CPH 931 — Fall 2009 — Dr. Charnigo

Written Assignment 2

Written Assignment 2 is due on Friday 02 October at the end of class.

Please visit {http://lib.stat.cmu.edu/DASL/Stories/HealthyBreakfast.html} and obtain the cereal data by following the link to {http://lib.stat.cmu.edu/DASL/Datafiles/Cereals.html}. Read these data into SAS. Note that SAS will not understand that "-1" represents a missing value, so you will need to run code like the following.

```
DATA CEREAL;
SET CEREAL;
IF CARBO = -1 THEN CARBO = .;
[[more lines for other variables]];
RUN;
```

In addition, please define variables representing calories from fat, calories from protein, calories from carbohydrates, and calories from sugars.

```
DATA CEREAL;
SET CEREAL;
CALFAT = 9*FAT;
CALPRO = 4*PROTEIN;
CALCAR = 4*CARBO;
CALSUG = 4*SUGARS;
RUN;
```

Finally, create an alternate version of the rating variable in which there is a "mistake", namely a rating of 507.64999 for Cheerios.

```
DATA CEREAL;
SET CEREAL;
RATINGALT = RATING;
IF NAME = 'Cheerios' THEN RATINGALT = 507.64999;
RUN;
```

In exercises 1 through 4, you do not need to take any action regarding the problem of missing values. In exercise 5, you will address the problem of missing values through multiple imputation.

[20] 1. Consider two linear regression models: (i) RATING is the response variable and CALFAT is the sole explanatory variable; (ii) RATINGALT is the response variable and CALFAT is the sole explanatory variable.

[10] a. Fit models (i) and (ii) using ordinary least squares. Then fit models (i) and (ii) using M estimation. Construct a table that compares the four fitted models with respect to the estimated slope and its corresponding p-value.

[10] b. Are the two sets of results for model (i) similar? How about the two sets of results for model (ii)? In each instance, explain why you think the results turned out to be similar or dissimilar.

[20] 2. Once again, let RATING be the response variable and CALFAT be the sole explanatory variable. However, let us no longer assume that the expected value of RATING is a linear function of CALFAT.

[10] a. Apply LOESS smoothing to obtain a plot of the estimated expected value of RATING as a (possibly nonlinear) function of CALFAT. Please submit this plot for my inspection.

[10] b. Describe the pattern revealed by LOESS smoothing. In this instance, do you think that a linearity assumption is completely reasonable, somewhat reasonable, somewhat unreasonable, or completely unreasonable?

[20] 3. Consider a linear regression model in which RATING is the response variable and the explanatory variables include CALFAT, CALPRO, CALCAR, CALSUG, CALORIES, SODIUM, FIBER, POTASS, and VITAMINS.

[10] a. Report variance inflation factors for the model fitted using ordinary least squares. Comment on the nature of the multicollinearity, if any.

[10] b. Construct a ridge trace that displays coefficient estimates for values of the ridge parameter between 0 and 0.02 by increments of 0.001. Please submit this plot for my inspection. Then refit the model using ridge regression with a suitable value of the ridge parameter. Identify the major differences, if any, between ridge regression results and ordinary least squares results.

[20] 4. Consider a linear regression model in which RATING is the response variable and the explanatory variables include FAT, PROTEIN, CARBO, SUGARS, CALORIES, SODIUM, FIBER, POTASS, and VITAMINS.

[10] a. Fit the model using ordinary least squares. Report the coefficient estimates, standard errors, and p-values. Comment on the nature of the heteroscedasticity, if any.

[10] b. Refit the model using weighted least squares. Report the coefficient estimates, standard errors, and p-values. Identify the major differences, if any, between weighted least squares results and ordinary least squares results.

[20] 5. Generate five complete data sets through multiple imputation. Confine attention to RATING, FAT, PROTEIN, CARBO, SUGARS, CALORIES, SODIUM, FIBER, POTASS, and VITAMINS.

[10] a. Report the filled-in values of CARBO and SUGARS for Quaker Oatmeal.

[10] b. Using the five complete data sets, fit a linear regression model in which RATING is the response variable and the explanatory variables include FAT, PROTEIN, CARBO, SUGARS, CALORIES, SODIUM, FIBER, POTASS, and VITAMINS. Derive overall parameter estimates and standard errors from the five sets of results. Compare the overall parameter estimates and standard errors obtained here to those acquired in part a of exercise 4.