Supplementary Material F for "Beyond Normal: Preparing Undergraduates for the Work Force in a Statistical Consulting Capstone" by Smucker and Bailer

F.1 Motivation

Reservoirs store water and are critical for human activities like drinking, irrigation, recreation, and flood control. Over time, these bodies of water tend to accrue sediment due to a variety of factors, and this buildup reduces reservoir capacity. Of interest to geographers, then, is the rate at which the sediment accumulates in reservoirs around the U.S., in order to monitor erosion and evaluate water supplies.

F.2 Data Description

Dr. Bill Renwick, Professor of Geography at Miami University, was the client for this project in early 2013. The publicly-available dataset of interest included about 3,900 observations on roughly 1,900 different reservoirs, collected between 1755 and 1992. The desired analysis was somewhat preliminary, since some new data was expected to be made available soon. Within the dataset, five variables are of interest (Table F.1).

A measurement in this dataset specified a particular reservoir in a particular region (HUC2; Figure F.1), and included an estimated sedimentation rate that was calculated based upon estimated reservoir volumes measured at two different times. A sedimentation rate was associated with the midpoint between the beginning and ending measurements.

F.3 Problem Statement

There is a general assumption among reservoir managers that modern sedimentation rates are more-or-less unchanged when compared to the past. However, some hypothesize that due to improved environmental conservation, sedimentation rates have decreased. The goal, as communicated by Dr.

Renwick, was to model the sedimentation rate as a function of year, to determine if there was a significant change in sedimentation rates across different regions within the United States.

Table F.1. List of variables, descriptions, and purpose of each variable in the model. Adapted from STA 475 written report to client in spring 2013.

Variable Name	Description	Purpose in Model
SedRate (S)	Sedimentation yield (in cubic meters of sediment per square kilometers of drainage area per year)	Response
MidYear (M)	Midpoint between beginning and ending measurements	Predictor
HUC2 (R)	Geographic region (1 to 18)	Predictor (Categorical)
RESSED ID	Reservoir identification number	Correlation structure
Duration (D)	Time between beginning and ending measurements (in years)	Weighting

F.4 Solution

The challenges presented were considerable. The dataset was messy, including duplicate and conflicting observations that had to be resolved. There were also at least four critical complications that precluded a straightforward, standard analysis of sedimentation rate regressed on time, for each region.

First, the response was highly skewed due to the natural bound of 0 on the sedimentation rate. This issue was largely remediated by a log transformation (though, since the bound of 0 was not inviolable, nonpositive observations were omitted based on the argument made by Dr. Renwick that it is unusual for a reservoir to naturally exhibit a negative sedimentation rate and more likely to be the result of a man-

made intervention such as dredging). Scatterplots of the log-transformed sedimentation rates reveal many regions with apparently increasing trends (Figure F.2).

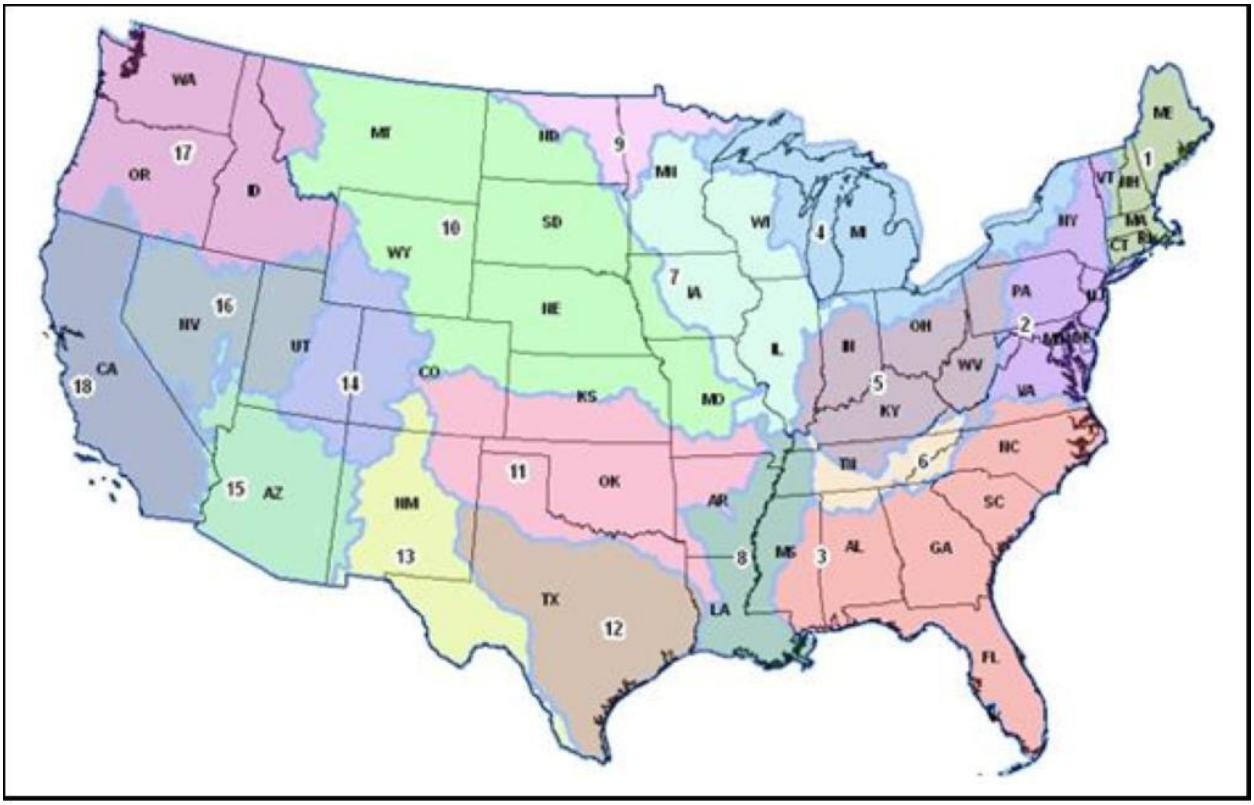

Figure F.1. Hydrologic Unit Code (HUC2) Map (Jones et al., 2010). Each shading indicates one of the 18 regions.

The second difficulty had to do with how the sedimentation rate of a reservoir was measured: At two different times, sometimes years apart, the volume of a particular reservoir was calculated and the two measurements used to estimate the sedimentation rate. As is clear from Figure F.3, rates are more volatile when the two measurements were taken close together and less variable when more time had elapsed between the two measurements. This attenuation of variability is, indeed, what was observed (a pattern that persists even when the sedimentation rate was log-transformed), which suggests that observations should be weighted as a function of the duration.

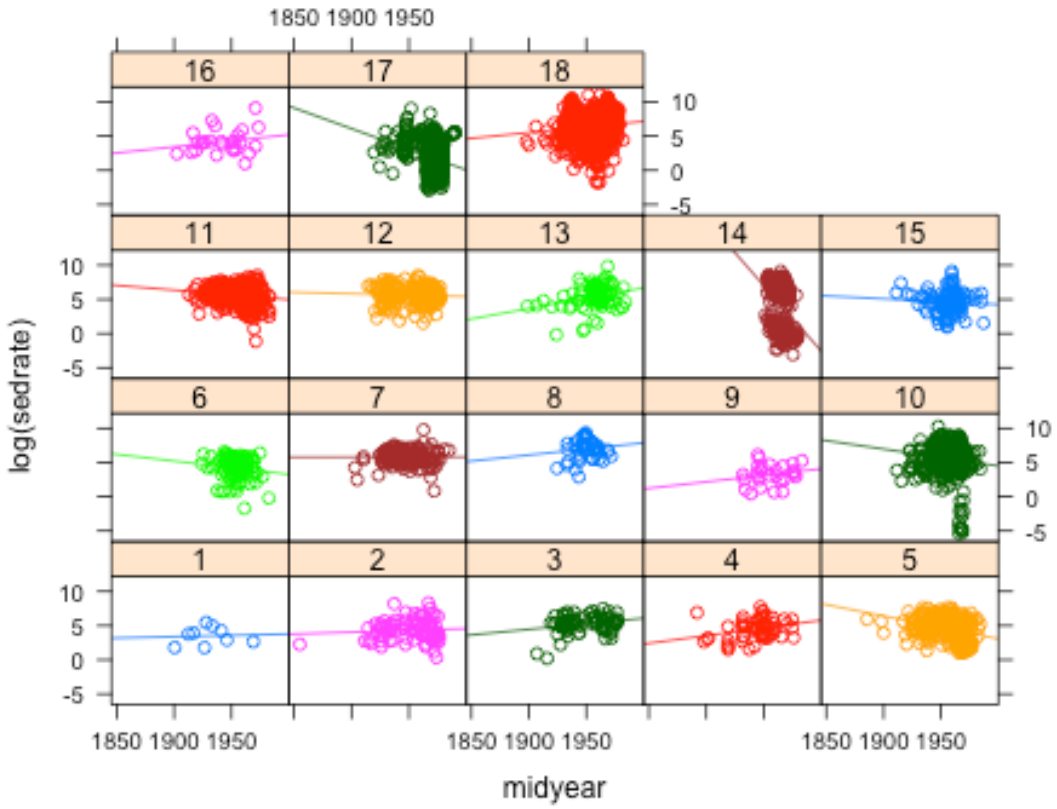

Figure F.2. Scatterplots of the log of the sedimentation rates versus MidYear (with simple linear regression line superimposed), for regions 1-18. Taken from STA 475 written report to client in spring 2013.

The third important complication is that many reservoirs were measured multiple times, meaning that the correlation between observations on the same reservoir needed to be incorporated into the model. (The simplifying assumption was made, however, that observations from different reservoirs were independent.) None of the students in the class had ever fit a model with such structure.

Finally, multiple comparisons was an issue that needed to be addressed, since there were 18 regions for which inference was desired.

The model fit was

$$\log(S_{jk}) = \sum_{i=1}^{18} \beta_i R_{ijk} + \sum_{i=1}^{18} \beta_{i+18} R_{ijk} M_{jk} + \varepsilon_{jk}$$

where

- j represents reservoir j and k represents the kth measurement taken on the jth reservoir;
- R_{ijk} is an indicator variable for Region i;
- errors are $\varepsilon_{jk} \sim N\left(0, \sigma^2 |D_{jk}|^{2\delta}\right)$, with D_{jk} the difference between the beginning and end of the sedimentation rate measurement (the "duration"), and σ^2 and δ are parameters estimated from the data.
- correlation between multiple measurements on reservoir j is $corr(\varepsilon_{jk}, \varepsilon_{jk'}) = \phi^s$ where ϕ is estimated from the data and s is the amount of time between measurements.

The model was parameterized to allow for direct hypothesis tests on each region slope parameter in the form of $H_0: \beta_i = 0$ vs. $H_a: \beta_i \neq 0$ for *i=19,20,…,36*. It was fit using the `gls` function in the `nlme` R package (Pinheiro et al., 2014).

The complications in the data posed substantial challenges for the undergraduate consultants. Clearly, the necessary procedure was beyond the material that the students had studied; in fact, they were not even equipped to *identify* all of the complications. Thus, the instructor highlighted the necessary sort of model, briefly introduced it, and suggested some possible software implementations that could fit it. After the brief introduction, the students had to research the details and how to use the appropriate R package. The complex data required them to carefully learn several aspects of statistical modeling with which they were unfamiliar, while concurrently building up the model in an unfamiliar R function. Then, they had to demonstrate their adequate knowledge by writing and orally communicating clearly about the project.

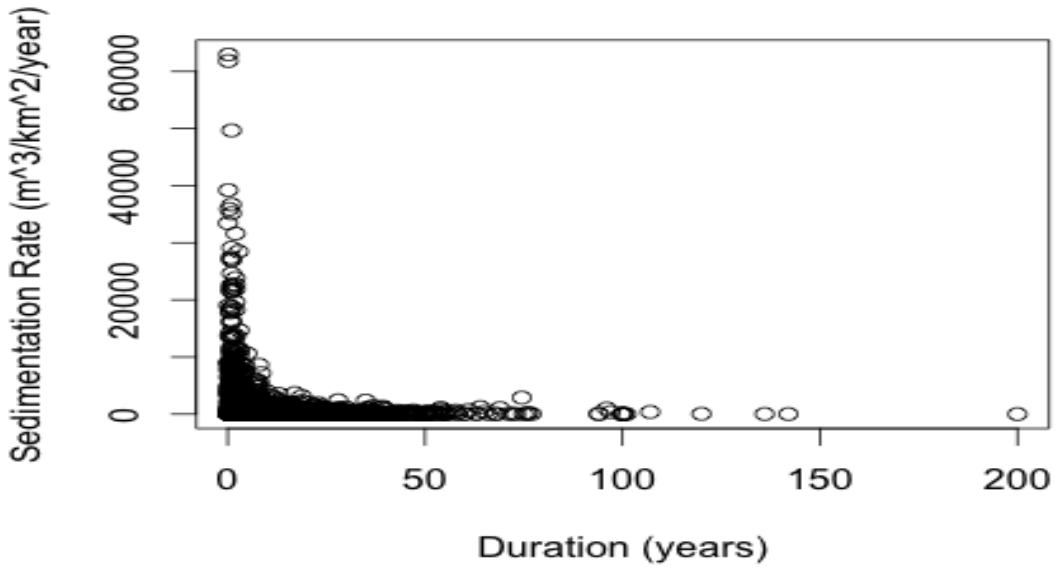

Figure F.3. Sedimentation rate versus the duration used to calculate the sedimentation rate. Taken from STA 475 written report to client in spring 2013.

F.5 Results

Once the complications described in the previous section were accounted for, the resulting fitted model produced a standardized residuals vs. fitted plot (Figure F.4) that gave a reasonable level of confidence in the subsequent inference. Upon fitting the model, there were eight regions with p-values less than 0.05, indicating possible changes in sedimentation rates over time. However, once the p-values were adjusted to account for the multiple hypothesis tests, four regions (Mid-Atlantic; South-Atlantic Gulf; Rio Grande; California) exhibited a slope parameter significantly different from 0. The procedure due to Holm (Holm, 1979), a method to control the family-wise error rate that is more powerful than the Bonferroni correction, was used to adjust for multiple tests.

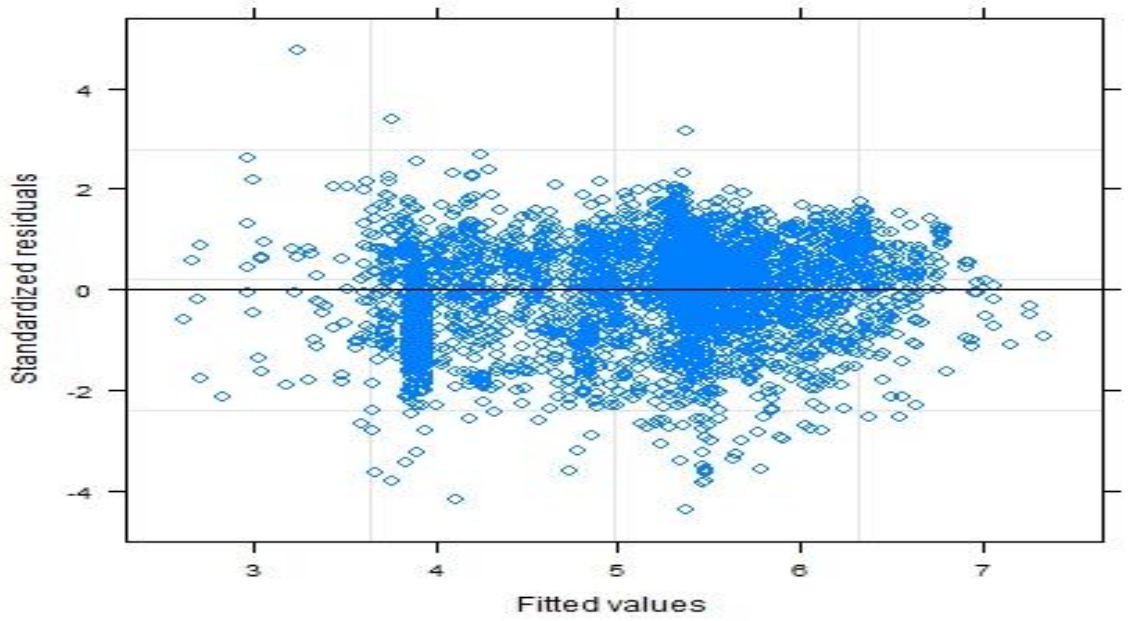

Figure F.4. Standardized residuals vs. fitted values for the final model. Taken from STA 475 written report to client in spring 2013.

Interestingly, each of the four significant slope parameters were positive, implying that the sedimentation rate was *increasing* across time, in contrast to the initial hypothesis. Though the primary objective of the analysis was to establish which relationships exist and the direction of those existing relationships, the slope estimates can also be interpreted. For instance, the significant slope estimate for the Mid-Atlantic region was 0.017, which can be interpreted roughly as follows: For this region, a one year increase is associated with an estimated change in the sedimentation rate of about a factor of $e^{0.017} = 1.017$, i.e. a 1.7% change (more precisely, the estimated change is in the *median* of the sedimentation rate, though this was not specified in the student report).

F.6 Limitations

In most consulting projects, there are caveats that temper confidence in the results. Here, one limitation was the elimination of the sedimentation rates less than or equal to 0, in order to facilitate the log transformation. Though there were scientific justifications for truncating the data in this way, it did reduce

the sample size by several hundred. A possible work-around would have been to add a constant to all responses in order to ensure that they all were positive. This would not have affected the slope parameter estimates and would have allowed the use of the entire dataset.

Another limitation was that correlation between neighboring reservoirs was ignored. Though the HUC2 region predictor accounts in a crude way for similarities in regions, it is possible that reservoirs in close proximity might be associated with each other. This possible correlation structure could have been accounted for by incorporating a spatial covariance component, but likely would have made the project even more challenging.

References

Holm, S. Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2) (1979), 65-70.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2014), *nlme: Linear and Nonlinear Mixed Effects Models*. Available at http://CRAN.R-project.org/package=nlme.