Professor Notes About the "Bivariate EDA - Quant" Homework

- You must provide labeled tables and figures to support your results and refer to these tables in your sentences.
- Do not use the word "correlation" unless you are specifically referring to "r." For example, you would NOT say "the correlation between suspended sediments and discharge is positive, linear, etc." In this case, it is better to replace the word "correlation" with "relationship."
- Make sure to use `xlab=` and `ylab=` to provide better labels for the x- and y-axes, respectively, on your scatterplots.
- You must explicitly state where the outliers are located. In this case, it is adequate to note that they are "in the upper-left" corner of the plot. Alternatively, you could note the approximate coordinates of the points.
- Note the sentences that explicitly state whether the correlation coefficient (r) could be used to assess strength or not.
- It is correct to not calculate or report the correlation coefficient because of the presence of outliers. However, you still need to comment on the strength of the relationship. Your comment will be more subjective based on your interpretation of the clustering of the points but it still needs to be made.
- Computing the correlation in the last question is inappropriate because of the outliers in the data. If you were to compute the correlation, then you would need to use the code below, including the `use="pairwise.complete.obs"` argument because there are missing data in the *SuspSed* variable. **Again, this is inappropriate in this situation**.

```
> corr(~SuspSed+DschrgCFS,data=d,use="pairwise.complete.obs")
```

Animal Fat and Breast Cancer

The relationship between age-adjusted death rate and animal fat intake is positive, linear, absent of outliers, and very strong (r=0.949; Figure on homework handout). It was valid to assess strength with the correlation coefficient because of the linear form and lack of outliers.

North Fish Creek Discharge

The relationship between total suspended solids and discharge in Fish Creek is mostly positive and mostly linear (Figure 1). Several outliers are apparent, with the two at a total suspended solids greater than 800 mg/L and discharge less than 600 cfs and the one with a total suspended solids of 0 mg/L and a discharge at approximately 250 cfs being most apparent (Figure 1). The strength, excluding the outliers, is moderately strong. I did not compute a correlation coefficient because of the presence of the outliers.

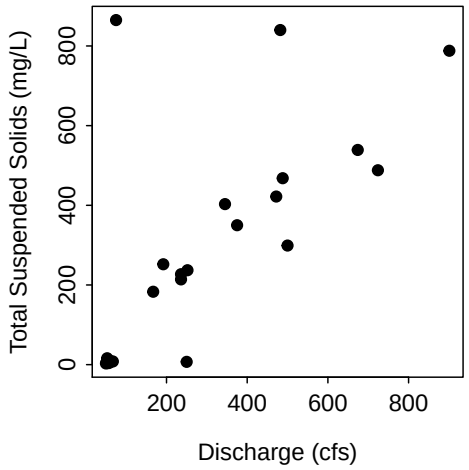

Figure 1. Scatterplot of total dissolved solids versus discharge in Fish Creek.

R Appendix

```
library(NCStats)
setwd('C:/aaaWork/Books/IntroStats/HW/')
d <- read.csv("FishCrNWaterQuality.csv")
plot(SuspSed~DschrgCFS,data=d,pch=19,xlab="Discharge (cfs)",
     ylab="Total Suspended Solids (mg/L)")
```