NAME:

STAT 503 – Statistical Methods for Biology

Homework 7

24 Points (27 available). Due at 11:59 PM on **Wednesday**, April 28, 2021.

**Part 1:** Alzheimer's disease is a degenerative neurological condition that affects approximately 50 million people worldwide. It has no cure and eventually leads to dementia, memory loss, and death. Early intervention and treatment may slow the disease's progression. However, because an Alzheimer's diagnosis requires patients to display symptoms of dementia, the disease is usually not recognized until it has become relatively advanced. There is therefore considerable interest in the development of blood tests or imaging tools that would allow patients to be diagnosed before symptoms appear.

Physically, Alzheimer's disease is closely associated with a protein called amyloid-b (Ab), which builds up and develops plaques in the hippocampus and other parts of the brain. Plaques begin to develop 10-20 years before patients show symptoms, so in theory, Ab levels could serve as an early diagnostic marker for Alzheimer's. However, the proteins are not normally detectable in the blood. Under natural conditions they can only be measured by autopsy.

In 2001, DeMattos *et al*. (*Proceedings of the National Academy of Sciences* 98:8850) described a method to make Ab detectable in blood samples from mice. Their technique used a monoclonal antibody to Ab. When injected into the mice, the antibody made it easier for Ab to enter the bloodstream and also slowed the proteins' degradation. As a result, the researchers saw a large increase in detectable Ab within 24-hours after injection.

In this homework, you will analyze data from a follow-up study (DeMattos *et al*. 2002, *Science* 5563:2264-2267) in which the researchers investigated the relationship between Ab levels in blood plasma and Ab loads in the brains of 49 mice. In each mouse, the plasma concentration of Ab was measured in pg/ml using ELISA (a standard laboratory method for quantifying proteins), and the Ab load in the hippocampus was measured as a percentage of total area using immunofluorescent staining (see Fig. 2 in the paper for example images).

You can find the data on the website in the file demattos\_et\_al\_2002\_amyloid.csv. In addition, you will find a PDF file containing the "original" data records (I read these data from a figure in the paper; they are not the author's actual notes). The data include two potential explanatory variables. The variable brain gives the percentage Ab load in the mouse's brain, and the variable severity classifies each mouse into one of four ordinal classes based on its Ab load. Please use these files to answer Questions 1–15.

**We will analyze the data using both ANOVA and regression (or their non-parametric alternatives). Our goals are to determine whether plasma concentrations of Ab are related to brain Ab load, and to describe the relationship if it exists.** Our methods will be somewhat different from those used by the authors.

* 1. [1 point] State the null and alternative hypotheses for the one-way ANOVA, using severity as the explanatory variable.

Null hypothesis: There is no relation between the plasma concentration of Ab and the brain Ab load in the brains of 49 mice.

Alternative hypothesis: there is a relationship between the plasma concentration of Ab and the brain Ab load in the brains of 49 mice.

* 1. Load demattos\_et\_al\_2002\_amyloid.csv into R. **Make sure that severity is a factor**. Then fit the model, obtain the residuals, and add them to your data frame as a new column (mutate() may be useful). Use the residuals to complete the following preliminary tasks:
     1. [1 point] **Check for possible outliers**. Present any graphs that you use (with captions), and briefly explain what you found. If the plot(s) suggest that an outlier exists, (i) identify the row number for each potential outlier in the data, and (ii) explain how you have addressed the issue. If you find a mistake in the .csv file, correct it. If you leave an outlier unchanged, say so, and explain your reasoning.

Some guidance:

* Review Lecture 2.9. Remember the following principles:

1. Outliers are only a problem if they cause the data to violate the model’s distributional assumptions.
2. You should always check for outliers using the same values that you will check the distributional assumptions. You may want to run these checks (in 2b) before you decide how to handle any outliers.
3. Consider the overall distribution of the data. If a point looks like a possible outlier in a boxplot, but is not very extreme relative to the outermost point in the opposite direction, and the data appear to be normal, then the "outlier" may not have a substantial effect on the analysis.
4. Data should **never** be removed from an analysis unless it is unambiguously flawed **and** you cannot fix it.
5. If you change anything in the data, you **must** recheck all of the model assumptions, including rechecking for outliers. **Your answer should discuss any follow-up checks that you run, but you only need to include your original plot(s)**.

* The arrange(), which.min(), or which.max() functions can help you figure out which row the outlier is in.
* If you need to fix a data-entry error, you can either correct the .csv file in Excel and reload it, or you can correct it in R using code similar to:

myData$myVariable[rowIndexOfOutlier] # check this is right row

myData$myVariable[rowIndexOfOutlier] <- correctValue

myData$myVariable[rowIndexOfOutlier] # confirm fix

Chart, box and whisker chart

Description automatically generated

Figure 1: Boxplot of plasma concentration level versus severity in Demattos’ (2002) amyloid data. Overall, as plasma level increased, the severity also increased.

1. There appears to be 2 outliers in the 4th severity level at around 9000 plasma level.
2. There was no mistake found when comparing the current data with the original data. This outlier is left unchanged as the data should not be changed without justification. In addition, the outlier doesn’t actual violate the assumptions of normality.
   * 1. [1 point] **Check the model's assumptions** (you may assume independence and random sampling). Please (i) clearly identify each assumption that you are checking, (ii) state whether or not it has been met, and (iii) present the evidence that you are using to check it. Evidence can take the form of plots, formal goodness-of-fit tests, or calculations based on statistics. Please only provide the evidence that you feel is really needed to verify the assumptions.

i. One assumption is checking for normal errors or normal populations, another assumption is to check for equal variance, the third assumption is to check for no outliers

ii. The normal errors assumption is met. The equal variance assumption is met. The outlier assumption is not met.

iii. The equal variance assumption is met as the is calculated to be 2.071516 which is fairly close to 2. In addition, a Levene’s Test was ran and it indicated that the assumption of equal variance was not violated as the Pvalue was 0.4266.

* 1. [1 point] Present the ANOVA table for your analysis, and report your interpretation of this table: decide whether or not to reject the null hypothesis at , and explain what your conclusion means. Your answer should cite specific information from the table.

You may use R to get the ANOVA table, or you may calculate it by hand (use the pf() function in R to obtain a -value). If you want to export the ANOVA table from R, first assign the output from anova() to an object, and then treat it like a regular table:

anovatable <- anova(myModel)

# follow procedure in Tutorial 3 to output anovatable to Word.

Table 2: One-way ANOVA for severity on blood plasma concentration.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **Df** | **SS** | **MS** | **F** | **P** |
| **Severity** | 3 | 89855388 | 29951796 | 24.507 | 1.494e-09 |
| **Residual error** | 45 | 54997095 | 1222158 |  |  |
| **Total** | 48 | 144852483 |  |  |  |

From the table, the F3,45 = 24.507 and P = 1.494e-09 and the global F-test would be significant for . Therefore, the null hypothesis is rejected.

* 1. Suppose that we are interested in a planned contrast between severity class 0 and class 1.
     1. [1 point] Using the output from summary() and confint(), report a point estimate, standard error, and 95% confidence interval for the contrast.
     2. [1 point] The standard error for this contrast is calculated as,

Please explain why this structure makes sense. Your answer should explain what each term in the equation represents (the terms are , , and ), give the value of each term, and explain what a contrast represents. It may be helpful to compare this equation with the equations for the SE of and the SE for the exact two-sample -test.

* 1. Look at the results from the summary() method.
     1. [0.5 points] At , which treatments have mean plasma levels that differ significantly from the class 0 group (if none of them are different, say so)?
     2. [0.5 points] Why does the result in 6a **not** indicate that all of the groups are different?
  2. There are a total of 5 possible pairwise comparisons (or unplanned contrasts) that we can make among the four treatment levels in this study. This exercise will illustrate how our choice of multiple testing correction can affect the results of these comparisons. The cheat sheet titled *Methods to control error inflation in multiple comparisons* (available on the website) may be helpful.
     1. [1 point] Each of the five contrasts is listed in the table below. Please fill in the -values that you obtain for each comparison using no correction, the Bonferroni correction, the Bonferroni-Holm correction, and the Tukey-Kramer method. The first three of these may be found using the pairwise.t.test() function. For Tukey-Kramer, use aov(fit) %>% TukeyHSD(), where fit is the output from lm(). I have filled in the first row so that you can confirm that your code is correct. Note that I am using E to indicate scientific notation. **Give values to at least 2 significant figures**.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Contrast | None | Bonferroni | Bonferroni-Holm | Tukey-Kramer |  |
| 1 – 0 | 6.7E-5 | 4E-4 | 2.7E-4 | 3.77E-4 |  |
| 2 – 0 |  |  |  |  |  |
| 3 – 0 |  |  |  |  |  |
| 2 – 1 |  |  |  |  |  |
| 3 – 1 |  |  |  |  |  |
| 3 – 2 |  |  |  |  |  |

* + 1. [0.5 point] Which method is the best choice for the current analysis, and why? Your answer should explain the -values that you got in 6a, but should not be based on those -values.
    2. [0.5 point] Suppose that instead of an ANOVA, we had decided to run a non-parametric Kruskal-Wallis test. Which multiple-testing method would be preferable in this scenario, and why?
  1. [2 points] Use the predict() method with interval = "conf" to estimate the mean preference index for the each of the four treatment groups (see Tutorial 9, Section 2.5.6). To do this, a data frame with one row for each severity level should be passed to the newdata argument in predict().

Use the point estimates and confidence intervals from predict() to generate a graph that shows (i) the data, (ii) the mean for each group, and (iii) a 95% confidence interval for the mean of each group, as illustrated in Tutorial 9, Section 2.6. Your caption should clearly identify each type of symbol in the plot. Finally, (iv) label your graph with letter codes to indicate which group means are significantly different from each other (see Lecture 5.10 for example code). The cheat sheet on the website titled *Identifying differences in pairwise comparisons* explains how to figure out the set labels.

* 1. [1 point] In these data, what proportion of the variation in plasma Ab concentration is accounted for by differences among the means for the different severity classes (i.e., by the effect of plaque severity)? Please identify the statistic used to find this answer.
  2. [1 point] Briefly explain your biological interpretation of the results. For evidence, your discussion can cite the ANOVA table results, hypothesis testing results for pairwise comparisons, and/or the graph in question 8. You do not need to repeat any statistics here. Simply explain what they mean, biologically. In particular, consider the goals of the analysis.
  3. [1 point] In general terms, how confident are you in the repeatability of these results? Please explain your reasoning.
  4. [1 point] Fit a new linear model for plasma, this time using the percent coverage of plaques in the brain column as your explanatory variable. Because brain is numeric, lm() will fit a regression instead of a one-way ANOVA. As you did in Question 2, add the residuals for the new model to the dataset and use them to check the model assumptions. Provide any plots that you use, along with any test results and your conclusions. Note that you should be able to reuse most of your code from question 2 here.
  5. [1 point] Report the results of the regression analysis. Specifically, your answer should indicate whether or not a linear relationship exists and should estimate the slope of that relationship at . (Note: there are two different statistics that can be used to test the null hypothesis that the slope is equal to zero; you only need to cite one of them).
  6. [1 point] What is the expected plasma concentration of Ab in a mouse that has a 12% hippocampal Ab load? **Please show your work**, including the symbolic equation that you are using to calculate your answer.
  7. Use the predict() method to find a 95% **prediction interval** for the rate of heat loss that would be seen in a mouse with a 12% Ab load, versus a mouse with a 30% load.
     1. [0.5 points] Report the intervals here.

* + 1. [0.5 points] How do these intervals differ from a confidence interval?
    2. [0.5 points] What explains the difference in width at 12% versus 30%?
  1. [1 point] In your opinion, does the ANOVA or the regression represent a better model for these data? Please briefly explain your answer. (There is not necessarily a right or wrong answer to this question; it is open to interpretation).

**Part 2:** Wilson et al. (2011, *Journal of the National Cancer Institute*, 103:876–884) used a dataset that tracked the status of 10,383 male health professionals for 20 years, from 1984 to 2004, to look for associations between coffee consumption and the occurrence of prostate cancer. The data that we have available in wilson\_et\_al\_2011\_coffee\_and\_cancer.csv compares the men who drank no coffee to the heaviest coffee consumers, who drank 6 or more cups of coffee a day. The subjects' coffee-drinking habits are classified in the variable coffeeStatus. The second variable, cancerStatus, has the value "cancer" if the individual developed advanced prostate cancer during the study, and "no cancer" if he did not develop advanced cancer during the study.

* 1. [1 point] Present a 2 2 contingency table for this dataset, with the explanatory variable represented as columns. Contingency tables are discussed in **Tutorial 3, Section 3.4**. To export your table to Word, use the as.data.frame.matrix() function, as shown below,

tableForWord <- as.data.frame.matrix(myContingencyTable)

and then follow the usual procedure for exporting tables. This function is required because as.data.frame() will produce unexpected results when it is run on a object with class *table* (try it!). If ***officer*** is being uncooperative, simply type your table by hand.

* 1. [1.5 points] Use chisq.test() to conduct a contingency test on your table from Question 16. (a) Cite evidence checking the assumptions of the test. (b) Report the null hypothesis that you are testing, the test statistic (with degrees of freedom), the -value, and your conclusion, and (c) interpret the result: is there an association between coffee consumption and advanced prostate cancer?

* 1. [1 point] Use the Agresti-Coull method to estimate probability of developing cancer in the non-coffee drinkers and in the coffee drinkers, at a confidence level of 95%. You may do this in R or by hand. Please write 1-2 sentences reporting your results.
  2. [1 point] Is it possible to use relative risk instead of an odds-ratio with this dataset? Why, or why not (hint: is this a case-control study?)?
  3. [1 point] Estimate the odds ratio for the coffee group versus the no coffee group, with a 95% confidence interval. By what factor does coffee consumption change the odds of developing advanced prostate cancer, according to your analysis? Show your work.
  4. [1 point] Would it be legitimate to claim that this study shows that coffee drinking lowers the odds of developing advanced prostate cancer? Why or why not? (Note: your answer to this question should not depend on the results of your earlier analysis).
  5. [1 point] Please attach your code.