Name _____ **DataSet Number**_____

Due: October 16, 4PM

Work on your own

Demonstrate your mastery of the material. Each student has a unique data set. Don't share code or discuss the questions or problems here. It is acceptable for students to "Google" for information and to learn. However, it is not acceptable to post requests for help, contact others about these particular problems, or provide help for others.

Instructions:

Data sets have been uploaded in a folder, `http://pj.freefaculty.org/stat/psyc790/student-test1`. Pick the one that is correct for you. Code hint:

```
dat <- read.table("student-??.txt", header = TRUE)
## Maybe add stringsAsFactors = FALSE, I'm not sure if that helps
head(dat)
str(dat)
```

The variables are

hand Either "left" or "right", which is the respondent's dominant hand

pref Either "Clinton" or "Trump", the respondent's favorite presidential candidate

iq The respondent's intelligence quotient

weight The respondent's weight, in pounds

dbp The respondent's diastolic blood pressure

The first 3 lines of the dataset I am using look like this:

```
   hand    pref gender  iq weight  dbp
1 right   Trump   male  98  152.6 85.6
2 right Clinton female 116  160.1 86.6
3 right Clinton female 103  153.1 83.9
```

Questions

1. (20 pts) Describe the data. Refer to a summary table(s) and figures(s) where necessary. As you write, think about what information an interested professional reader wants to know about your data.
2. (10 pts) Create a cross tabulation table that demonstrates the effect of handedness (left or right) on choice between candidates. Write a paragraph to summarize the table and interpret it.
3. (30 pts) Use OLS regression to analyze the predictive effect of diastolic blood pressure and weight. My doctor says heavier people have higher blood pressure, and I wonder if she is correct.
   (a) (5 pts) Write down the theoretical model you estimated, being careful to identify parameters and define terms. How is the mismatch between the model and the data summarized in your estimation process?
   (b) (5 pts) Introduce a summary table for your estimated regression.
   (c) (5 pts) In your summary table, number the following items and interpret them
      1. Estimate of the intercept
      2. Estimate of the slope
      3. Estimated standard error of the slope
      4. Estimated standard deviation of the error term
      5. The R^2.

(d) (5 pts) Prepare a scatterplot, including a predictive representation of the regression model.

(e) (5 pts) Write a paragraph about the observed effect of weight on blood pressure.

(f) (5 pts) Conduct a hypothesis test of the estimated slope. Let the null value for the slope be 1.2. Write out the steps of your analysis.

4. (30 pts) I wonder if men or women have higher blood pressure. Conduct a regression analysis.

   (a) (5 pts) Create some graphs to represent the relationship between dbp and gender.

   (b) (5 pts) Estimate your regression model. Warning: turn gender into a factor variable, if it is not a factor already. Introduce a summary table for your estimated regression

   (c) (5 pts) For men and women, calculate the predicted dbp values, along with 95% confidence and prediction intervals for your estimates.

   (d) (5 pts) Write a paragraph that discusses the differences between estimation and analysis of a numeric predictor (as in previous question) and a categorical predictor. Sometimes this is easier to explain if you check how R sees your variable by running "contrasts(mydata$gender)" or "model.matrix(m1)" or even "rockchalk::predictOMatic(m1)" (if your fitted model is named "m1" and your data is "mydata", that is.

   (e) (5pts) Make the most informative plot you can to represent the regression relationship. Write a paragraph to discuss the plot, and why you chose it. Compare, possibly, against a boxplot, scatterplot, or something else.

   (f) (5pts) Calculate the means for the 2 genders and conduct a t-test, like this

```
t.test(dbp ~ gender, data = mydata, var.equal = TRUE)
```

   Compare this result with the regression analysis. Discuss any insights you gather from this exercise.

5. (10pts) I am always a little bit unsure about how I ought to convey numeric variables. Lets focus on the iq variable and compare some types of graphs. Do I prefer a histogram? A box plot? A kernel density smoothing plot? Or maybe a "spike plot" (also known as lollipop plot). Create some graphs for the "iq" variable and compare/contrast the strengths and weaknesses. Where is the central tendency? How to represent it in a graph? Does this appear to be drawn from a Normally distributed data source? Prepare several options and try to persuade me that your favorite is the best.

   I'll help you get started with some code hints. My data frame is named "dat". You need to adjust parameters to meet your needs, I made it ugly on purpose:

```
hist(dat$iq,xlab="Intelligence Quotient", ylab="Observed Proportion", stat=
   TRUE, lwd=0.5, main="", breaks=7, xlim=c(0,150))
m1 <- mean(dat$iq)
s1 <- sd(dat$iq)
legend("topright", legend=c(paste("mean=",m1),paste("std.dev.=",s1)))
```

   Here's how I made the spike plot for comparison

```
t3 <- table("iq" = dat$iq)
t3df <- as.data.frame(t3, stringsAsFactors=FALSE)
plot(Freq ~ iq, data = t3df, type = "h")
points(t3df$iq, t3df$Freq)
## I believe prev same as points(Freq ~ iqn, data = t3df)
legend("topright", legend=c(paste("mean=",m1),paste("std.dev.=",s1)))
```