CPH 931 — Fall 2009 — Dr. Charnigo

Written Assignment 2 Solutions

1a. See Table 1.

Table 1:

Model, estimation method	Slope estimate	p-value
(i), ordinary least squares	-0.6347	0.0002
(ii), ordinary least squares	0.0161	0.9817
(i), M estimation	-0.5386	0.0010
(ii), M estimation	-0.5660	0.0005

1b. The two sets of results for model (i) are similar because there are no gross outlying values on RATING to disrupt estimation by ordinary least squares. The two sets of results for model (ii) are not similar because there is a gross outlying value on RATINGALT. This outlier disrupts estimation by ordinary least squares but has little effect on M estimation, since M estimation places a finite upper bound on the contribution each observation can make to the criterion whose minimization determines the estimates.

2a. [Show your LOESS plot.]

2b. The expected value of RATING decreases abruptly as CALFAT moves from 0 to 9 and then decreases more modestly as CALFAT moves from 9 to 45. On the other hand, the nonlinear pattern in the LOESS plot is substantially driven by two observations: All Bran with Extra Fiber, which has CALFAT = 0 and RATING = 93.70, and 100% Natural Bran, which has CALFAT = 45 and RATING = 33.98. If one or both of these observations were removed, then the pattern in the LOESS plot would have been close to linear. Overall, the linearity assumption is somewhat reasonable: the nonlinear pattern in the LOESS plot is not so strong that we feel compelled to abandon the benefits of parametric regression (ease of interpretation, confidence intervals, hypothesis tests).

3a. [Note: For the next three exercises, I modified the original instructions by removing VITAMINS from consideration.] The variance inflation factors are shown in Table 2. The explanatory variables with the largest variance inflation factors are CALORIES (17.16), CALSUG (15.96), and CALCAR (14.08). Borderline high variance inflation factors occur for FIBER (10.26) and POTASS (9.97). The main source of the multicollinearity is that adding up the different types of calories (from fat, protein, carbohydrates, and sugars) roughly approximates the total calories.

3b. [Show your ridge trace.] Choosing λ based on the ridge trace entails some subjectivity. I chose $\lambda = 0.02$ (and seriously considered expanding past the original range of values proposed for λ) because the curve for FIBER had not appeared to stabilize at any earlier value of λ. The coefficient estimates and standard errors are shown in Table 3. Compared to ordinary least squares, the coefficient estimates for CALFAT and CALSUG increased in magnitude by 32.4% and 28.0% respectively, while those for POTASS and CALCAR decreased in magnitude by 60.2% and 27.3% respectively. Also compared to ordinary least squares, the standard errors for CALSUG, CALORIES, CALCAR, and CALFAT decreased by 41.8%, 40.6%, 40.5%, and 29.0% respectively.

Table 2:

Variable	Variance inflation factor
Intercept	NA
CALFAT	5.16
CALPRO	2.74
CALCAR	14.08
CALSUG	15.96
calories	17.16
sodium	1.22
fiber	10.26
potass	9.97

Table 3:

Variable	Coefficient estimate	Standard error
Intercept	57.7750	1.0176
CALFAT	-0.2534	0.0228
CALPRO	0.7043	0.0447
CALCAR	0.1754	0.0184
CALSUG	-0.2641	0.0171
calories	-0.1750	0.0159
sodium	-0.0553	0.0018
fiber	2.7381	0.1345
potass	-0.0129	0.0046

4a. The ordinary least squares results are reported in Table 4. The plot of residuals against fitted values shows that there are six (negative) residuals more than twice as large in magnitude as the rest. They correspond to fitted values between 30 and 50, which are neither extremely small nor extremely large. So, there is heteroscedasticity, although not of the type where the error variance increases with the mean response.

4b. The weighted least squares results, using weights inversely proportional to the squares of linearly smoothed absolute values of ordinary residuals against fitted values from part a, are reported in Table 5. There are no remarkable changes from the ordinary least squares results, which is not surprising because this strategy for choosing weights is designed to work when the error variance increases with the mean response. Results using weights inversely proportional to the squares of quadratically smoothed absolute values of ordinary residuals against fitted values from part a are reported in Table 6. The only remarkable change from the ordinary least squares results is that the standard error for FIBER has decreased by 25.5%.

5a. My filled-in values of CARBO and SUGARS for Quaker Oatmeal are shown in Table 7. Since there is an element of randomness to multiple imputation, your results may be different.

Table 4:

Variable	Coefficient estimate	Standard error	p-value
Intercept	56.2544	0.9838	<.0001
FAT	-1.7224	0.2892	<.0001
PROTEIN	3.1266	0.1973	<.0001
CARBO	0.9655	0.1236	<.0001
SUGARS	-0.8252	0.1175	<.0001
calories	-0.2163	0.0268	<.0001
sodium	-0.0575	0.0017	<.0001
fiber	3.3588	0.1695	<.0001
potass	-0.0325	0.0057	<.0001

Table 5:

Variable	Coefficient estimate	Standard error	p-value
Intercept	56.3045	0.9979	<.0001
FAT	-1.7051	0.2906	<.0001
PROTEIN	3.1250	0.1970	<.0001
CARBO	0.9591	0.1245	<.0001
SUGARS	-0.8241	0.1179	<.0001
calories	-0.2164	0.0269	<.0001
sodium	-0.0571	0.0018	<.0001
fiber	3.3746	0.1781	<.0001
potass	-0.0332	0.0058	<.0001

5b. My overall parameter estimates and standard errors from the five sets of results are shown in Table 8. Since there is an element of randomness to multiple imputation, your multiple imputation results may be different. The multiple imputation results in Table 8 exhibit no remarkable differences from the ordinary least squares results in Table 4.

Table 6:

Variable	Coefficient estimate	Standard error	p-value
Intercept	56.3911	0.9086	<.0001
FAT	-1.7239	0.2696	<.0001
PROTEIN	3.1088	0.1770	<.0001
CARBO	0.9630	0.1166	<.0001
SUGARS	-0.8341	0.1095	<.0001
calories	-0.2146	0.0248	<.0001
sodium	-0.0577	0.0015	<.0001
fiber	3.3547	0.1262	<.0001
potass	-0.0332	0.0048	<.0001

Table 7:

Imputation Number	Filled-in CARBO	Filled-in SUGARS
1	5.9078	9.3418
2	6.2970	9.6260
3	5.6956	8.8093
4	8.1684	8.1633
5	6.9972	9.4313

Table 8:

Variable	Coefficient estimate	Standard error
Intercept	56.1523	1.0028
FAT	-1.6672	0.2903
PROTEIN	3.1416	0.2049
CARBO	0.9889	0.1251
SUGARS	-0.8154	0.1195
calories	-0.2184	0.0272
sodium	-0.0577	0.0018
fiber	3.4237	0.1692
potass	-0.0350	0.0056