

HOMEWORK 13: LINEAR MODELS

Instructions

This assignment is due in Canvas by **Wednesday, December 8 at 11:59pm**.

Your submission needs to include **your R code, the corresponding R output, and your narrative** interpretation/responses to the questions (in complete sentences). The easiest way to do this is to work in an **R Notebook**, as described in the first homework guide and demonstrated in the corresponding video. If you use an R Notebook, you will submit the notebook's **HTML file** to Canvas.

Exercise

Refer to the datasets indicated on the homework assignment page on Canvas.

1. A bacteriologist is studying the antibiotic properties of four solutions (labeled T1 through T4 in the dataset `antibiotics` on Canvas). The bacteriologist measures the reduction in bacterial cultures when each solution is applied and would like to know which solutions are more effective than others as antibiotics.
 - a. Examine the data by making a side-by-side boxplot for the reduction in bacterial cultures for each treatment group (refer to the Homework 2 Guide for a refresher on making box plots).
 - b. What are the null and alternative models for an ANOVA in this context?
 - c. Evaluate the assumptions for ANOVA (refer to the Lecture 11a Prereading and the Homework 2 Guide for reminders about the relevant R code).
 - d. What is the value of the test statistic for an ANOVA of the antibiotic data? What does this test statistic represent, in the context of the scenario? Name two conditions that, if changed, would result in a **smaller** test statistic.
 - e. Explain what conclusion the bacteriologist should draw, based on the data.
 - f. Find Tukey-adjusted 95 percent confidence intervals for the difference in each pair of solutions. Which solutions are significantly different from the others?
 - g. Would the unadjusted confidence intervals be wider or narrower? Explain.
2. An epidemiologist studying factors associated with glucose levels in the United States would like to know whether glucose levels differ by race/ethnicity. Because glucose levels are highly skewed, the researcher decides to investigate the median, rather than the mean, for each group. The researcher plans to conduct a Kruskal-Wallis test with a Bonferroni adjustment for post-hoc comparisons to keep the family-wise Type I error rate to no more than .05. Refer to the `glucose` dataset on Canvas.
 - a. What is the sample median for each race/ethnicity group?
 - b. What are the null and alternative models for this test?

- c. Find the p-value for the Kruskal-Wallis test. Explain what conclusion you will draw from the statistical test.
- d. Explain what significance level will you use for each post-hoc comparison to keep the family-wise Type I error rate to no more than .05?
- e. The p-values for each Wilcoxon pairwise comparison are listed below. Which groups are significantly different from the others?
- f. How would your answer to (d) differ if you were not adjusting your conclusions to account for multiple comparisons? Why is it important to consider the consequences of conducting multiple tests in a research study?

Comparison	Wilcoxon p-value
Black vs Hispanic	.3260
Black vs Other	.9157
Black vs White	.0195
Hispanic vs Other	.3737
Hispanic vs White	.0009
Other vs White	.0149

3. A public health worker wants to use air quality (as measured by the number of particulates in the air in parts per million) to explain variation in the childhood asthma rate for various cities. Refer to the dataset `asthma` on Canvas.
 - a. Estimate a linear model for this analysis. What is the estimated linear equation for the model? Explain the interpretation of the slope.
 - b. Create scatterplots (see p. 3) for (i) asthma rate vs. air quality and (ii) the residuals of the linear model vs. air quality. Evaluate the assumptions of the linear model.
 - c. The public health worker wants to know whether there is strong evidence of a relationship between air quality and childhood asthma. What are the null and alternative models for the statistical test that can address this research question?
 - d. Explain what conclusion the public health worker should draw, based on your analysis.
 - e. What is the prediction of childhood asthma for a city that has a particulate air quality of 10 ppm? Show how this is calculated.
 - f. Find a 95 percent confidence interval for the mean childhood asthma rate in cities that have particulate air quality of 10 ppm.
 - g. The public health worker is visiting a city with a particulate air quality of 10 ppm. What is a 95 percent interval for the prediction of that city's childhood asthma rate? Explain why this interval is different from the interval in part (f).
 - h. If the public health worker were visiting a city with a particulate air quality of 15 ppm, would the prediction interval be narrower or wider? Explain.

R Code

Scatterplots can be made with the ggplot function `geom_point()`, like so:

```
ggplot(crabs) + geom_point(aes(x=width, y=weight))
```

This example plots crab weights on the vertical (y) axis and widths on the horizontal (x) axis.

You can plot the residuals of a linear model (either for ANOVA or linear regression) by using the `residuals()` function with the y aesthetic. For example,

```
myModel <- lm(weight ~ width, data = crabs)
ggplot(crabs) + geom_point(aes(x=width, y=residuals(myModel) ))
```

Similarly, you can use `residuals()` to examine a Q-Q plot of the residuals. For example,

```
qqnorm(residuals(myModel))
qqline(residuals(myModel))
```

Refer to the lecture 11a, 12a, 13a, and 13b slides for additional commands related to linear models.