Homework 5 – Stat 230 – Fall 2022

Due date: Friday, October 14

Complete the following exercises and submit your assignment via gradescope (linked on the course webpage).

Problems to start after class Oct 7

Q1

How do bats make their way about in the dark? Echolocation requires a lot of energy. In this problem, you will explore how energy expenditure is related to body mass from 20 energy studies on three types of flying vertebrates: echolocating bats, non–echolocating bats and non–echolocating birds.

```
bats <- read.csv("https://aloy.rbind.io/data/bats.csv")
```

(a) Fit a multiple linear regression model with `log(Energy)` as the response variable and `log(Mass)` and `Type` as the predictor variables. Report the fitted regression equation.

(b) What indicator variables did R create to represent the categorical variable `Type`?

(c) Based on the fitted model you reported in part (a), write a fitted model equation for each type of flying vertebrates (echolocating bats, non–echolocating bats and non–echolocating birds).

(d) Conduct the sums of squares F-test that can be used to determine whether `Type` is associated with the energy after accounting for mass. State the hypotheses, p-value, and conclusion in terms of the problem (that is, say things about the flying vertebrates).

Q2

Data were collected on the volume of users on the Northampton Rail Trail in Florence, Massachusetts. Variables in the data set include the number of crossings on a particular day (measured by a sensor near the intersection with Chestnut Street, `volume`), the average of the min and max temperature in degrees Fahrenheit for that day (`avgtemp`), and a dichotomous indicator of whether the day was a weekday or a weekend/holiday (`weekday`).

```
railtrail <- read.csv("http://aloy.rbind.io/data/RailTrail.csv")
```

Consider the following full linear model predicting the volume on the Northampton Rail Trail.

```
rail_lm <- lm(volume ~ hightemp + lowtemp + cloudcover + precip, data = railtrail)
```

(a) Test whether `cloudcover` can be dropped from the regression model given that `precipitation`, `hightemp`, and `lowtemp` are retained. Use an appropriate F test. State the hypotheses, p-value, and conclusion in terms of the problem (that is, say things about the rail trails and an appropriate population). [Note: you should know how to do this by hand given the ANOVA table. However, R will do the test for you with the code `anova(model1, model2)`.]

(b) Test whether both `lowtemp` and `cloudcover` can be dropped from the model given that `hightemp` and `precipitation` are retained. Use an appropriate F test. State the hypotheses, p-value, and conclusion in terms of the problem (that is, say things about the rail trails and an appropriate population).

Problems to start after class Oct 10

Q3 (adapted from *Statistical Sleuth* 9.19)

R.A. Miech and M.J. Shanahan[1] investigated the association of depression with age and education, based on a 1990 nationwide (U.S.) telephone survey of 2,031 adults aged 18 to 90. Of particular interest was their finding that the association of depression with education strengthens with increasing age—a phenomenon they called the "divergence hypothesis."

The researchers constructed a depression score from responses to several related questions. Education was categorized as (i) college degree, (ii) high school degree plus some college, or (iii) high school degree only.

[1] Miech, R. A., & Shanahan, M. J. (2000). Socioeconomic status and depression over the life course. *Journal of health and social behavior*, 162-176.

For parts (a) and (b) carefully write the mean function (i.e., signal) that you would use to represent the following situations. Be sure to clearly define any indicator variables that you use—don't make us guess!

(a) The depression score should change linearly with age in all three education categories, with possibly unequal slopes and intercepts.

(b) Modify your model from part (a) to specify that slopes of the regression lines are equal in categories (i) and (ii) of education but possibly different in category (iii). The y-intercepts should still be possible unequal for all three categories.

(c) Based on your model in part (a), identify a single regression coefficient that measures the diverging gap between categories (iii) and (i) with age. That is, what coefficient allows the difference in mean depression score for college (i) and high school only (iii) to widen as people age.

(d) Based on your model in part (b), identify a single regression coefficient that measures the diverging gap between categories (iii) and (i) with age.

Q4

The `RailTrails` data set contains information about 104 homes sold in Northampton, Mass., in 2007. In this problem, you will use the following variables:

Variables	Description
`Adj2007`	Estimated 2007 price (in thousands of 2014 dollars)
`Distance`	Distance (in miles) to the nearest entry point to the rail trail network
`SquareFeet`	Square footage of interior finished space (in thousands of square feet)
`NumFullBaths`	Number of full bathrooms in the house

Suppose that researchers are interesting in determining whether the selling price of a home in Northampton is impacted by its proximity to a bike trail, and if so how is it impacted. The researchers also want to control for the size of a home and the number of full bathrooms.

```
rails <- read.csv("https://aloy.rbind.io/data/RailsTrails.csv")
```

(a) Fit the multiple linear regression model using `log(Adj2007)` as the response and `log(Distance)`, `log(SquareFeet)`, and `NumFullBaths` as the predictors. Report the fitted regression equation.

(b) Assess the model conditions for the model using residual plots and write a brief summary

of your findings.

(c) Fit a model that adds interactions between all of the explanatory variables, so there will be three two-way interaction terms and one three-way interaction term. Report the fitted regression equation.

(d) Use an appropriate F-test to determine whether the model complex model in part (c) is a substantial improvement over the simpler model you fit in part (a). Report your findings.