This data involves predicting a person's weight (the response variable y), based on a number of factors (x1 through x8 below).

- y: weight (pounds)
- x1: height (inches)
- x2: gender (male = 1; female = 0)
- x3: average number of meat servings (4 oz.) consumed each day
- x4: average number of fruit or vegetable servings consumed each day
- x5: age (years)
- x6: carries cell phone regularly (yes = 1; no = 0)
- x7: last digit of home phone number
- x8: will Cubs win the World Series? (Yes = 1; No = 0)

Two separate sets of data are involved: The test set (in Lab2Test.csv) are the preceding variables collected for 22 students in 2019 (each row = one person; each column = one variable). The training set (in Lab2Training.csv) are the same variables collected for 136 students in other years. You will fit a regression model to the training data. Then, you will use the model fitted to the training data to see how well you can predict the responses for the data in the test data. Do NOT include the test data when fitting the regression model.

Test.csv:

y	x1	x2	x3	x4	x5	x6	x7	x8
173	73.6	1	4	1	21	1	2	1
185	72	1	3	2	20	1	7	0
135	65	0	2	4	20	1	2	0
127	64	0	3	2	20	1	4	0
135	68	1	4	3	20	1	4	0
170	72	1	4	4	22	0	8	1
115	61	0	2	0	20	1	3	0
150	70	1	4	0	20	1	4	0
150	72	1	4	1	21	1	9	1
155	73	1	4	1	21	1	7	0
115	62	0	1	5	22	1	7	0
160	75	1	4	2	22	1	9	1
135	65	1	1.5	2	23	1	6	0
165	71	1	1	3	21	1	4	1
123	67	0	1.5	4	20	1	8	0

177	73	1	1.5	2	22	1	1	1
165	69	1	3	2	21	1	7	0
155	70	1	3	3	21	1	8	1
180	75	1	3	3	20	1	2	0
175	70	1	4	1	20	1	8	1
165	68	1	4	1	20	1	6	0
180	71	1	4	3	20	1	8	0

Training.csv:

y	x1	x2	x3	x4	x5	x6	x7	x8
155	73	1	4	1	21	0	2	1
125	70	1	2	3	21	1	9	0
195	72	1	4	5	20	1	6	1
185	72	1	2	5	20	1	9	1
123	62	0	1	3	20	1	4	1
155	71	1	2	0	20	1	7	1
120	65.5	0	0	3	19	1	7	1
170	71	1	2	0.5	21	1	3	1
155	71.5	1	2	1	21	1	3	1
150	68.7747036	1	2	1	23	1	3	1
112	65	0	2	2	21	1	1	1
160	68	1	4	0.5	20	1	6	1
130	66	1	4	0.5	20	1	8	1
160	65	1	4	2	20	1	4	1
139	70	1	3	1	23	0	7	1
176	73	1	2	3	25	1	8	1
109	64	0	2	5	21	1	6	1
215	75	1	3	2	20	1	7	1
200	74	1	3	4	22	1	8	1
165	70	1	3	3	21	1	4	1
160	69	1	5	2	20	1	9	1
100	62	0	2	3	20	1	1	1
115	66	0	2	2	21	1	5	1
165	72	1	5	2	23	1	1	1
165	72	1	3	2	20	1	6	1
125	63	0	2	5	20	1	4	1
182	62	1	2	1	20	1	6	0

130	66	0	1	2	20	1	2	0
113	60	0	2	2	21	0	4	1
160	72	1	4	4	21	1	5	0
165	73	1	2	0	21	1	1	0
215	68	1	1	0	21	1	7	0
160	68	1	3	2	21	1	4	1
185	71	1	5	1	21	1	0	1
160	71	1	2	1	21	0	4	1
180	74	1	1	1	20	1	2	1
165	75	1	2	2	20	1	4	0
170	74	1	1	1	27	1	6	1
180	71	1	2	2	20	1	5	1
160	67	1	2	1	20	1	5	1
119	61	0	2	3	19	1	4	1
160	71	1	6	3	20	1	1	1
165	70	1	5	5	29	1	3	1
218	73	1	6	5	21	1	7	1
200	78	1	3	2	19	0	6	1
230	73	1	1	1	30	1	0	0
155	66	1	2	1	20	1	9	0
120	66	0	2	1	19	1	0	0
150	71	0	1	4	22	1	0	0
170	71	1	2	5	21	1	1	0
170	70	1	3	4	21	1	0	0
200	74	1	3	4	23	1	0	0
95	63	0	2	4	19	1	0	1
170	71	1	2	2	20	1	1	0
160	70	1	2	3	20	1	1	0
180	61	1	3	5	21	1	8	1
140	71	1	3	4	19	1	2	0
185	74	1	3	2	20	1	4	1
145	70	1	3	3	23	1	3	1
145	67	1	3.5	0.5	22	1	2	1
170	72	1	2	4	21	1	3	0
215	71	1	3	1	21	1	4	0
105	62	0	1	0	19	1	8	1
170	74	1	3	4	20	1	4	0
200	73	1	2	2	20	1	3	0
160	70	1	4	6	19	1	1	1

200	74	1	4	3	20	1	5	0
120	64	0	1	2	21	1	8	1
168	71	1	2	2	21	0	2	0
175	70	1	2	4	24	1	3	0
130	64	0	2	3	24	1	4	0
127	62	0	1	4	24	1	9	1
150	66	0	2	4	20	1	3	1
155	67	1	4	1	20	1	1	0
148	67	1	4	3	20	1	0	1
155	68	0	3	2	20	1	8	1
163	68	1	2	3	20	1	0	0
160	70	1	3	2	20	1	8	0
160	69	1	2	2	22	1	8	1
150	67	1	1.5	1.5	22	1	1	0
140	65	1	4	5	20	1	5	1
135	68	1	2	3	23	1	0	1
176	73	1	7	2	21	1	6	1
140	61	0	2	4	20	1	8	0
110	60	0	2	1	20	1	0	1
160	71	1	3	1	20	1	9	0
215	72	1	3	3	20	1	2	0
155	72	1	0	4	19	1	0	1
135	70	1	0	2	19	1	2	0
200	72	1	1	2	21	1	1	1
165	68	1	2	2	21	1	9	1
123	63	0	2	4	19	1	9	1
150	69	0	2	4	20	1	8	0
175	72	1	4	4	20	1	3	1
150	68	1	4	4	20	1	8	1
165	72	1	4	6	20	1	3	0
168	71	1	8	6	20	1	6	0
145	64	1	2	2	19	1	4	0
138	65	1	0	8	20	1	9	0
170	73	1	3	4	20	1	7	0
200	70	1	3	1	21	1	0	1
180	73	1	3	2	22	1	9	0
165	70	1	4	2	22	1	3	0
175	73	1	2	3	21	1	6	1
165	71	1	1	1	20	1	2	0

140	69	1	2	1	20	1	7	0
110	61	0	0	1	20	1	1	0
151	67	1	2	4	23	1	8	0
125	67	0	2	1	23	1	3	0
173	74	1	2	2	21	1	1	0
180	72	1	3	4	22	1	2	0
115	68	0	2	6	21	1	1	0
160	70	1	3	3.5	21	1	9	0
150	70	1	3	3	20	1	6	0
148	71	1	4	3	21	1	5	1
190	75	1	5	1	20	1	7	0
170	71	1	5	5	21	1	6	1
130	65	0	1	2	20	1	6	1
130	71	1	2.15	3	24	1	7	0
163	67	1	3	4	23	1	1	0
200	76	1	5	4	20	1	1	0
170	69	1	3	3.41	22	1	9	0
155	71	1	2	2	20	1	5	0
115	64	0	2	3	21	1	3	0
118	66	1	2	2	20	1	4	0
159	71	1	2	1	24	1	8	1
125	63	0	2	2	21	1	5	0
150	72	1	2	3	22	1	2	0
115	61	0	2	4	21	1	7	0
160	68	1	2	2	22	1	7	0
110	64	0	2	4	20	1	0	0
154	66	1	2	2	19	1	0	0
170	73	1	3	4	20	1	7	0
200	70	1	3	1	21	1	0	1
180	73	1	3	2	22	1	9	0
165	70	1	4	2	22	1	3	0

1. For the training data, construct and interpret a scatter-plot matrix of all variables (response and predictors). Are any relationships between y and the predictor variables apparent?

2. Using the training data, fit a regression model that includes all eight predictor variables. Use this fitted model to predict the weights for everyone in the test set. Construct a scatter

plot of y versus ˆy (y on the vertical axis) for all observations in the training set and in the test set together (i.e., a single plot), but distinguish the training group and the test group using two different symbols. Discuss the significance of what you see.

3. Calculate 95% prediction intervals for the weights of everyone in the test set, and also calculate the prediction errors. The ”prediction errors” are defined as the residual errors (y− ˆy) for predicting the test data, using the coefficients estimated from the training data. Hence, you have one prediction error for each person in the test set. Note that you are NOT supposed to fit a new model to the test data. Use the same model that you fit to the training data in Problem 2 but apply it to predicting the test data. Do the actual response values for the test data seem consistent with the prediction intervals? Explain.

4. Plot the standardized residuals versus the fitted values for the training data. Interpret the results.

5. Repeat Problems 2—4, but this time include only the two predictor variables x1 and x2 in the regression model. Construct side-by-side box plots of the two sets of prediction errors for the test data (one set using the eight-predictor model and the second set using the two-predictor model). Discuss what you see.

6. Using only the training data, conduct an ”Extra Sum of Squares” F-test of whether x3 through x8, together, have an effect on y.

7. Based on Problems 2—6, would you recommend the 8-predictor model or the 2-predictor model for predicting the weight of some new person (say from a different class)? Provide a detailed justification for your answer, incorporating all relevant findings from Problems 2-6 into your discussion. Use quantitative arguments, as well as qualitative arguments based on the plots (e.g. from Problems 2 and 5) that you constructed. If evidence from different parts of Problems 1 through 6 contradicts each other, you will have to determine which to weigh more heavily. Finally, give some concluding remarks summarizing and generalizing what you found in this lab. Use the following data: