i **Department of mathematical Sciences**

Examination paper for ST2304 Statistical modelling for biologists and biotechologists

Examination date: 9th May 2020

Examination time (from-to): 09:00 – 13:00

Permitted examination support material: All support material is allowed

Academic contact during examination: Bob O'Hara
Phone: 915 54 416

Technical support during examination: Orakel support services
Phone: 73 59 16 00

OTHER INFORMATION

- If a question is unclear/vague – make your own assumptions and specify in your answer the premises you have made. Only contact academic contact in case of errors or insufficiencies in the question set.
- **Saving:** Answers written in Inspera are automatically saved every 15 seconds. If you are working in another program remember to save your answer regularly.
- **Cheating/Plagiarism:** The exam is an individual, independent work. Examination aids are permitted. All submitted answers will be subject to plagiarism control. *Read more about cheating and plagiarism here.*
- **Notifications:** If there is a need to send a message to the candidates during the exam (e.g. if there is an error in the question set), this will be done by sending a notification in Inspera. A dialogue box will appear. You can re-read the notification by clicking the bell icon in the top right-hand corner of the screen. All candidates will also receive an SMS to ensure that nobody misses out on important information. Please keep your phone available during the exam.
- **Weighting:** Weighting of the questions is given for each question.

ABOUT SUBMISSION

- **Your answer will be submitted automatically when the examination time expires and the test closes**, if you have answered at least one question. This will happen even if you do not click "Submit and return to dashboard" on the last page of the question set. You can reopen and edit your answer as long as the test is open. If no questions are answered by the time the examination time expires, your answer will not be submitted.
- **Withdrawing from the exam:** If you wish to submit a blank test/withdraw from the exam, go to the menu in the top right-hand corner and click "Submit blank". This can not be undone, even if the test is still open.
- **Accessing your answer post-submission:** You will find your answer in Archive when the examination time has expired.

Lazy Mole Rats

Mole-rats are small eusocial mammals native to Africa. They have a high level of social organisation, where different individuals play different roles, called castes. There are two castes of workers: workers and lazy.

Researchers wanted to know if the lazy caste really was lazy, i.e. if it used less energy. They collected data for 35 mole-rats, observing caste, and measuring the amount of energy individuals used. They also measured body mass, which we will use later.

The data are all analysed with energy use and body mass log-transformed, using a natural log.

First, we can compare the two casts, to test if they use the same amount of energy. This was done by calculating the maximum likelihood estimate of the difference in means, assuming the data were normally distributed, with the same variance for all observations.

1 Which of these is the best description of a maximum likelihood estimate?

Select one alternative:

- The most likely value of the parameter
- The estimate most likely to to be true
- The estimate of the data that makes the parameter most likely
- X The value of the parameter that makes the data most likely

Maximum marks: 1

2 This is the likelihood curve for the difference in means between the worker and lazy castes.

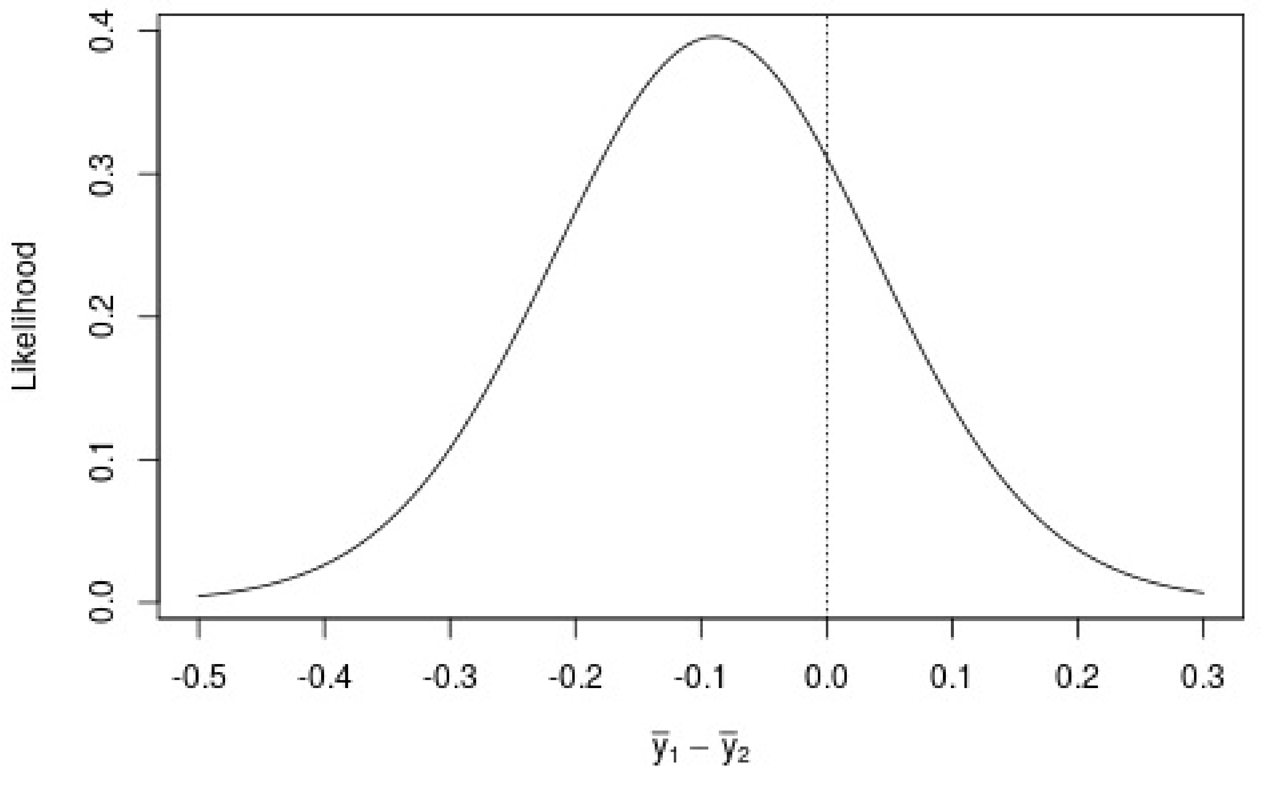

Figure 1: Likelihood for difference in log(energy use) between worker ($\bar{y}_1$) and lazy ($\bar{y}_2$) mole rat castes.
MIGHT WANT TO REMOVE ZERO LINE?

How to do this: look for where the highest value of the curve is.

What is (approximately) the maximum likelihood estimate? ______ (to no more than 2 decimal places).

-0.1 (about!)

Maximum marks: 1

3

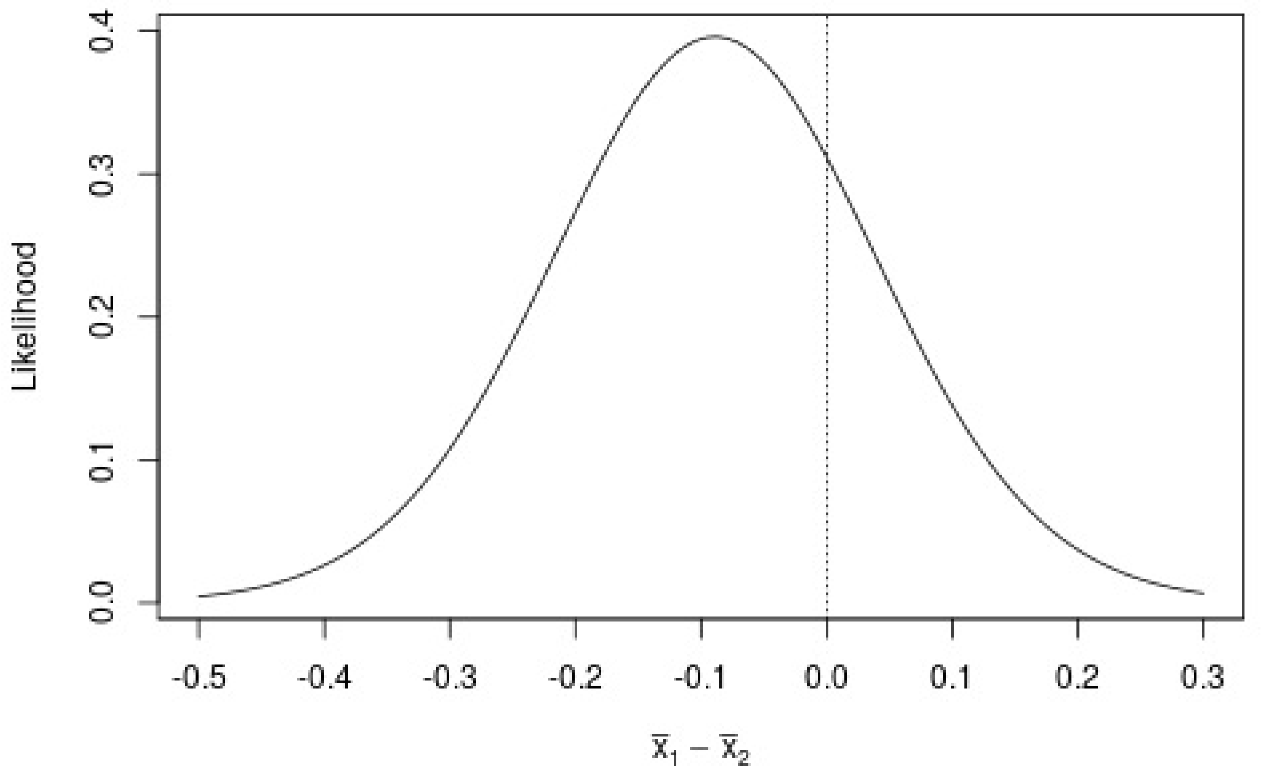

Figure 1: Likelihood for difference in log(energy use) between worker ($\bar{y}_1$) and lazy ($\bar{y}_2$) mole rat castes.
MIGHT WANT TO REMOVE ZERO LINE?

What is the probability that the difference would be great than 0?
[AAGH, 0.249 & 0.312 ARE TOO CLOSE TO EACH OTHER]
Select one alternative:

- 0.499
- 0.623
- 0.312
- X 0.249

How to do this: estimate what proportion of the area under the curve is above the x_1-x_2 line. The total area is 1, and the area above 0 is obviously less than 0.5. So it's either 0.312 or 0.249. But it's too difficult to decide between the two.

Note: this is a Bad Question: as noted, the numbers are too close together, and it mainly tests the wrong things (the ability to estimate proportions of areas).

Maximum marks: 1

4 From this analysis, would you conclude that there is a difference in energy use between the castes?
Fill in your answer here

I would conclude that there is no evidence for a difference in energy use between the castes. The likelihood curve suggests that we cannot tell what direction any difference is: the p-value is about 0.5

Maximum marks: 4

Lazy Mole Rats

The researchers also measured the body mass of the mole rats, because they expected that larger animals would use more energy, and this might obscure any relationship with caste. The data are plotted in Figure 2, with energy use and mass both transformed with a natural log.

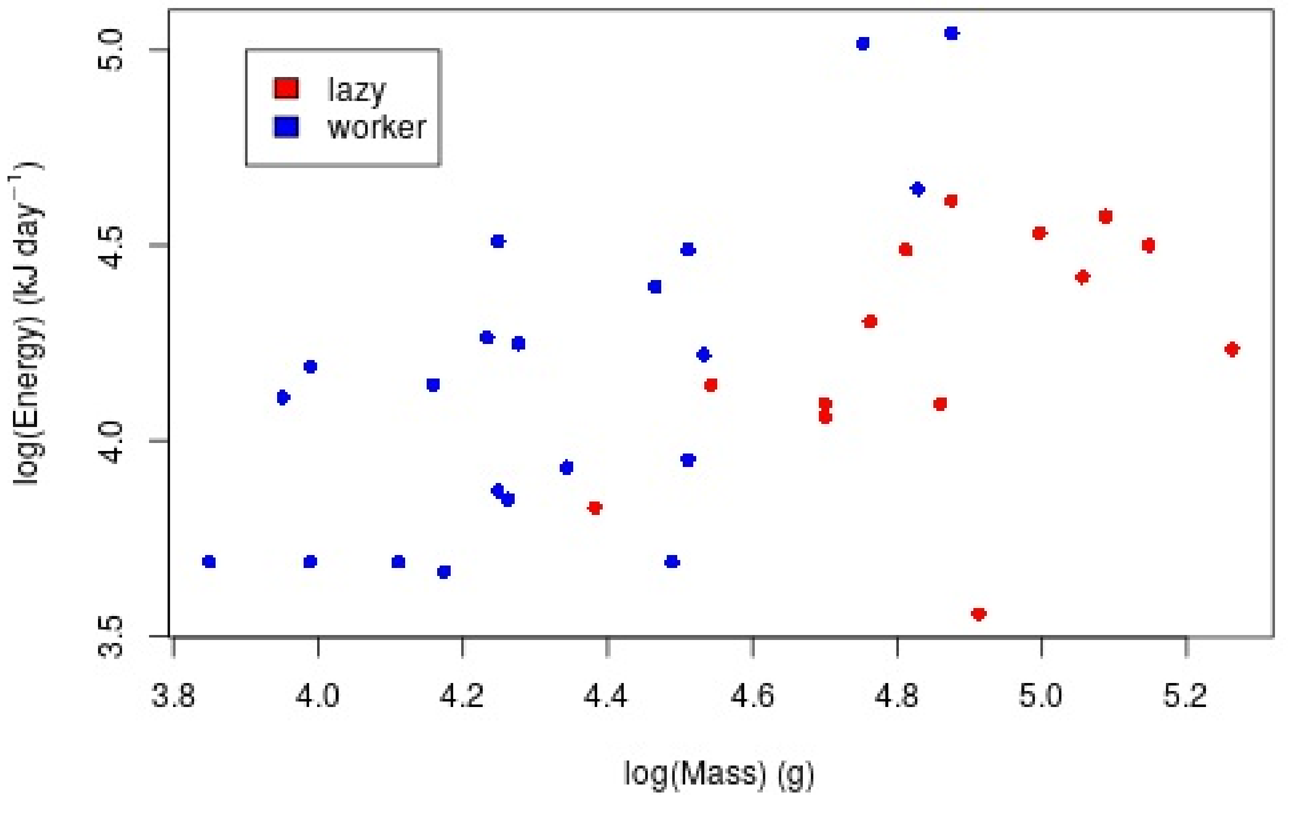

Figure 2: Energy use, body mass (both on natural log scale) and caste of naked mole rats.

A linear model was fitted with log(Mass) and caste as explanatory variables. it gave the following summary:

```
Call:
lm(formula = energy ~ mass + caste, data = Data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.73388 -0.19371  0.01317  0.17578  0.47673 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.09687    0.94230  -0.103   0.9188    
mass         0.89282    0.19303   4.625 5.89e-05 ***
casteworker  0.39334    0.14611   2.692   0.0112 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2966 on 32 degrees of freedom
Multiple R-squared:  0.409,	Adjusted R-squared:  0.3721 
F-statistic: 11.07 on 2 and 32 DF,  p-value: 0.0002213
```

5 What is the estimated effect of caste, to 2 decimal places? 0.39 .

Maximum marks: 1

6 What is the 95% confidence interval for the effect of caste, to 2 decimal places?

Lower value: 0.10 , upper value: 0.69

(hint: the 2.5% quantile for a t-distribution with 33 degrees of freedom is -2.03)

Maximum marks: 4

7 How much of the variation is explained by the model (as a percentage, to the nearest whole number)? 41 .

Maximum marks: 1

8 An analysis of variance was conducted, and gave the following readout.

```
Analysis of Variance Table
Response: energy
          Df  Sum Sq Mean Sq F value    Pr(>F)
mass       1 1.31061 1.31061 14.9013 0.0005178 ***
caste      1 0.63747 0.63747  7.2478 0.0111984 *
Residuals 32 2.81450 0.08795
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Would you consider the test to be exploratory, confirmatory or something else?

Select one alternative:

- Something else
- Exploratory
- X Confirmatory

Maximum marks: 1

9

An analysis of variance was conducted, and gave the following readout.

```
Analysis of Variance Table
Response: energy
          Df  Sum Sq Mean Sq F value    Pr(>F)
mass       1 1.31061 1.31061 14.9013 0.0005178 ***
caste      1 0.63747 0.63747  7.2478 0.0111984 *
Residuals 32 2.81450 0.08795
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Does it suggest a significant effect of caste? How do you come to that conclusion?

Fill in your answer here

Yes, it suggests a significant effect of caste. I came to that conclusion by looking at the p-value of 0.01, which is below the usual threshold of 0.05.

Words: 0

Maximum marks: 4

10 The formula used to fit the model for the effects of caste and body mass was

```
energy~mass+caste
```

What code would we use if we thought the effect of mass could differ between castes?

energy~mass*caste

(please don't use spaces!)

Maximum marks: 2

11 Based on the model of the data in Figure 2, and the ANOVA in Question 8, interpret the effect of caste on log(energy) use."

Fill in your answer here

Format B I U x_2 x^2 I_x Ω Σ

The ANOVA shows that there is an effect of both caste (p=1%) & body mass (p=0.05%). We can see from the figure that energy use increases with mass, but also that lazy ants tend to be bigger. Thus, although there is no effect of cast alone (Q4), once we take mass into account, we see that for ants of the same mass, lazy ants use less energy, i.e. they are indeed lazy.

Words: 0

Maximum marks: 4

Some ancient Greek researchers were surveying Mediterranean islands to work out how many people there were, so they could work out how much ancient Greek wine they could sell to them. In addition to counting how many people there were on each island, hey also recorded the following:

- the area of the island (in km^2)
- the maximum height (in m)
- whether it is visited by the gods (if you know about the Greek gods, you can understand why this might not be desirable)
- the proportion of the land that is pasture

They want to see whether any of these variables explain the current population size, so that they don't have to count the number of people on every island.

Is this exploratory or confirmatory?

Select one alternative:

X Exploratory

Confirmatory

Maximum marks: 1

All combinations of models were fitted, and different statistics calculated to compare the models. The summaries are below: the models with R^2 > 10% are not presented.

Table 1: AIC, BIC and R^2 for models with R^2 > 10%.

Model	AIC	BIC	R^2 (%)
Area	-598.1	-589.2	52.2
Area + PropPasture	-610.7	-598.9	56.9
Area + MaximumHeight	-596.2	-584.3	52.2
Area + PropPasture + MaximumHeight	-608.8	-594.0	56.9
Area + Gods	-599.3	-587.4	53.3
Area + PropPasture + Gods	-612.2	-597.4	57.9
Area + MaximumHeight + Gods	-597.3	-582.5	53.3
Area + PropPasture + MaximumHeight + Gods	-610.2	-592.4	57.9

13 Which statistic is best to use to compare these models, if we want to predict the population size
Select one alternative:

- X AIC
- ○ R^2
- ○ BIC

Maximum marks: 1

14 Which is the best model?
Select one alternative:

- ○ Area + MaximumHeight + Gods
- ○ Area + PropPasture + MaximumHeight + Gods
- X Area + PropPasture + Gods
- ○ Area + Gods
- ○ Area + PropPasture + MaximumHeight
- ○ Area
- ○ Area + MaximumHeight
- ○ Area + PropPasture

Maximum marks: 2

15 Why do you think this is the best model? Which statistics did you use to come to this conclusion and why?

Fill in your answer here

Format B I U x_2 x^2 I_x Ω Σ

I decided that AIC was the statistic to use, and this model has the lowest AIC. Some others, e.g. 'Area + PropPasture' and 'Area + PropPasture + MaximumHeight + Gods' had AICs that were 1.5 and 2 higher, so could also be considered: 'Area + PropPasture' is also simpler, so we could also try that model (although ignoring the Gods does not seem wise).

Words: 0

Maximum marks: 4

The residuals for the best model are plotted below.

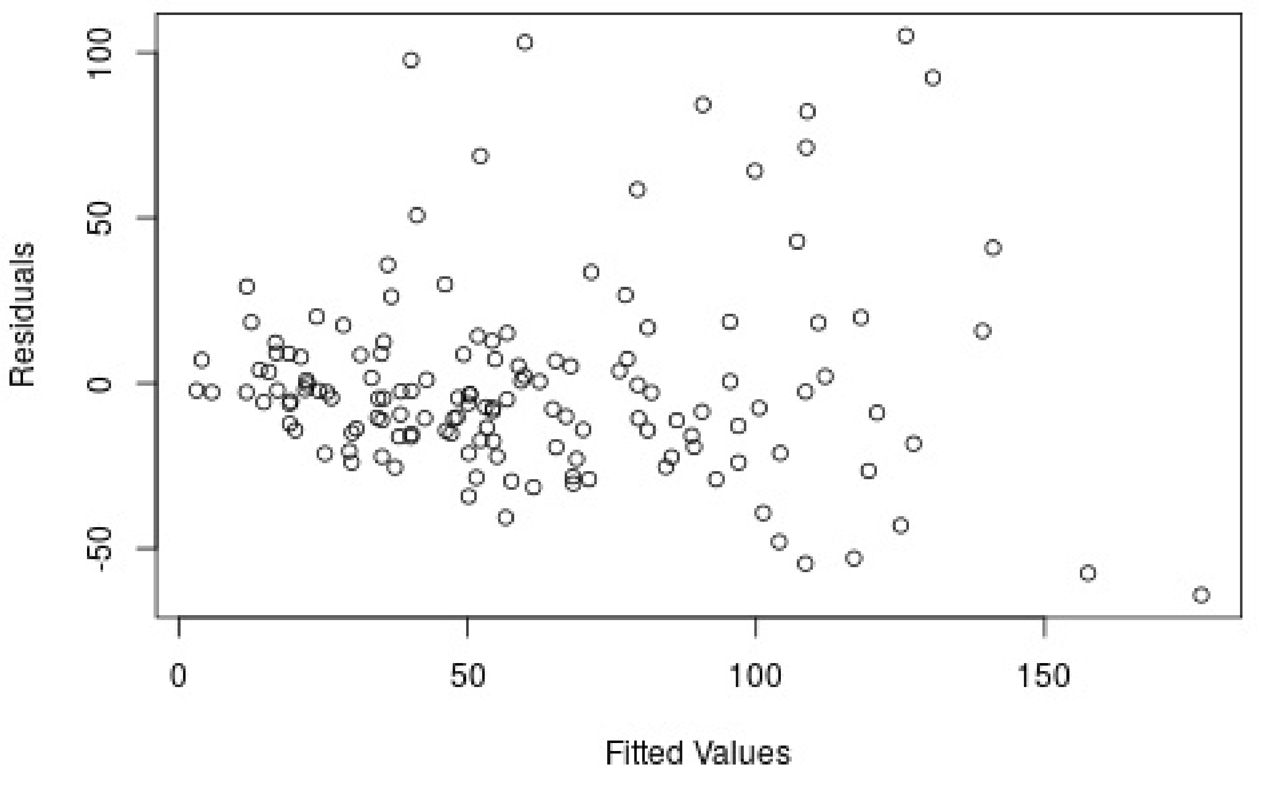

Figure 3: Residual Plot for model of population size on mythical Greek islands

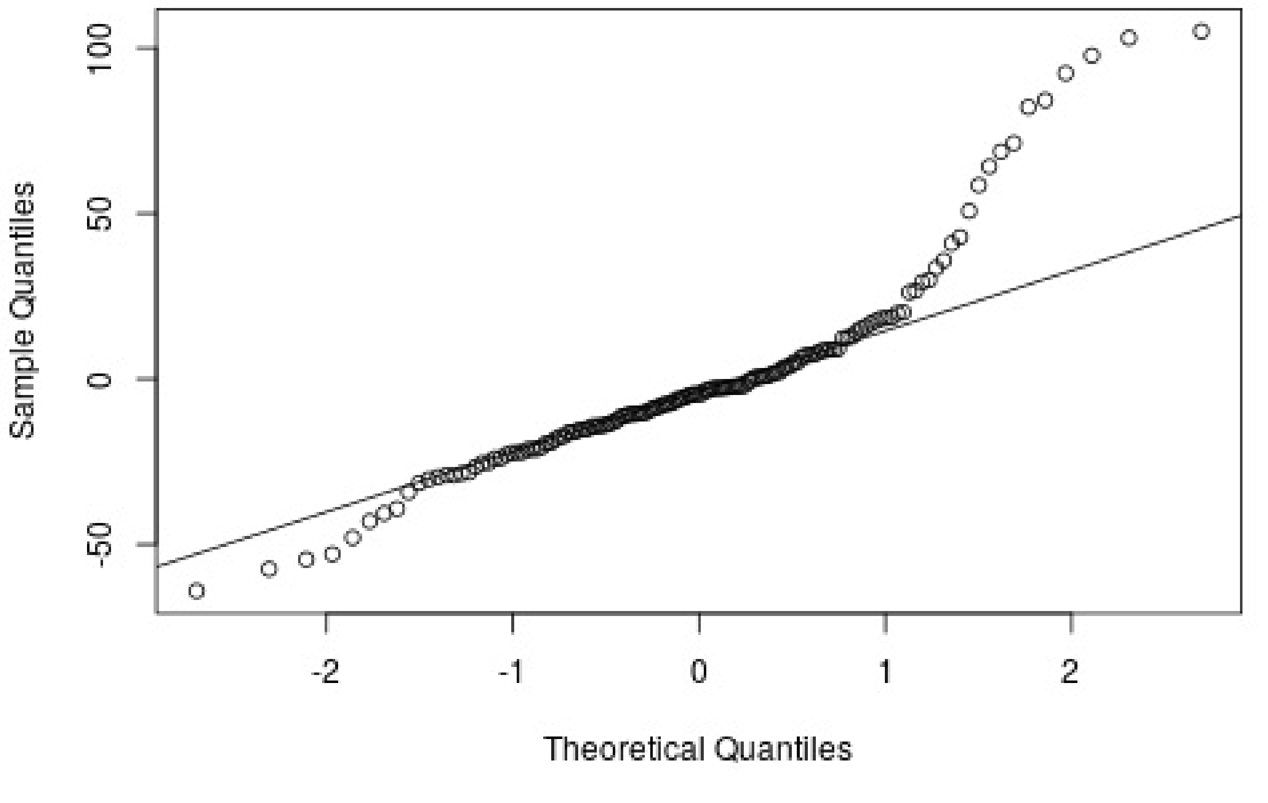

Figure 4: Normal probability plot for residuals from a model of population size on mythical Greek islands

16 Which model assumptions can you use the plot in Figure 3 to check for?
Select one or more alternatives:

- X The relationship is linear
- Normally distributed error (residuals)
- X There are no outliers
- X Error has equal variance along line

Maximum marks: 2

17 Which model assumptions can you use the plot in Figure 4 to check for?
Select one or more alternatives:

- X Normally distributed error (residuals)
- The relationship is linear
- X There are no outliers
- Error has equal variance along line

Maximum marks: 2

18 Based on Figures 3 and 4, do you think the assumptions of this model are met? If not, why?

Format B I U x_2 x^2 I_x Ω Σ

No the assumptions are not met: it's horrible. the main problem, from Fig. 3, seems to be the heteroscedasticity, and possibly the skewness & thick tails we can see in Fig. 4.

The thick tails may partly be the result of the heteroscedasticity.

Words: 0

Maximum marks: 4

19 Based on your answer to question 18, how could you try to improve the model?

Fill in your answer here

Sacrificing a goat to the right gods may be a first step. After that, a Box-Cox transformation is the first thing to try: something like a square root or log transformation might work. We would have to check the residuals after transformation, though.

The alternatives (not covered in the course!) would be to use weighted least squares: weighted so that the larger variance data points have a smaller weight. Or we could try a different GLM, if we can find one that weights the means and variance correctly. (NB: just saying "use a GLM" is not really good enough: you would have to show you understood enough about finding the correct sort of GLM. As we haven't covered that in the course, I wouldn't expect it as an answer)

Words: 0

Maximum marks: 4

20

What other plots could be used in addition to those in Fig 3 and Fig 4 to check if the model assumptions are met? What assumption would these plots check?

Fill in your answer here

We could also plot the residuals against covariates. These would tell us if the effect of a covariate was not linear. We could also plot the leverage, to see if there are any influential observations. Whilst this does not check an assumption, it is useful to see if the model is robust to weird data points, i.e. if the estimates are mainly affected by one or two data points.

Words: 0

Maximum marks: 4

The researchers looking at the island found a problem: many of the islands also had cyclops on them. Cyclops are one-eyed giants who are also shepherds, and don't like humans. So the researchers also recorded whether cyclops occurred on the islands.

The researchers fitted a model assuming a binomial distribution with one trial for whether cyclops were on each island. They used island area and the proportion of pasture as predictors. They got the following summary from the model:

```
Call:
glm(formula = Cyclops ~ log(Area) + PropPasture, family = binomial("cloglog"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2138  -0.9333   0.5043   0.8363   1.7790

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8333     0.4630  -3.960 7.51e-05 ***
log(Area)     0.7072     0.1517   4.662 3.14e-06 ***
PropPasture   1.6608     0.7840   2.118   0.0341 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 183.84  on 142  degrees of freedom
Residual deviance: 149.45  on 140  degrees of freedom
AIC: 155.45

Number of Fisher Scoring iterations: 6
```

Which link function was used in this binomial model?

Select one alternative:

- X cloglog
- logit
- probit
- log
- identity

Maximum marks: 1

A wandering scholar, on a 10 year trip home from a conference, is considering landing on one of two islands. From his maps he has the following statistics for the islands:

- Island 1: area of 5 km^2, 50% pasture
- island 2: area of 10 km^2, 20% pasture

He wants to know which island to land on to have the smallest chance of meeting a cyclops.

Note that in the analysis the pasture was used as a proportion, and also the natural log of the area was used.

On link scale: eta = -1.8333 + log(5)*0.7072 + 0.5*1.6608 = 0.14
log(5) and 0.5 come from the data for Island 1. And we get -1.8333, 0.7072 & 1.6608 from the summary output (there may be some variation depending on when you round)

22 What is the prediction on the link scale for there being a cyclops on Island 1? **0.14** (answer to 2 decimal places).

Probability: P = 1 - exp(exp(-eta)) = 1 - exp(exp(-0.14))

What is the corresponding probability of cyclops being on Island 1? **0.68** (answer to 2 decimal places)

(reminder in the analysis the pasture was used as a proportion, and also the natural log of the area was used)

On link scale: eta = -1.8333 + log(10)*0.7072 + 0.2*1.6608 = 0.13
log(5) and 0.5 come from the data for Island 1. And we get -1.8333, 0.7072 & 1.6608 from the summary output (there may be some variation depending on when you round)

Maximum marks: 4

23 What is the prediction on the link scale for there being a cyclops on Island 2? **0.13** (answer to 2 decimal places)

Probability: P = 1 - exp(exp(-eta)) = 1 - exp(exp(-0.13))

What is the corresponding probability? **0.67**

(answer to 2 decimal places)

(reminder in the analysis the pasture was used as a proportion, and also the natural log of the area was used)

Maximum marks: 4

24 Which island do you recommend he land on and why?
(to help you: the standard errors for the predicted probabilities of cyclops being on each island were estimated to be about 0.06)

Fill in your answer here

There is a slightly smaller predicted probability of a cyclops on Island 2, it is 0.01 smaller, but that is well within the sampling variation (one standard error is 0.06). So Island 2 might be slightly better, but it does not matter too much. Beyond this we should consider more than the numbers, e.g. one explorer preferred island 1, because there is more pasture, so they would be able to see the cyclops from further away, and either run or throw sheep at them.

Maximum marks: 4