David Smith
STAT 4630
Homework 2

Part A

See R file.

Part B

Question 1.

Note that this process is simply ridge regression.

a) As we increase λ, the training RSS will steadily increase, since we are constraining the coefficients, causing the model to deviate from the least-squares regression line.

b) As we increase λ, the test RSS will decrease initially, and then eventually start increasing in a U shape. This is because the test MSE is a sum of the variance, the squared bias, sand the irreducible error. As we will show, the variance steadily decreases, the squared bias steadily increases, and the irreducible error remains constant as we increase λ, so the sum of the three will at first decrease, then eventually increase.

c, d) As we increase λ, the variance will steadily decrease, since the flexibility of the fit decreases. For the same reason, the bias will steadily increase, as we are constraining our model.

e) As we increase λ, the irreducible error will remain constant, as it does not depend upon the chosen model.

Question 2.

a) When $x \leq \xi$, we have $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4(0) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$. Thus, $f_1(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

b) When $x > \xi$, we have $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4(x - \xi)^3 = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4(x^3 - 3x^2\xi + 3x\xi^2 - \xi^3) = (\beta_0 - \beta_4\xi^3) + (\beta_1 + 3\beta_4\xi^2)x + (\beta_2 - 3\beta_4\xi)x^2 + (\beta_3 + \beta_4)x^3$. Thus, $f_2(x) = (\beta_0 - \beta_4\xi^3) + (\beta_1 + 3\beta_4\xi^2)x + (\beta_2 - 3\beta_4\xi)x^2 + (\beta_3 + \beta_4)x^3$.

c) We have that $f_2(\xi) = (\beta_0 - \beta_4\xi^3) + (\beta_1 + 3\beta_4\xi^2)\xi + (\beta_2 - 3\beta_4\xi)\xi^2 + (\beta_3 + \beta_4)\xi^3 = \beta_0 + \beta_1\xi + \beta_2\xi^2 + (-\beta_4 + 3\beta_4 - 3\beta_4 + \beta_3 + \beta_4)\xi^3 = \beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3 = f_1(\xi)$.

d) We have that $f_1'(x) = \beta_1 + 2\beta_2 x + 3\beta_3 x^2$ and $f_2'(x) = (\beta_1 + 3\beta_4\xi^2) + 2(\beta_2 - 3\beta_4\xi)x + 3(\beta_3 + \beta_4)x^2$. Thus, $f_2'(\xi) = (\beta_1 + 3\beta_4\xi^2) + 2(\beta_2 - 3\beta_4\xi)\xi + 3(\beta_3 + \beta_4)\xi^2 = \beta_1 + 2\beta_2\xi + (3\beta_4 - 6\beta_4 + 3\beta_3 + 3\beta_4)\xi^2 = \beta_1 + 2\beta_2\xi + 3\beta_3\xi^2 = f_1'(\xi)$.

e) We have that $f_1''(x) = 2\beta_2 + 6\beta_3 x$ and $f_2''(x) = 2(\beta_2 - 3\beta_4\xi) + 6(\beta_3 + \beta_4)x$. Thus, $f_2''(\xi) = 2(\beta_2 - 3\beta_4\xi) + 6(\beta_3 + \beta_4)\xi = 2\beta_2 - 6\beta_4\xi + 6\beta_3\xi + 6\beta_4\xi = 2\beta_2 + 6\beta_3\xi = f_1''(\xi)$.

Question 3.

1. See R file.

2. I decided to use lasso regression to predict y. In order to come to that conclusion, I used a validation set approach where I trained ordinary least squares, ridge, and lasso regression on most of the training data and tested each on a small portion. First, I divided the training data into random training and test subsets, with the test subset containing 10% of the training data. I then ran ordinary least squares, ridge, and lasso regression on the training subset of the training data, predicted for the test subset using each fit, and calculated the root mean square error for each, using the predictions and the known y values of the test subset. Since the RMSE for lasso regression was the smallest by a significant margin, I chose to use lasso regression to predict y values for the test data. It makes sense that lasso regression would result in vastly better predictions that ordinary least squares regression since the data is very high-dimensional. With ridge and lasso regression, I used cross validation to choose the best value of the tuning parameter λ.

3. Rdata file submitted.

Question 4.

1. See R file.

2. I decided to use a smoothing spline to predict the total value of oil imports. First, I observed that it wouldn't make sense to use local regression, since I was predicting for a wide range of CPI values, or generalized additive models, since it is a more general form of the other methods explained in Chapter 7. After graphing the data using a scatterplot, the y values were seen to be changing rapidly, and increasing quickly as the CPI grew large, indicating that step functions would not yield good results. Thus, I limited my methods to polynomial regression, regression splines, natural splines, and smoothing splines. I tested each using a validation set approach, where I divided the training data into a training subset, consisting of 90% of the original data, and a test subset, containing the remaining 10% of the data. I ran each of polynomial regression, regression splines, natural splines, and smoothing splines on the training subset, predicted for the test subset using each fit, and calculated the root mean square error for each, using the predictions and the known y values of the test subset. Since the RMSE for smoothing splines was the smallest, I chose to use it to predict the total value of oil imports for the test data. For polynomial regression, the scatterplot indicated that while there were fluctuations in the data, it generally followed a cubic curve, so I chose to use polynomials of at most degree 3 for the regression. I trained and predicted using a linear, quadratic, and cubic polynomial regression separately, and found the RMSE for each. For regression and natural splines, I used the default of a cubic spline, and used three evenly spaced knots for each, as is standard. The data didn't indicate the necessity to change these values, especially the cubic degree of the polynomials, and I couldn't determine how to use cross validation to determine the optimal number of knots to use, so I stuck with the default, suspecting

that a smoothing spline would work best anyway. Finally, I trained and predicted with a smoothing spline, using cross validation to determine the best value of λ.

3. Rdata file submitted.