Cover page

Department of Mathematical Sciences

Examination paper for ST2304 Statistical Modelling for Biologists and Biotechnologists

Examination date: 21stMay 2024

Examination time (from-to): 15:00 – 19:00

Permitted examination support material: C. Hand-written (either digitally witten or by hand, not typed) support material is allowed on both sides of one yellow A4 exam paper. A calculator of the following types is allowed: Casio FX-82CW, Casio FC100 V2, Casio fx-82ES PLUS og Casio fx-82EX, Citizen SR-270X, Citizen SR-270X College, Hewlett Packard HP30S.

Academic contact during examination: Bert van der Veen
Phone: 91343314

Academic contact present at the exam location: NO

OTHER INFORMATION

Get an overview of the question set before you start answering the questions.

Read the questions carefully and make your own assumptions. If a question is unclear/vague, make your own assumptions and specify them in your answer. The academic person is only contacted in case of errors or insufficiencies in the question set. Address an invigilator if you suspect errors or insufficiencies. Write down the question in advance.

No hand drawings: This exam does not include hand drawings. If you receive hand drawing sheets, this is by mistake. **You will not be able to submit the sheets, and they will not be graded.**

Weighting: Maximum marks per question are indicated next to the question.

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.

Withdrawing from the exam: If you become ill or wish to submit a blank test/withdraw from the exam for another reason, go to the menu in the top right-hand corner and click "Submit blank". This cannot be undone, even if the test is still open.

Access to your answers: After the exam, you can find your answers in the archive in Inspera. Be aware that it may take a working day until any hand-written material is available in the archive.

During the course we have learned a lot about sampling data, fitting models to data, and how parameters of models are estimated. These first few questions relate to some of these most foundational concepts in the course.

1 Q1: Likelihood

Which of these best describes a likelihood? (1pt)

Select one alternative:

○ Probability of the data p(y)

○ Probability of a model p(model)

● Probability of the data due to a model p(y;model)

Maximum marks: 1

2 Q2: MLE

Describe the mathematical process for finding the estimators of parameters based on a likelihood (4 pts)

Fill in your answer here

The maximum is at the place on the likelihood with zero curvature.
We want to find the estimator that corresponds to that location.
To do that, we calculate the first derivative with respect to our parameter, which represents the equation for the slope of the likelihood.
We set this equal to zero, and isolate the parameter.

Maximum marks: 4

3 Q3: Assumptions

Which of these are assumptions of linear regression? (5pt)
Select one or more alternatives:

☑ Independence of errors

☑ Heteroscedasticity

□ Dependence of errors

☑ Linearity

☑ Normality

□ Variance changes with mean

□ Significant p-values

☑ Lack of perfect multicollinearity

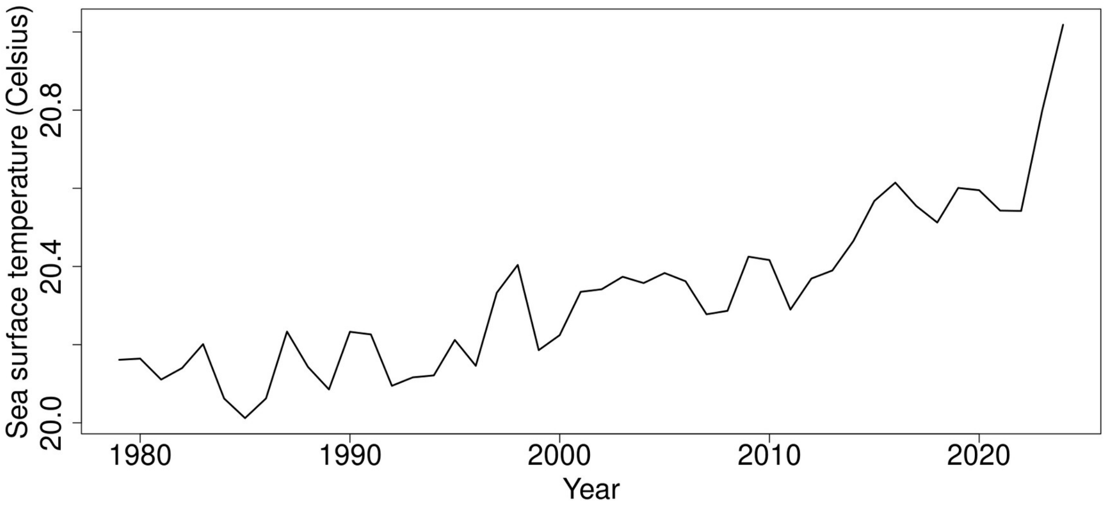
□ Normality of covariates

□ Low AIC

□ Low Cook's distance

□ Homoscedasticity

Maximum marks: 2



2024 is an El Niño year. As a consequence, sea water scientists have observed higher sea surface temperatures than usual. The Copernicus project provides climate data, so you decide to download the data and fit a model to that data, in order to see for yourself how much sea surface temperatures have changed between 1979 and now. The data is collected each month, but you calculate the yearly mean temperature to simplify the analysis. For 2024 there is only 3 months of data available.

The dataset includes:

1. sst: the (averaged) yearly temperatures since 1979 in the top layer of the ocean (numerical)
2. year (numerical)

SST lm

You fit a linear model with year as explanatory variable and receive the following summary output:

```
Call:
lm(formula = sst ~ year, data = seadatayear)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.16739 -0.08076 -0.00467  0.05973  0.38473 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.855363   2.272799  -3.016  0.00424 ** 
year         0.013582   0.001136  11.961 2.02e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1022 on 44 degrees of freedom
Multiple R-squared:  0.7648,	Adjusted R-squared:  0.7594 
F-statistic: 143.1 on 1 and 44 DF,  p-value: 2.022e-15
```

and associated confidence intervals:

```
(Intercept) -11.43588819 -2.27483770
year          0.01129305  0.01587004
```

4 Q4: LM interpretation

What does the model say about changes in sea surface temperature over time? (3pt)

Fill in your answer here

1) sea surface temperature increases over time
2) it increases by 0.014 degrees celsius per year
3) in year 0 the sea surface temperature was nearly -7 degrees celsius

Maximum marks: 3

5 Q5: Technically correct CI

What is a technically correct interpretation of a confidence interval? (2pt)

Select one alternative:

○ 95% range of the true parameter

○ The range that has 95% chance to contain the true parameter

● The range that will contain the true parameter 95% of the time

○ The range of the true parameter with 95% confidence

Maximum marks: 2

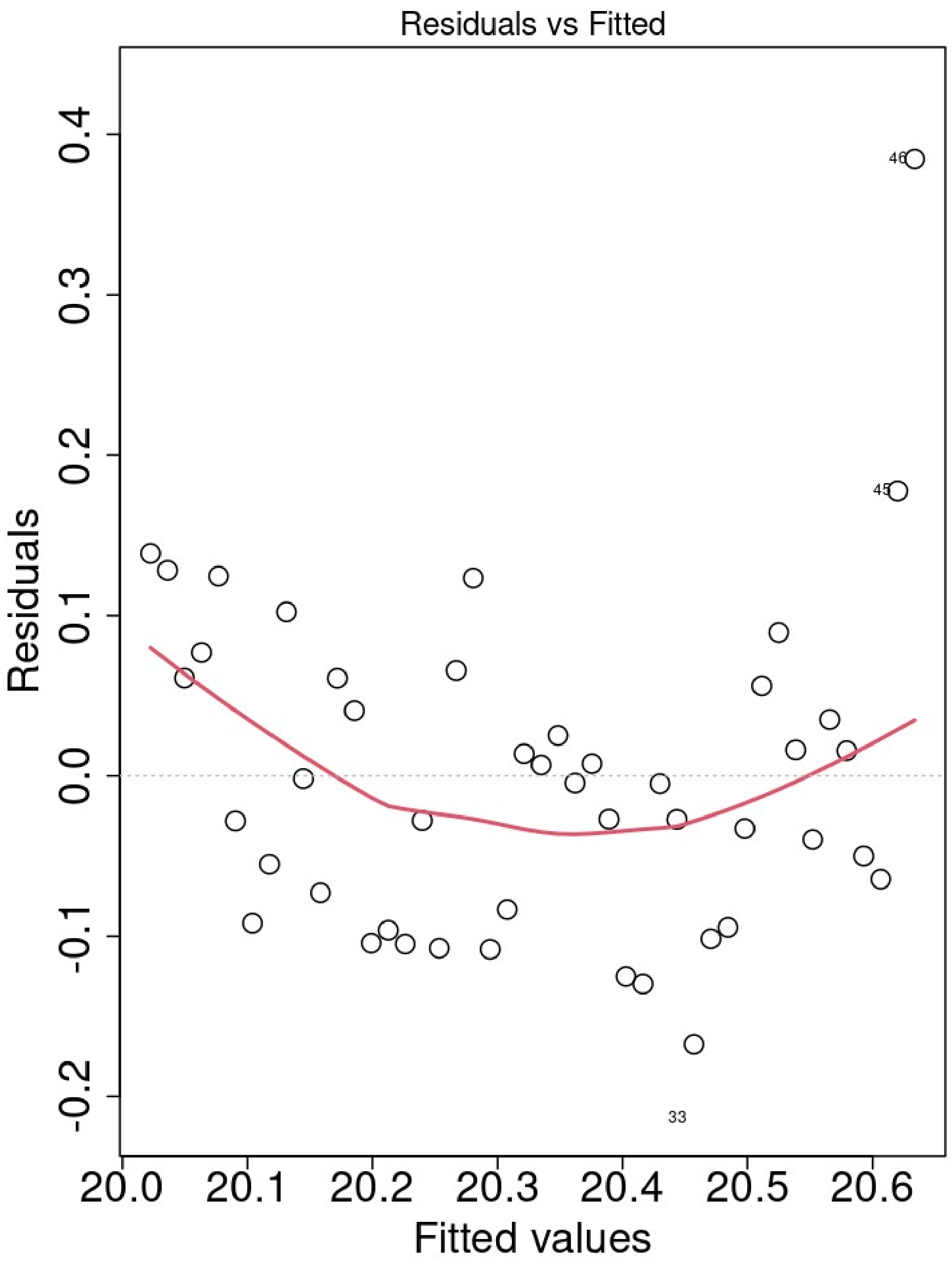
6 Q6: CI importance

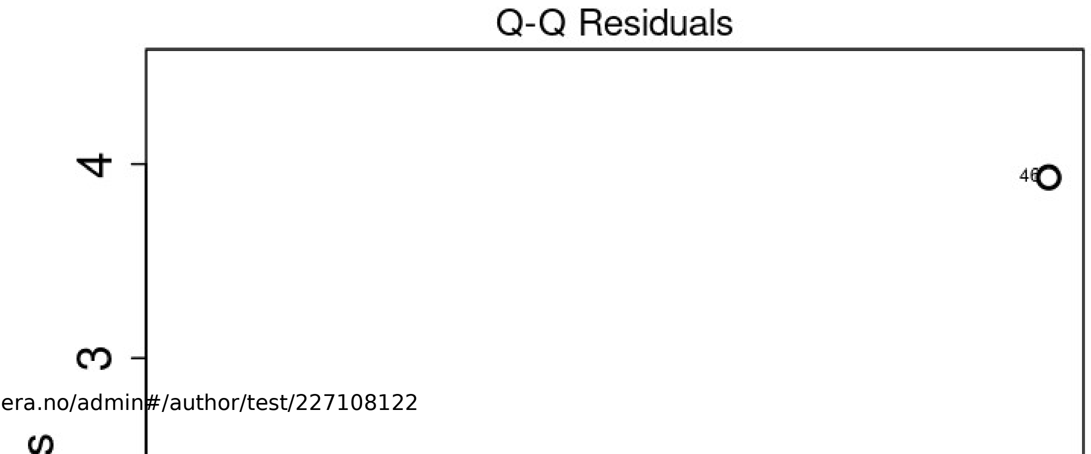
Why is it important to consider a CI along with a parameter estimate? (2pt)

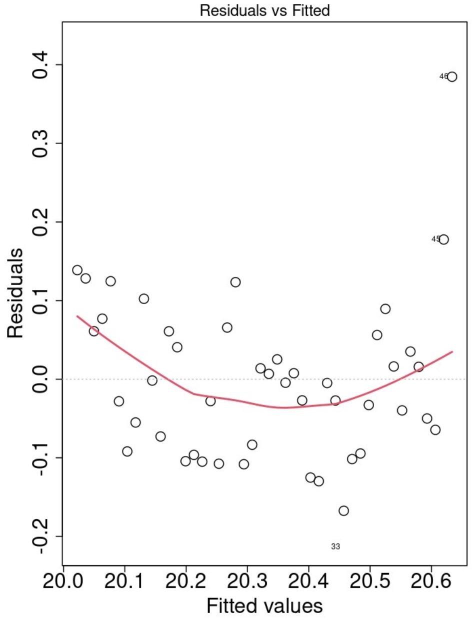
Fill in your answer here

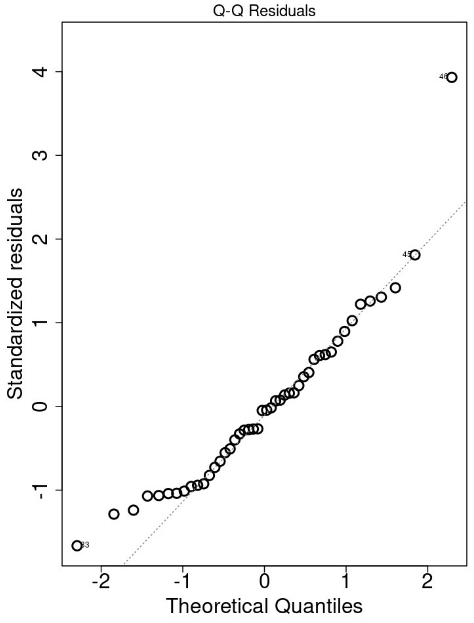
The parameter estimate can change if we collected data again.
We want to quantify how much it might change, so that we can get an impression of how robust our answer is to sampling variation.

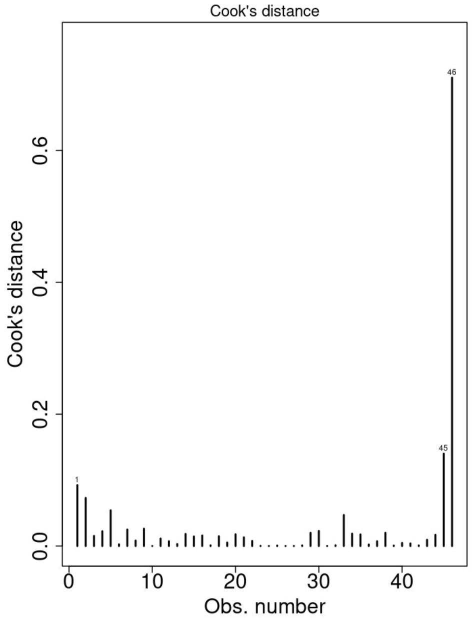
Maximum marks: 2


Residuals vs Fitted
Residuals
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
46
45
33
20.0
20.1
20.2
20.3
20.4
20.5
20.6
Fitted values


Q-Q Residuals
4
3
46
s


Residuals vs Fitted
Residuals
Fitted values
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
20.0
20.1
20.2
20.3
20.4
20.5
20.6
46
45
33


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
4
3
2
1
0
-1
-2
-1
0
1
2
46
45
33


Cook's distance
Cook's distance
Obs. number
0.6
0.4
0.2
0.0
0
10
20
30
40
46
45
1

You plot the model and receive three residual plots.

7 Q7: QQ-plot

Which assumption do we check with a QQ-plot?
Select one alternative:

○ Lack of perfect multicollinearity

○ Independence of errors

● Normality

○ Constant variance

Maximum marks: 1

8 Q8: Cook

Which assumption do you check with the Cook's distance? (1pt)

Fill in your No outliers here.

Maximum marks: 1

9 Q9: 46

Do the statistics suggest that observation 46 is dubious?

Select one alternative:

● Yes

○ No

Maximum marks: 1

10 Q10: Motivation 46

Motivate your answer: why (not) should 46 be removed? (2pt)

Fill in your answer here

Point 46 represents the average sst for 2024, which is lower than most other years due to only a few months of data being included. It is not comparable tot he other observations and should be removed.

Maximum marks: 2

11 Q11: Describe quadratic

Describe why the residual plots indicates that a model with a quadratic curve might fit better. (2pt)

Fill in your answer here

The residuals vs. fitted plot exhibits curvature. This represents systematic departure from the model, and if we include this trend the model will improve.

Maximum marks: 2

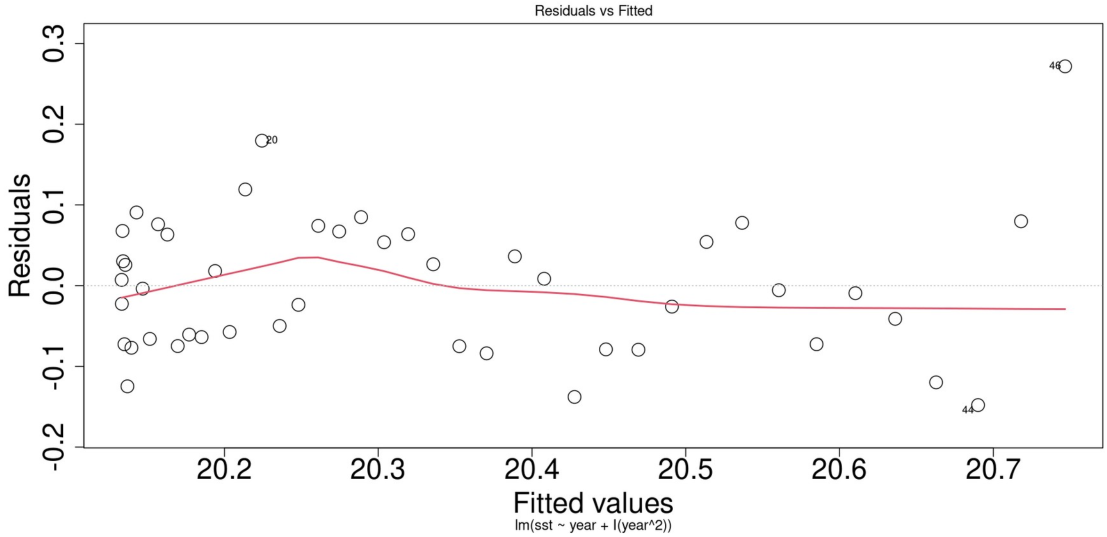
12 Q12: Other things

Are there things other than removing the 2024 data, or including a quadratic term, that you could do? (2pt)

Fill in your answer here

Instead of using averages, we could analyse the monthly sea surface temperatures so that the 2024 data can be included.

Maximum marks: 2



You decide to fit a new model with a quadratic function of year, and removing the data for 2024. The new model gives the following summary:

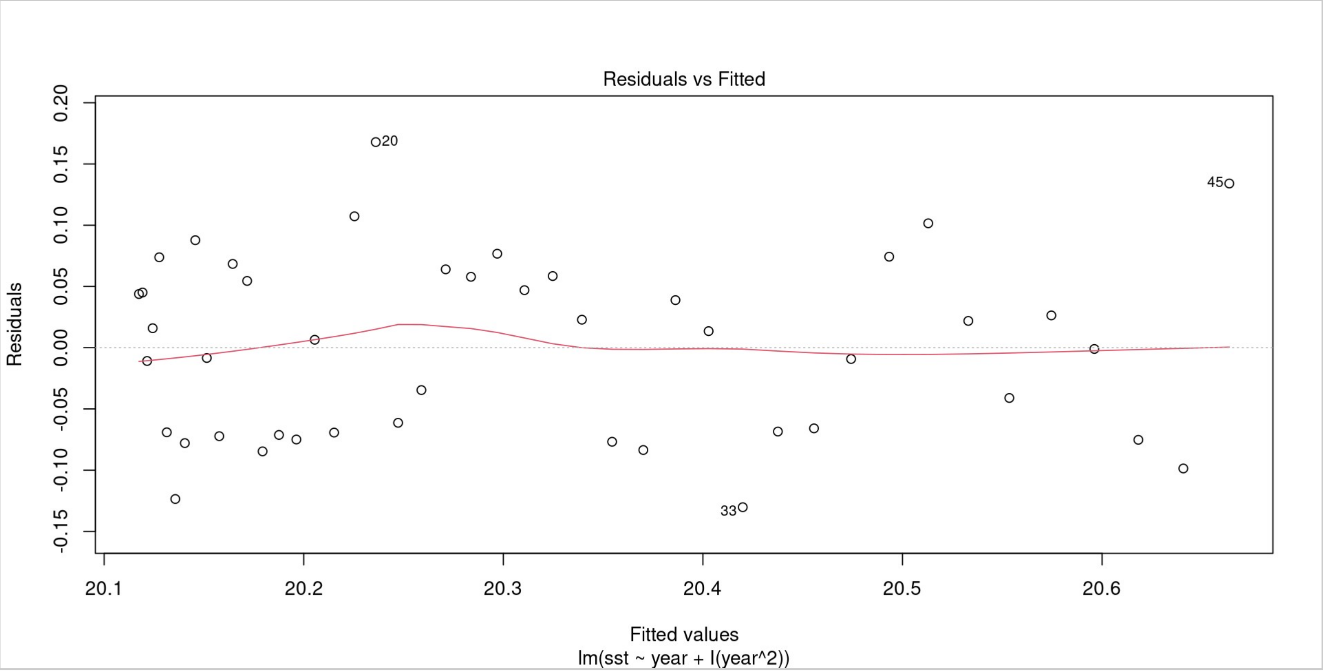
```
Call:
lm(formula = sst ~ year + I(year^2), data = seadatayear[seadatayear$year !=
     2024, ])

Residuals:
       Min        1Q    Median        3Q       Max
-0.130237 -0.069316  0.006456  0.057883  0.167873

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.837e+02  2.968e+02   3.314  0.00190 **
year        -9.753e-01  2.967e-01  -3.287  0.00205 **
I(year^2)    2.468e-04  7.413e-05   3.329  0.00182 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07497 on 42 degrees of freedom
Multiple R-squared:  0.8392,    Adjusted R-squared:  0.8316
F-statistic: 109.6 on 2 and 42 DF,  p-value: < 2.2e-16
```

the following confidence intervals:

```
(Intercept)  3.846491e+02  1.582691e+03
year        -1.574072e+00 -3.765939e-01
I(year^2)    9.720443e-05  3.964232e-04
```

and the attached residual vs. fitted plot.

Quadratic summary



You decide to fit a new model with a quadratic function of year, and removing the data for 2024. The new model gives the following summary:

```
Call:
lm(formula = sst ~ year + I(year^2), data = seadatayear[seadatayear$year !=
     2024, ])

Residuals:
       Min        1Q    Median        3Q       Max 
-0.130237 -0.069316  0.006456  0.057883  0.167873 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)   
(Intercept)  9.837e+02  2.968e+02   3.314  0.00190 **
year        -9.753e-01  2.967e-01  -3.287  0.00205 **
I(year^2)    2.468e-04  7.413e-05   3.329  0.00182 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07497 on 42 degrees of freedom
Multiple R-squared:  0.8392,	Adjusted R-squared:  0.8316 
F-statistic: 109.6 on 2 and 42 DF,  p-value: < 2.2e-16
```

the following confidence intervals:

```
(Intercept)  3.846491e+02  1.582691e+03
year        -1.574072e+00 -3.765939e-01
I(year^2)    9.720443e-05  3.964232e-04
```

as well as the attached residual vs. fitted plot.

13 Q13: Intercept

The intercept is very large, why is this? (3pt)

Fill in your answer here

The quadratic curve in the model opens upward, so the model predicts sea surface temperature to Indefinitely increase to the past, as well as to the future.

Maximum marks: 3

14 Q14: Improvement

How can you improve the model to make the intercept more meaningful? (3pt)

Fill in your answer here

We can center the year variable, or subtract a particular year, so that the intercept is the value of temperature when year is zero, or alternatively, equal to the year that we subtracted.

Maximum marks: 3

15 Q15: Better fit

What information in the summary indicates that this model might be a better fit than the model in question 4? (2pt)

Select up to two alternatives:

□ Smaller t value

☑ Lower residual variance

□ More significant p-values

☑ Higher R^2

□ Higher degrees of freedom

□ Larger median residual

Maximum marks: 2

16 Q16: Residual plot

Are you satisfied with the residuals vs. fitted plot? (2pt)

Fill in your answer here

Yes, there are no clear patterns. The points look randomly distributed.

Maximum marks: 2

17 Q17: Prediction

What does the quadratic model predict the sea water temperature to be in 2050? (up to 2 decimal places, 4pts)

21.47

Maximum marks: 4

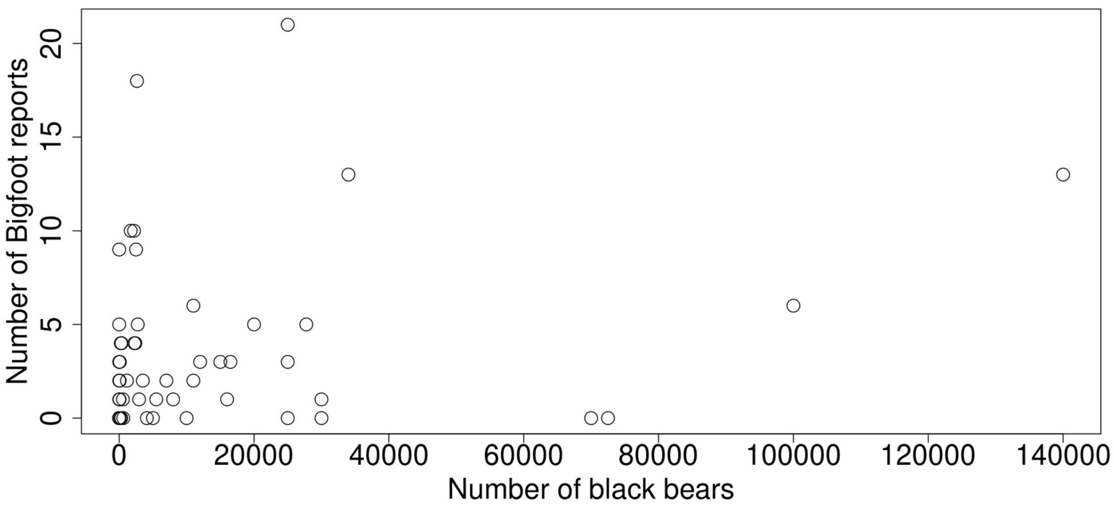
18 Q18: Data properties

What particular property of data collected through time might violate linear regression assumptions? (4pt)

Fill in your answer here

We have data of multiple years, but the sea surface temperature of one year is depends on the sea surface temperature of the previous years. So the independence of errors assumption is violated.

Maximum marks: 4



A “groundbreaking” new study hypothesizes that reports of Bigfoot (a.k.a. Sasquatch) actually represent reports of black bear (*Ursus americanus*) in the United States of America and Canada. Fortunately, you know better, Bigfoot is real! You decide to redo the statistical analysis of the scientists and prove them wrong.

The dataset includes:

1. Number of bigfoot reports (count, response variable)
2. Black bear population (count of individuals)
3. Human population (number of individuals in a given state or province)
4. Forest area in kilometers squared (numerical)
5. Country (US or Canada, categorical)

Since the number of Sasquatch is a count, you decide to fit a Generalized Linear Model with a negative-binomial distribution and log-link function. A negative-binomial distribution is a distribution for counts, like the Poisson distribution, but unlike the Poisson distribution it does not assume that the variance is equal to the mean, so that it accommodates issues with overdispersion that can arise in Poisson regression.

19 Q19: lm vs glm

What is one of the main differences between a lm and a glm? (1pt)
Select one alternative:

● GLM can have a different distribution

○ Nothing, GLM is a linear model ("General Linear Model")

○ GLMs include more parameters

○ The name: Generalized Linear Model

Maximum marks: 1

20 Q20: Overdispersion

What does it mean if there is “overdispersion” in count data? (1pt)
Select one alternative:

○ The data is skewed

○ There is not enough data to answer the research question

● Variance is larger than assumed

○ An inappropriate link function was used

○ The constant variance assumption is not met

Maximum marks: 1

21 Q21: Assumptions

What additional assumptions does a GLM make compared to an LM? (4pt)

Select one or more alternatives:

□ No multicollinearity

□ Constant dispersion parameter

■ Correct distribution

□ Just fewer assumptions

□ Normality

□ Dependence of errors

■ The appropriate variance function

■ The right link function

□ No outliers

□ Constant variance

Maximum marks: 4

22

Q22: link

Which of these is a link function? (1pt)

Select one alternative:

○ Cosine

○ Floor

● Logit

○ Absolute

Maximum marks: 1

23

Q24: Data property

What property of the data as, shown in the figure, is an indicator that the number of black bears cannot fully explain the number of Sasquatch reports? (2pt)

Fill in your answer here

There are some states without black bears that do have Sasquatch reports, and some states that have many Sasquatch reports but few black bears.

Maximum marks: 2

You decide to fit a model that represents the number of Sasquatch reports as a function of the black bear population, country, and their interaction. The summary output of the model is:

```
Call:
glm.nb(formula = Number_of_Bigfoot_Reports ~ Black_Bear_Population *
     Country, data = data, init.theta = 0.8965479658, link = log)

Coefficients:
                                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -1.524e+00  1.081e+00  -1.410   0.1584
Black_Bear_Population                  2.917e-05  1.185e-05   2.461   0.0139 *
CountryUS                              2.656e+00  1.099e+00   2.416   0.0157 *
Black_Bear_Population:CountryUS -1.055e-05  1.735e-05  -0.608   0.5431
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.8965) family taken to be 1)

     Null deviance: 67.749  on 52  degrees of freedom
Residual deviance: 58.040  on 49  degrees of freedom
AIC: 253.76
```

with the following confidence intervals:

```
(Intercept)                     -3.973774e+00 4.839182e-01
Black_Bear_Population            8.642128e-06 5.571107e-05
CountryUS                        6.038026e-01 5.137361e+00
Black_Bear_Population:CountryUS -4.991301e-05 2.945083e-05
```

24 Q24: Bigfoot US

What is the expected number of Sasquatch reports in the US where there are no black bears? (up to 1 decimal, 2pt)

3.1

Maximum marks: 2

25 Q25: Bigfoot and bears

What can you conclude from the model about the relationship between Sasquatch reports and the number of black bears? (2pt)

Fill in your answer here

1) There are more Sasquatch reports where there are more black bears
2) This effect is (more or less) the same in both countries, but slightly smaller effect for the US

Maximum marks: 2

26 Q26: CI practical

How do you use a confidence interval in practice, to draw conclusions from your model? (2pt)

Fill in your answer here

1) A wider confidence interval means that the estimate might change more when we collect more data so our inference might not be robust
2) If the CI crosses zero, we are not sure of the direction of the effect most of the times

Maximum marks: 10

You decide to fit another model that additionally includes the human population and forest area explanatory variables, because you expect more Sasquatch reports in places where there are more humans to report them, and you expect more Sasquatch reports where there is more forest (because then there are probably more black bears). You receive the following model summary:

```
Call:
glm.nb(formula = Number_of_Bigfoot_Reports ~ 0 + Black_Bear_Population *
     Country + Forest_Area_Square_km + Human_Population, data = data,
     init.theta = 1.352133715, link = log)

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
Black_Bear_Population            4.555e-05  1.503e-05   3.031  0.00244 **
CountryCA                       -7.900e-01  1.125e+00  -0.702  0.48262
CountryUS                        5.839e-01  2.268e-01   2.575  0.01003 *
Forest_Area_Square_km           -5.200e-06  2.755e-06  -1.888  0.05903 .
Human_Population                 1.065e-07  2.233e-08   4.769 1.86e-06 ***
Black_Bear_Population:CountryUS -1.497e-05  1.639e-05  -0.914  0.36096
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.3521) family taken to be 1)

     Null deviance: 182.406  on 53  degrees of freedom
Residual deviance:  56.536  on 47  degrees of freedom
AIC: 242.98
```

with the confidence intervals:

```
                                         2.5 %       97.5 %
Black_Bear_Population            1.743181e-05 8.097061e-05
CountryCA                       -3.256686e+00 1.239971e+00
CountryUS                        1.819119e-02 1.132216e+00
Forest_Area_Square_km           -1.120960e-05 2.273355e-07
Human_Population                 4.913036e-08 1.752065e-07
Black_Bear_Population:CountryUS -5.113339e-05 1.989511e-05
```

You notice that this model has an AIC value of 242.98 and the first model of 253.76.

27 Q27: AIC

What is AIC used for? (2pt)

Select one or more alternatives:

☑ It is used to evaluate a candidate set of models

□ It is used to find the model closest to the truth

□ It is used to find a good model

□ It is used to test a hypothesis

☑ It it used to find the model that predicts best

Maximum marks: 2

28 Q28: Best AIC

From the two models ranked by AIC which is the best? (1pt)

Select one alternative:

○ The model without Forest area and human population

○ Neither

● The model with Forest area and human population

Maximum marks: 1

29 Q29: AIC again

Does AIC give you a model that fits the data well? (1pt)

Select one alternative:

○ We need BIC to find a model that fits the data well

○ Yes, the model with lowest AIC is always a model that has a good fit to the data

● No, the model with lowest AIC might fit the data poorly

Maximum marks: 1

30 Q30: AIC vs BIC

What is the difference between AIC and BIC? (2pt)

Fill in your answer here

1) AIC finds the best predictive model, BIC the model closest to the truth
2) The penalty is different. 2*k for AIC and log(n)*k for BIC

Maximum marks: 2

31 Q31: hypothesis

What hypothesis does the interaction term between black bear population and country represent? (2pt)

Select one alternative:

● The effect of black bear population on the amount of Sasquatch reports is different in the US and Canada

○ The US has a different probability for Sasquatch and black bear occurrence than Canada

○ Black bear population affects the country that Sasquatch reports are found in

○ The number of black bears is different in the US and Canada

Maximum marks: 2

32 Q32: Summary interpretation

What is your conclusion from the second model's summary? (6pt)

Fill in your answer here

1) The number of Sasquatch reports increase with black bear population (as before)
2) Still a minor difference in black bear effect for US from Canada
3) Sasquatch reports decrease with the amount of forest, CI crosses zero
4) On average the US has more Sasquatch reports than CA, CI crosses zero

Maximum marks: 6

33 Q33: ranger encounter

A forest ranger offers to collect more data for an improved analysis. What should he collect? (4pt)

Fill in your answer here

There are many sensible answers, but I would ask him to collect also data on other animal species. I doubt that black bear is the only animal accused of being a humanoid. Perhaps also information on the reporter. How many reports belong to a select few individuals..?

Maximum marks: 4