Prof. Dr. Altwicker

Fall Semester 2023

Solutions
Empirical Legal Studies

January 5, 2024

Duration: 120 minutes

- Please check at receipt of the exam the number of question sheets. The examination contains 9 pages and 15 questions.
- Students are allowed to use the Formula Sheet attached to this examination.

Notes on marking

- When marking the exam each question is weighted separately. Points are distributed to the individual questions as follows:

Questions 1-10	30 points	30%
Question 11	5 points	5%
Question 12	20 points	20%
Question 13	15 points	15%
Question 14	15 points	15%
Question 15	15 points	15%
Total	100 points	100 %

- A total of 10 additional points may be awarded (max. 5 for question 12 and max. 5 for question 15).

Notes on multiple-choice questions

- Select the only correct or most appropriate statement out of five statements.
- Each correctly answered multiple-choice question gives 3 points, each incorrectly answered multiple-choice question gives 0 points.

Notes on completing the separate multiple-choice solution-sheet

- We strongly recommend that you transfer the solutions to the separate solution sheet shortly before the end of the exam (see below). This is advisable, because possibly an answer to a question gives you reason to return to a previously answered question and to answer that question differently.

Notes concerning the separate multiple-choice solution sheet

- Answers to the multiple-choice questions **must be made on the separate multiple choice solution sheet according to the guidelines.** Only this solution sheet will be revised**.**

We wish you a lot of success!

Part I: Multiple Choice Questions

1. Select the only correct or most appropriate statement from the following five.

a) Empirical Legal Studies have no relevance for doctrinal legal scholarship (i.e. for the doctrine of the positive law), since Empirical Legal Studies concerns “is”-statements and the dogmatic legal scholarship relates to “ought”-statements about the law.

b) Empirical Legal Studies can be useful when it comes to finding natural groups or patterns in legal data.

c) Empirical scientific data cannot contribute to the interpretation of complex, undefined legal concepts.

d) Due to the quality of legal data, Empirical Legal Studies cannot contribute to the prediction of legal decisions.

e) In the U.S., Empirical Legal Studies are primarily conducted by the courts themselves when they have to clarify complex issues (e.g. the question of whether certain video games encourage violence among young people).

2. Select the only correct or most appropriate statement from the following five.

a) Since many phenomena (such as social status) are not observable, proxies (such as occupation, income, level of education) are often used as variables by which the phenomenon of interest is approximated.

b) In Empirical Legal Studies, the definition of variables always follows the positive law.

c) Even partially automated data collection is not possible in Empirical Legal Studies due to the complexity of the data basis.

d) Hypotheses are formulated as propositions containing words like “should” or “could”.

e) In Empirical Legal Studies, hypotheses are developed on the basis of doctrinal legal scholarship (i.e. for the doctrine of the positive law); other sources cannot be considered because otherwise there would be no reference to the law.

3. Select the only correct or most applicable statement from the following five.

a) The variable "Complainant is a natural person: yes / no" is ordinal scaled.

b) The variable "Nationality of the complainant (country codes 1 to 49)" is a quantitative discrete variable.

c) The variable "Temperature in Celsius" is a quantitative interval-scaled variable.

d) The variable "Weather: 1 = sun, 2 = rain, 3 = cloudy" is a qualitative ordinal-scaled variable.

e) The variable "Compensation (in euros)" is a qualitative variable.

4. Select the only correct or most applicable statement from the following five.

a) The statistical concept of probability and the legal concept of probability are identical, as this is the only way that the law can take into account findings from statistics.

b) In criminal proceedings, the judge must follow the statistical concept of probability when assessing the evidence; otherwise their assessment would count as arbitrary.

c) The Federal Supreme Court requires that the court must be convinced that the facts of the case have been established "with a probability bordering on certainty"; this corresponds to a probability of approx. 90%.

d) In criminal proceedings and liability proceedings, the frequentist concept of probability in statistics is becoming increasingly important.

e) In the system of risk-oriented enforcement of sanctions, a recidivism prognosis for prison inmates is calculated on the basis of data. The aim of calculating the probability of recidivism is to prevent future offenses.

5. Select the only correct or most applicable statement from the following five.

a) The idea of a Bayesian probability can be described by the following question: How likely are our prior assumptions or our prior knowledge (priors) given a certain course of events?

b) Bayes' probabilities are irrelevant for criminal proceedings.

c) Bayes' probability can be used for the following problem: What is the probability that a driver was drunk if the test result was positive?

d) Bayes' probability results from the number of favorable events divided by the number of possible events.

e) A Bayesian probability P(B|A) > 1 is possible in exceptional cases.

6. A law firm analyzes the time its paralegals spend reviewing legal documents. Review times are normally distributed with a mean (μ) of 90 minutes and a variance (σ^2) of 200 minutes. The firm wants to ensure efficiency. In particular, they are interested in the probability that a paralegal takes more than 100 minutes to review a document. How can this probability be calculated in R? Select the only correct or most applicable statement from the following five.

a) `1-pnorm(100, 90, sqrt(200))`

b) `pnorm(100, 90, sqrt(200))`

c) `1-pnorm(100, 90, 200)`

d) `1-pnorm(90, 100, sqrt(200))`

e) None of the above.

7. Select the only correct or most applicable statement from the following five.

a) Inferential statistics involves drawing conclusions from the characteristics of the population to the characteristics of the sample.

b) The parameters of the population are never accessible to statistical analysis.

c) If samples of size n are drawn from the same population, the mean values M of these samples follow a normal distribution as the number of samples increases.

d) Key figures of the sample, such as the mean value, are denoted with Greek letters.

e) Inferential statistics are used to make statements about the sample.

8. Select the only correct or most applicable statement from the following five.

a) Some hypothesis tests allow three hypotheses to be tested.

b) If you were to apply the terminology of hypothesis tests to the criminal trial by analogy, you could formulate the following alternative hypothesis: "The defendant is innocent".

c) The aim of hypothesis testing is to show that the measured data is very unlikely under the null hypothesis.

d) A type I error refers to the situation where the null hypothesis is retained even though it should be rejected.

e) There is no connection between the type I error and the type II error.

9. Select the only correct or most accurate statement from the following five.

a) Confidence intervals are a method of descriptive statistics.

b) If you were to take 100 samples of the population, calculate their mean values and the 95% confidence interval for this mean value, the confidence interval constructed in this way would contain the true value of the mean value of the population for 95 of these samples.

c) Confidence intervals are irrelevant for legal data.

d) The standard error increases with increasing number n.

e) The standard error concerns the question of how representative the mean value is of the data in the sample.

10. Select the only correct statement from the following five.

a) Multiple linear regression is used to measure the linear relationship between a target variable and a single input variable.

b) Multiple linear regression can be used to measure how much the property value changes when the level of aircraft noise changes and other input variables are kept constant (e.g. plot area, apartment size).

c) Multiple linear regression is not suitable for making predictions in law because there cannot be linear relationships in legal data.

d) In multiple linear regression, the hypothesis test refers to the error term.

e) In multiple linear regression, neither the input nor the target variables may be logarithmized, as this could otherwise lead to distorted results.

Part II: Questions

11. List the words to fill in the gaps (A-E), choosing the correct or most appropriate technical terms from the following list: ratio, sample, value(s), sample space, variable(s), sample, population, observation, selection.

In a recent study, researchers focused on examining workplace discrimination in the tech industry within a specific region. The study involved a detailed analysis of discrimination lawsuits to understand patterns and commonalities. The researchers chose a _____**(A)** of 75 workplace discrimination lawsuits filed against tech companies in this region over the past two years. They were interested in exploring the wider _____ **(B)** of all workplace discrimination lawsuits filed in that region, to which their sample belonged. One specific piece of information they looked at in each case was the compensation amount awarded in a particular lawsuit. Each of the 75 discrimination lawsuits studied by the researchers represents an individual _____ **(C)**. Important _____ **(D)**, which the researchers paid special attention to, are the type of discrimination alleged in the lawsuits as well as the gender of the complainant. The latter takes on the _____ **(E)** “female”, “male”, or “diverse”.

A: sample	*1*
B: population	*1*
C: observation	*1*
D: variables	*1*
E: values	*1*
Total	***5***
A: selection (wrong, because it is not a technical term) ***E: sample space (correct, although not most appropriate term)***	

12. The ICT dataset contains information on defendants brought up on charges in front of the International Criminal Tribunal for Rwanda, the International Criminal Tribunal for Yugoslavia, and the Special Court for Sierra Leone from 1994 to 2010. In particular, the dataset contains data on the sentence lengths (in months) that were imposed for international crimes.

In the Codebook, you find the following information on the variable `sentence`: "Sentence (in months) imposed at trial: "The length of the sentence imposed at trial, measured in months. Life sentences are coded as 624 months which is equal to the longest non-life sentence in the database."

Here is a random sample of seven values:

`624, 30, 132, 300, 624, 600, 96`

These values are stored in a vector `sentence_sample`.

a) Calculate the mean (rounded to two decimal places) and the median of these data. What does the difference between the mean and the median tell you about the skewness of the data?

mean = $\frac{624+30+132+300+624+600+96}{7}$ *= 343.71 (rounded to two decimal places)*	*2*
median = $x_{((7+1)/2)} = x_{(4)} = 300$	*2*
The mean is greater than the median, which tells us that there may be some higher values (outliers) pulling the mean higher relative to the median, because the mean is more sensitive to such outliers. This indicates that the distribution might be right-skewed (though this is not always true).	*2*

b) Calculate the range and the interquartile range of these data.

range = highest value – lowest value = 624 – 30 = 594	*2*
interquartile range = qU - qL = $x_{(3/4(7+1))} - x_{(1/4(7+1))} = = x_{(6)} - x_{(2)}$ *= 624 – 96 = 528 (96 to 624)*	*2*

c) Suppose two more datapoints `100, 3600` are added to the list above. Which of the values calculated above (a and b) will not or not heavily be affected by the additions? Why?

The median will not be affected as one new datapoint lies below, the other above the original media.	*1*
The interquartile range will not be heavily affected. In general, both the median and the interquartile range are more robust (against outliers) than the mean and the range, which will both become a lot greater, especially by adding the very large datapoint 3600.	*1*

d) It is suggested that the datapoint `3600` is an outlier. How can you verify this graphically? How should this datapoint be treated? Why?

*For a graphical verification, we can draw a boxplot to have a look at the data and spot possible outliers. In a boxplot, potential outliers would be displayed as datapoints that fall outside the range of the whiskers. The whiskers extend from the box to the highest and lowest values within 1.5*IQR of the quartiles. Outliers are the points outside the range of the whiskers.* *[Numerically: Outliers can be found like this. Identify the bounds for outliers: The lower bound for outliers is defined as Q1 - 1.5 * IQR. The upper bound for outliers is defined as Q3 + 1.5 * IQR. Any data points that fall below the lower bound or above the upper bound (here 1416) are considered outliers. Please note that this is not a graphical verification as required here.]* *[Alternatively, though less common, a histogram may be used. If the histogram shows a highly skewed distribution, it can indicate the presence of outliers. For instance, a long tail on one side of the histogram suggests outliers in that direction. Outliers may appear as isolated bars (bins with a very low frequency) far from the main body of the histogram. Furthermore, if the data is expected to be normally distributed, any significant deviation from the bell curve shape could indicate outliers.]*	*2*
First, we should try to find out, if the outlier is due to a mistake in the data collection/entry or if it represents a genuine pattern. Since the Codebook tells us that "Life sentences are coded as 624 months which is equal to the longest non-life sentence in the database", this datapoint is a mistake. If the outlier is a mistake, we can remove it.	*1*
If not, we could think about transforming the data (e.g., with a log transformation) or excluding a certain percentage of extreme data points at the low and high ends (like we do for the trimmed mean). The choice of treatment depends on the nature of the data and the analysis we want to perform with the data.	*1*
[Generally, we can try to use statistics that are not too sensitive to outliers, like the median and the interquartile range and avoid those which are more sensitive to outliers, like the mean and the range.]	

e) **Optional (additional marks):** The data is stored in a vector like this: `sentence_extended <- c(624, 30, 132, 300, 624, 600, 96, 100, 3600)`. Write down the R code which can be used to carry out the tests mentioned in d).

***Creating a Boxplot**:* *This will visually help you identify the outlier.* `sentence_extended <- c(624, 30, 132, 300, 624, 600, 96, 100, 3600)` `boxplot(sentence_extended, main="Boxplot of Sentence Lengths", ylab="Sentence Length (months)")`	1

***Investigating the Outlier**:* *You might want to check some summary statistics or specifically investigate the extreme values.* `summary(sentence_extended)`	1

***Using Robust Statistical Methods**:*
If you decide to use robust statistical methods, you might consider methods like median, IQR:
`median(sentence_extended)`
`IQR(sentence_extended)`
	1

Calculate the quartiles and IQR
`Q1 <- quantile(sentence_extended, 0.25)`
`Q3 <- quantile(sentence_extended, 0.75)`
`IQR_value <- IQR(sentence_extended)`
	1

Calculate the lower and upper bounds
`lower_bound <- Q1 - 1.5 * IQR_value`
`upper_bound <- Q3 + 1.5 * IQR_value`
Find indices of outliers
`outlier_indices <- which(sentence_extended < lower_bound | sentence_extended > upper_bound)`
Remove outliers
`sentence_without_outliers <- sentence_extended[-outlier_indices]`
Print the result
`sentence_without_outliers`
Then you can perform your analyses on this new dataset.
	1

f) The standard deviation of `sentence_sample` is 267.5 (to 1 d.p.). What does this mean?

The standard deviation is the square root of the variance. The value given means that, on average, the sentence lengths in the sample vary by approximately 267.5 months from the mean (average) sentence length. This is a measure of how spread out (dispersed) the sentence lengths are. A higher standard deviation indicates a greater spread or dispersion of values from the mean.	*1*
It can be used to assess how well the mean represents the data. The standard deviation is especially useful for comparing the spread of data in different datasets. A single value is not interpretable on its own (but should be compared to other measurements). The standard deviation is preferred over the variance as measure for the spread of the sample data because it retains the original unit of measurement (e.g. months, instead of the uninterpretable months^2).	*1*

g) Figure 1 contains a graphical representation of all available data (n = 148) on the variable `sentence`. Interpret Figure 1.

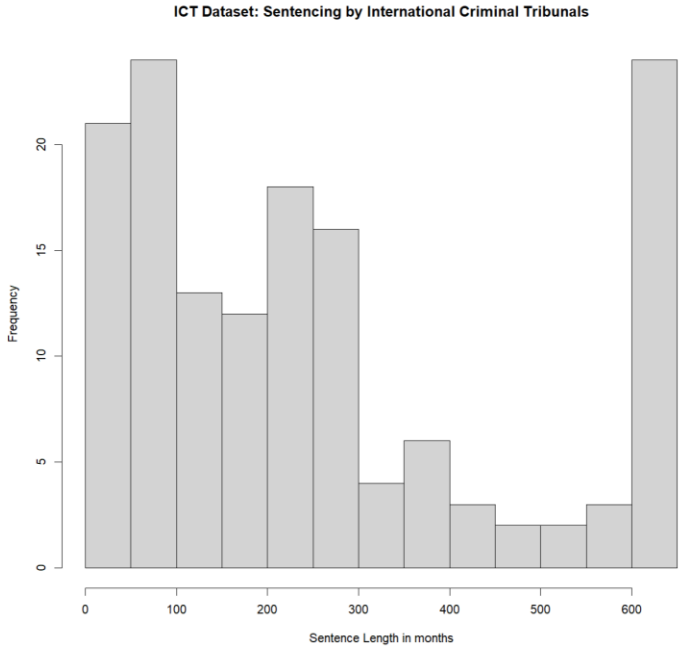


Figure 1

The histogram displays the (absolute) frequencies of the sentence lengths that are grouped into ranges (bins). The histogram shows the frequency of data points within each bin. Generally, there are more shorter sentence lengths than longer ones.	1
***X-axis (Sentence Length in months)**: The horizontal axis shows the sentence length in months. The range of sentences appears to start from 0 and extend to over 600 months.* ***Y-axis (Frequency)**: The vertical axis shows the frequency of each sentence length range. This indicates the number of cases falling within each bin or range of sentence lengths.*	1

***Bins**: The histogram is divided into bins or intervals that represent ranges of sentence lengths. Each bar's height corresponds to the number of sentences that fall within that range.* ***Distribution Shape**: The distribution appears to be unimodal, having a single peak (first bin, range 0-100), which suggests that most of the sentences cluster around a central range of months. There's a noticeable decline in frequency as the sentence length increases, which is typical in such data. [It was acceptable if students wrote that there are two modes.]* ***Outliers**: There seems to be a bar at the extreme right that is separated from the other bars. This could indicate the presence of outliers, or a few cases with extremely long sentences. In this case, it captures life sentences that have been coded as 624 months (see Codebook excerpt).* ***Mode**: The highest bar indicates the most common range of sentence lengths imposed. It seems to be in the first bin, suggesting that the mode of the data is relatively low compared to the maximum sentence length.* ***Skewness**: The distribution seems to be right-skewed, meaning there are a minority of cases with very high sentence lengths compared to the rest.*	
In conclusion, the histogram suggests that while most sentences are clustered in the lower range of months, there are a few cases with much longer sentences, and possibly some with the maximum sentence length as coded in the dataset.	

13.

a) In Switzerland, a random inspection of companies is conducted to check for environmental law compliance. Past data shows that 20% of all companies are in violation. When a violation is present, inspectors correctly identify it 90% of the time. If no violation has occurred, there is a 25% chance that the inspection will still result in a false positive due to misinterpretation of regulations or data errors. If a company has been flagged for a violation following an inspection, what is the updated probability that they were actually in violation (in percent, rounded to two decimal places)?

Given: *P(A) = Probability of violation = 20% or 0.20* *P(B|A) = Probability of correctly identifying a violation when it is present = 90% or 0.90* *P(A') = Probability of no violation = 1 - P(A) = 80% or 0.80* *P(B|A') = Probability of identifying a violation when it is not present (false positive) = 25% or 0.25* *We want to find P(A|B), the probability that a company is actually in violation given that they have been flagged for a violation.*	*2*

*Bayes' Theorem states that: P(A|B) = (P(B|A) * P(A)) / P(B)* *Where P(B) is the total probability of flagging a violation and can be calculated as: P(B) = (P(B|A) * P(A)) + (P(B|A') * P(A'))* *Plugging in the numbers we get:* *P(B) = (0.90 * 0.20) + (0.25 * 0.80) = 0.18 + 0.20 P(B) = 0.38*	*3*
Now we use Bayes' Theorem to find P(A|B): *P(A|B) = (0.90 * 0.20) / 0.38 = 0.18 / 0.38 = 0.4736842105 P(A|B) ≈ 0.4737* *Converted to a percentage and rounded to two decimal places: P(A|B) ≈ 47.37%* *So, the updated probability that a company was actually in violation given that they have been flagged for a violation is 47.37%.*	*3*
Remark: If concluding sentence was missing: -0.5 points. ***Rounding error: -0.5 points.***	

b) A legal aid organization in Switzerland is analyzing the outcomes of cases where they provided assistance. Historical data shows that 40% of the cases where the organization assisted resulted in a favorable outcome for their clients. What is the probability that in the next 15 cases they take on, exactly 6 will result in a favorable outcome (in percent, rounded to two decimal places)?

$P(X = k) = \binom{n}{k} \times p^k \times (1-p)^{n-k}$ *p = 0.4, n = 15, k = 6*	2
$P(X = 6) = \binom{15}{6} \times 0.4^6 \times 0.6^{15-6}$ *= 0.2065976053* *The chance that in the next 15 cases the legal aid organization takes on, exactly 6 will result in a favorable outcome is 20.66%.*	5
Remark: If concluding sentence was missing: -0.5 points. ***Rounding error: -0.5 points.***	

14. The ICT Dataset contains the variables `lifeAtTrial` and `genocide`. The Codebook provides the following information:

lifeAtTrial: Whether trial resulted in a life sentence: Whether a defendant was found guilty and given a life sentence at the trial stage (regardless of what happened on appeal). 1. Yes 0. No

genocide: Whether defendant was convicted of genocide: Was defendant convicted of genocide? 1. Yes 0. No.

a) What statistical test can you use to determine if there is a statistically significant association between the two variables? Which conditions need to be fulfilled for the test?

We want to test if there is a statistically significant association between two categorical variables. For this we can use the Pearson's chi squared test.	*1*
To use a Pearson's chi squared test the following conditions need to be met: The observations used in the calculation of the contingency table are independent. The sample has a size of n > 50. The expected frequency counts for 80% of cells in the table is at least 5.	*2*
Common mistake: The Pearson chi squared test does not require the observations to be normally distributed. The chi squared test is used with categorical data (nominal or ordinal) rather than numerical data that would be tested for normality.	

b) Formulate the hypotheses that you would test.

H_0 = There is no association between the trial resulting in a life sentence and the conviction of genocide. [Alternative: Receiving a life sentence is independent from the conviction of genocide.]	*1*
H_1 = There is a significant association between the trail resulting in a life sentence and the conviction of genocide. [Alternative: Receiving a life sentence is not independent from the conviction of genocide.]	*1*

c) The contingency table below is created based on the two variables `lifeAtTrial` and `genocide`.

(i) Are the conditions for the test you proposed under a) fulfilled in this case? Why? Why not?

The observations used in the calculation of the contingency table are independent: There is no reason to assume that the different trails (observations) are not independent of each other. This condition is fulfilled. The sample has a size of n > 50: The sample size is 149, which is > 50. This condition is fulfilled. The expected frequency counts for 80% of cells in the table is at least 5: All expected frequencies are > 5. This condition is fulfilled. All the conditions to use a Pearson's chi squared test are fulfilled.	*2*

(ii) How was the expected value of 82.03 calculated?

	No Genocide	Genocide	Total
No Life Imprisonment	95 (82.03)	2 (14.97)	97
Life Imprisonment	31 (43.97)	21 (8.03)	52
Total	126	23	149

Expected frequencies in brackets

*The expected frequency is calculated by (row total * column total) / grand total. For No Genocide and No Life Imprisonment it is (97*126)/149 = 82.03 (rounded to two decimal places).*	*1*

d) A statistical test is performed resulting in the following output:

```
data:  d_ict$genocide and d_ict$lifeAtTrial
X-squared = 38.087, df = 1, p-value = 6.766e-10
```

Interpret the result (in light of the hypotheses formulated under b).

The p-value is $6.766*10^{-10}$, which is very small. It is common to set the significance level at 0.05 (5%). According to the table "Interpretation of the p-value" in the Formulary, a p-value ≤ 0.001 means that we have very strong evidence against the null hypothesis (H_0).	*1*
Therefore, we can reject H_0 and accept H_1. This means, there is a statistically significant association between the trial resulting in a life sentence and the conviction of genocide.	*1*

e) How strong is the association between the variables `lifeAtTrial` and `genocide`? Show one way to calculate the relevant value and give an interpretation of that value.

To assess the strength of the association, we can calculate Cramer's V. The formula for this is $V = \sqrt{\frac{\chi^2}{n(min\,(k,l)-1)}}$ *with k and l being the number of rows and the number of columns of the contingency table.* *The contingency table (shown above under c)) is a 2x2 table, which means k = 2 and l = 2. The total number of observations n =149. According to the R output under d)* χ^2 *is 38.087.*	1
$V = \sqrt{\frac{38.087}{149*(min\,(2,2)-1)}} = \sqrt{\frac{38.087}{149}} = 0.5055862435$ *Cramer's V is 0.51 (rounded to two decimal places). According to the formulary, a value between 0.2 and 0.6 indicates that the variables are moderately associated*	4
Also acceptable: Compare* χ^2 *= 38.087 to value calculated with* $n(min(k,l)-1$ *= 149. As 38.087 is closer to 0 than to 149, there is only a weak association between the variables.	

15. Consider the following regression output created on the basis of the ICT dataset.

```
Call:
lm(formula = sentence ~ genocide + numGuil + mfTotal, data = d_ict_NA)

Residuals:
    Min      1Q  Median      3Q     Max
-307.46  -98.16   -8.93  120.16  436.60

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  169.885     36.662   4.634 1.28e-05 ***
genocide     256.132     36.955   6.931 7.59e-10 ***
numGuil       14.618      3.209   4.556 1.73e-05 ***
mfTotal      -13.894      7.579  -1.833   0.0703 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 146.2 on 85 degrees of freedom
Multiple R-squared:  0.4901,    Adjusted R-squared:  0.4721
F-statistic: 27.24 on 3 and 85 DF,  p-value: 1.941e-12
```

The Codebook provides the following information:

sentence: Sentence (in months) imposed at trial: The length of the sentence imposed at trial, measured in months. Life sentences are coded as 624 months which is equal to the longest non-life sentence in the database.

genocide: Whether defendant was convicted of genocide: Was defendant convicted of genocide? 1. Yes 0. No

numGuil: Number of guilty counts: Number of counts on which the defendant was found guilty.

mfTotal: Number of mitigating factors cited: The total number of mitigating factors taken into account for sentencing purposes.

a) Which statistical model was estimated here? Describe the relationship of interest in words.

The statistical model, which was estimated here, is a multiple linear regression model.	*1*
It examines the relationship between the length of the sentence imposed at trial (dependent or response variable) and the following (independent or explanatory) variables: whether the defendant was convicted of genocide, the number of guilty counts and the number of mitigating factors cited.	*2*

b) Describe the regression coefficients. What do they mean in each case? State the regression equation, using the output above. Which variables are statistically significant / not significant? What does this mean?

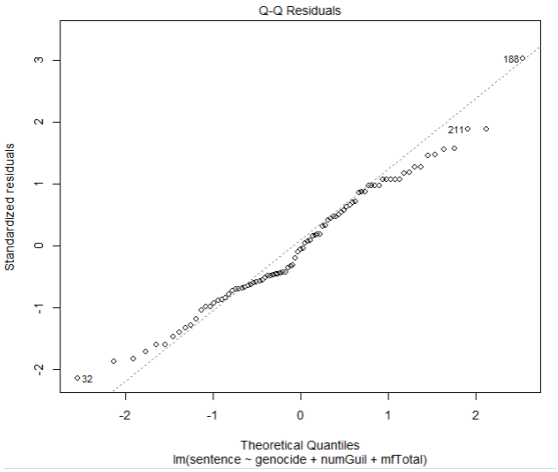
1. ***Intercept (169.885)**: This is the expected value of the sentence length when all the independent variables are 0. In this context, if a defendant was not convicted of genocide, had zero guilty counts, and there were no mitigating factors cited, the expected sentence length would be approximately 169.885 months. The intercept does not have a meaningful interpretation in this case. **[Remark: No need for students to explain this here.]***	
2. ***genocide (256.132)**: The coefficient for genocide is positive, indicating that being convicted of genocide is associated with an increase in the sentence length by 256.132 months, holding all other variables constant.*	1
3. ***numGuil (14.618)**: The coefficient for the number of guilty counts is also positive, suggesting that each additional count of guilt is associated with an increase in the sentence length by 14.618 months, holding all other variables constant.*	1
4. ***mfTotal (-13.894)**: The coefficient for the number of mitigating factors is negative, indicating that for each additional mitigating factor, the sentence length is expected to decrease by 13.894 months, holding all other variables constant.*	1
The corresponding regression equation is: $sentence\ length\ (months) = 169.885 + (256.132 \times genocide) + (14.618 \times numGuil) - (13.894 \times mfTotal)$	2
It is common to set the significance level at 0.05 (5%). Statistical significance is typically determined by looking at the p-values for each coefficient: • ***genocide (p < 0.001)**: The p-value for genocide is less than 0.001, indicating that this variable is statistically significant at common significance levels (e.g., 0.05, 0.01, 0.001).* • ***numGuil (p = 0.035)**: The p-value for numGuil is 0.035, which is less than 0.05, thus this variable is also statistically significant at the 0.05 level.* • ***mfTotal (p = 0.0703)**: The p-value for mfTotal is 0.0703, which is greater than 0.05. This suggests that mfTotal is not statistically significant at the 0.05 level, but may be considered marginally significant or a trend towards significance.* *Statistical significance indicates the likelihood that the relationship observed in the sample data occurred by chance. A statistically significant result suggests that there is evidence to believe that there is a true relationship in the population from which the sample was drawn. For genocide and numGuil, there is strong evidence that they have a significant impact on the*	2

sentence length. However, for mfTotal, the evidence is weaker, and it might not have a significant impact on the sentence length at the 0.05 significance level. *To conclude, the conviction of genocide and the number of guilty counts have a significant impact on the sentence length; the number of mitigating factors cited has not.*	

d) The researcher also tested the following model: `sentence ~ genocide, data = d_ict_NA` and obtains an adjusted R^2-value of 0.2969. What does this value tell you? How does it compare to the adjusted R^2-value of 0.4721 above? Which model is to be preferred and why?

The adjusted R^2-value indicated how well a regression model fits the data, taking into account the number of predictions in the model. Generally, a higher R^2-value indicates a better-fitting model, as it suggests that a larger proportion of the variance in the dependent variable is explained by the predictor(s).	*2*
The adjusted R^2-value of 0.2969 means, the variable `genocide` *explains 29.69% of the variance in the variable* `sentence`*. The first model also considered the variables* `numGuil` *and* `mfTotal`*. With this it explains 47.21% of the variance in the variable* `sentence`*. To conclude, the fist model is to be preferred, because it has a higher R^2-value, which means it fits better, without taking into account an unreasonably high number of variables.*	*3*
Remark: Students should state the difference between the concept of the R^2-value and the adjusted R^2-value.	

Optional (additional marks): Interpret the plot below.



To use a linear regression model, as we did, some assumptions need to be met. One of them is that the residuals need to be normally distributed. The Q-Q plot (Quantile-Quantile plot) is a graphical tool that can be used to assess whether a dataset follows a particular theoretical distribution, such as a normal distribution.	*1*
For this, theoretical quantiles which would be expected under the normal distribution (x-axis) are plotted against the sample quantiles from our dataset (y-axis). The dashed line (y=x) is included to make the interpretation easier: If the data perfectly followed the assumed distribution, the points on the Q-Q plot should fall along this line.	*1*
In our example the points roughly follow the line, with a better fit in the middle and a little worse fit at the tails.	*1*
There are some points at the upper end of the plot that deviate from the line. These are labeled with numbers like "2119" and "1869". These labels usually correspond to the indices of the observations in the dataset. These points are potential outliers since they do not follow the expected trend for data coming from a normal distribution.	*1*
The slight curvature in the lower tail (bottom left) suggests that the lower end of the data distribution might be slightly lighter-tailed than the normal distribution, while the upper tail (top right) suggests a heavier tail than normal distribution because the points deviate upward from the line. We can conclude that the data is roughly normally distributed.	*1*