STAT2050 Winter 2019
Assignment 4

Due: Wednesday, April 3, by 4:00pm.

---

1. (Adapted from *The Statistical Sleuth.* (2013). Ramsey and Schafer. Brooks/Cole.). In an experiment that aims to study the effect of a certain fatty acid (CPFA) on the level of protein in rat lives, there were 6 levels of CPFA considers. For each level, there was replications. The observed data were in the "rat.csv" from CourseLink. Load it into R as

```
rat.dat=read.table("xxxxxx/rat.csv", sep=",", header=TRUE)
head(read.dat)
```

Have a good look at the data set. You will notice that CPFA has been coded in two ways: "TREATN" which is a numerical variable with values of 0, 50, 150, 300, 450 and 600; "TrtGroup" which is a character or "factor" variable with levels of "CONTROL", "CPFA50", "CPFA150", "CPFA300", "CPFA450", and "CPFA600".

a) Use R, plot a scatter plot of the "PROTEIN agains TREATN".

```
PROTEIN=rat.dat[,1]
TrtGroup=rat.dat[,2]
TREATN=rat.dat[,3]
plot(TREATN, PROTEIN)
```

b) We consider a simple linear regression (SLR) model for the data. Write out the model, specify your model assumptions.

c) Fit the SLR model using R, include the code you use in R to fit the model, the R output, and insert the fitted regression line to the scatter plot you produced in a).

d) We now fit an ANOVA model to the data, use the multiple treatment group of CPFA for model fitting. Write out the model, specify the model assumption. Include the R code and the ANOVA table.

e) Perform a Lack-of-Fit (or Goodness-of-Fit) F-test to determine which of the SLR or ANOVA model is preferable. Follow the 5-step hypothesis test procedure to get full marks.

2. (Adapted from *The Statistical Sleuth.* (2013). Ramsey, F. L. and Schafer, D. W. Brooks/Cole publishing.) It has been theorized that developing countries cut down their forests to pay off foreign debt. Data on debt, deforestation and population appear in the data set `deforest.csv`, available on Courselink. The variables include country (COUNTRY), debt (in millions of dollars) (LDEBT), deforestation (in thousands of hectares) (LDEFOREST) and population size (in thousands of people) (LPOP). The variables of debt, deforestation, and population have been log transformed.

(a) Create a scatterplot matrix for the data set. Include the plot in your solutions. Do the relationships between LDEBT, LDEFOREST and LPOP appear to be linear?

(b) Does LDEBT exert any effect on LDEFOREST? Fit a simple linear regression model and include the output in your solutions. What conclusions can you draw about the effect of debt on deforestation? Justify your response with values from your output.

(c) Does LDEBT exert any effect on deforestation, after the effect of population on LDEFOREST is accounted for? Fit a multiple linear regression model and include the output in your solutions. What conclusions can you draw about the effect of LDEBT on LDEFOREST, after accounting for LPOP? Justify your response with values from your output.

(d) Provide a plausible explanation for why the coefficient for LDEBT can be negative (and non-significant) in the multiple linear regression model, but positive and highly significant in the simple linear regression model.

3. (Adapted from *The Statistical Sleuth.* (2013). Ramsey, F. L. and Schafer, D. W. Brooks/Cole publishing.) Natal dispersal distances are the distances that juvenile animals travel from their birthplace to their adult home. An understanding of dispersal distances will help to identify which species are vulnerable to habitat fragmentation. Information on maximum dispersal distance (in km), body mass (in kg), and diet type (O=Omnivore, C=Carnivore, H=Herbivore) is contained in the file `ex1124.csv` on Courselink.

The variables distance and body mass are continuous, while diet type is categorical. To analyze the data set, we will need to create indicator variables for diet type; we will also have to transform *both* continuous variables with the natural log function, in order to create a linear model. Here is the code to do it:

```
# Extract variables from data frame for easy access
ex1124=read.table("xxxxxx/ex1124.csv", sep=",", header=TRUE)
type <- ex1124[,"type"]
bodymass <- ex1124[,"bodymass"]
maxdist <- ex1124[,"maxdist"]

# Create indicator variables for diet type.
omni <- herbi <- carni <- rep(0,64) # make 3 variables of length 64
                           # all values equal to 0
omni[type=="O"] <- 1 # if type ==O, then omni <- 1; else omni <- 0;
herbi[type=="H"] <- 1
carni[type=="C"] <- 1
lbodymass <- log(bodymass)
lmaxdist <- log(maxdist)

# Attach new variables to data frame
ex1124 <- data.frame(ex1124,omni,herbi,carni,lbodymass,lmaxdist)
attach(ex1124,pos=1)
```

(a) Justify whether the log transformation of the body mass or/and dispersal distance variables are needed, by creating a scatterplot of (1) distance versus body mass; (2) distance versus log of body mass; (3) log of distance versus body mass; and (4) log of distance versus log of body mass. You can create all four plots in a single graphing window by using the following command *before* your `plot` commands:

```
par(mfrow=c(2,2))
# Specifying 2 rows and 2 columns allows to create a 2 by 2 matrix plot
# to accommodate four plots in one graphing window.
```

Include all plots in your answer. Which relationship appears to be linear?

Whatever your answer to a), from here on, use the log transformation for both the body mass and distant in the analysis.

(b) Fit a model that allows three separate lines with different slopes and intercepts. Include the output from a summary command and the anova command.

(c) Fit a parallel lines model. Include the output from a summary command and anova command.

(d) Is the parallel lines model adequate or does the three separate lines model provide a better fit? Perform an 5-step test to support your conclusion.

(e) Fit a single line model. Include the output from a summary command and anova command.

(f) Is the single line model adequate or does the parallel lines model provide a better fit? Perform an 5-step test to support your conclusion.

(g) In the parallel lines model, interpret all $\hat{\beta}$'s.