ***

    Null deviance: 62.5585  on 11  degrees of freedom
Residual deviance:  6.4157  on  9  degrees of freedom
AIC: 47.054
```

(a) Using the output above, assess the evidence for

(i) whether the model with both storage and fertilizer is an improvement on the null model; ***(3 marks)***

3 (continued)

(ii) whether storage is needed in the model if the model already includes amount of fertilizer. ***(2 marks)***

(b) In the fitted model with both storage and fertilizer:

(i) Give the odds ratio for success for seeds which have been stored for a year compared with those which are fresh. ***(3 marks)***

(ii) Give an approximate 95% confidence interval for the odds ratio you calculated in (b)(i). ***(2 marks)***

(c) In the fitted model with both storage and fertilizer:

(i) What is the estimated probability that a seed will be successful if it is fresh and 50g per square metre of the fertilizer is used? ***(2 marks)***

(ii) What is the estimated probability that a seed will be successful if it has been stored for a year and 200g per square metre of the fertilizer is used? ***(2 marks)***

(iii) Do you think your answers to (c)(i) and (c)(ii) would give good predictions if these experiments were actually carried out? Give reasons for your answers. ***(4 marks)***

(d) If the model given by `model.both` had given a poor fit, what suggestions might you have to find a better model for these data? Give two suggestions. ***(2 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