1. In this problem we will examine the performance of ridge regression and Lasso regression on the `seatpos` data set from Faraway. The response variable, `hipcenter`, measures the position of the seated driver's hips in the car. The covariates are `Age`, `Weight`, `HtShoes` (height in shoes), `Ht` (height without shoes), `Seated` (seated height), `Arm`, `Thigh`, `Leg` (arm / thigh / lower leg length).

   (a) Examine the correlations among the covariates and comment on what you see, and how you might expect this to affect the regression.

   (b) Begin by standardizing each covariate so that we don't run into issues of how the code treats each covariate. Center each covariate to have zero mean, and then rescale it so that $\sum_i X_{ij}^2 = n$.

   (c) Now we run ridge regression.

```
library(MASS)
model = lm.ridge(hipcenter ~ Age+Weight+HtShoes+Ht+Seated+Arm+Thigh+Leg, lambda = ????)
betahat = ~~c(model$Inter, model$coef)~~ # intercept, then coefficients on the covariates
```

   coef (model)

   Note that by convention, the intercept term is not penalized by the ridge regression procedure. Run this at $\lambda = 0, 0.1, 1, 2, 5, 10, 20, 50,$ and comment on what you see for the fitted coefficients. In particular, what do you see happening to the coefficients on the covariates `HtShoes`, `Ht`, and `Seated`? Discuss.

   (d) Next we will use leave-one-out cross-validation to select a good value of λ. Fix a grid of λ values ranging between 0 and ~~20~~ 50 (you should use a very fine grid, e.g. 0, 0.1, 0.2, ...) For each data point $i = 1, \ldots, n$, run ridge with each λ value on the data set with point i removed, find $\hat{\beta}$, then get the leave-one-out error for predicting Y_i. Average the squared error over all n choices of i. Plot the leave-one-out error against λ and find the best λ value. How do the estimated coefficients compare against least squares? Based on your leave-one-out analysis, does ridge appear to offer substantial improvement of the prediction error?

   (e) Next we'll turn to the Lasso.

```
library(glmnet)
model = glmnet(cbind(Age,Weight,HtShoes,Ht,Seated,Arm,Thigh,Leg), y = hipcenter, lambda = ????)
betahat = c(model$a0,as.matrix(model$beta)) # intercept, then coefficients on the covariates
```

   Again, the intercept is not penalized. Run this at $\lambda = 0, 0.1, 1, 2, 5, 10, 20, 50,$ and comment on what you see for the fitted coefficients. In particular, what do you see happening to the coefficients on the covariates `HtShoes`, `Ht`, and `Seated`? Discuss. [Note: the range of λ values that works well for ridge, will not necessarily be the right range of values for Lasso—the two λ's are penalizing different functions and are not comparable.]

   (f) Run the leave-one-out analysis again, now with Lasso, and answer the same questions as above.

   (g) Finally, let's look at variability. Bootstrap the sample 1000 times and record $\hat{\beta}$ using (1) least squares, (2) Ridge (with the value of λ selected by the leave-one-out analysis for ridge), (3) Lasso (with the value of λ selected by the leave-one-out analysis for Lasso). Plot histograms of $\hat{\beta}_{\texttt{HtShoes}}$ and compare—what do you see? (Each of the three methods is its own plot.)

   (h) Next plot scatterplots of $(\hat{\beta}_{\texttt{HtShoes}}, \hat{\beta}_{\texttt{Ht}})$. Discuss what you see for each plot, in detail. (Each of the three methods is its own plot. Each plot has 1000 points, one for each bootstrapped sample.)

2. Faraway (1st edition) problem 8.4. (See the equation in the middle of the page on pg. 122, to see what is meant by a second order polynomial model with interaction term included.)

3. You are welcome to collaborate in pairs or groups of three on this problem; if you choose to work in a group, please list your collaborators in your handed in HW.

   Why does the Lasso lead to solutions $\hat{\beta}$ with values $\hat{\beta}_j$ that are *exactly* equal to zero? To study this, let's look at an extreme case—for very large values λ, the solution is actually all zeros. We will use the definition

$$\hat{\beta} = \arg\min \left\{ \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \sum_j |\beta_j| \right\}$$

   for the Lasso. Here $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$, and the optimization is over $\beta \in \mathbb{R}^p$.

(a) Preliminary step: prove that for any X and Y and β,

$$|\langle Y, X\beta\rangle| \leq \max_j |X_j^\top Y| \cdot \sum_j |\beta_j|,$$

where X_j is the jth column of the matrix X.

(b) Now suppose we choose an extremely large λ, satisfying $\lambda \geq \max_j |X_j^\top Y|$. Using the preliminary calculation above, prove that $\hat{\beta} = \mathbf{0}_p = (0, \ldots, 0)$. In other words, prove that for any $\beta \in \mathbb{R}^p$, we have

$$\text{Loss}(\beta) \geq \text{Loss}(\mathbf{0}_p)$$

where the loss function is

$$\text{Loss}(\beta) = \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda \sum_j |\beta_j|,$$

i.e. the function we're trying to minimize. If indeed $\text{Loss}(\beta) \geq \text{Loss}(\mathbf{0}_p)$ for every $\beta \in \mathbb{R}^p$ then this means that $\mathbf{0}_p$ is a minimizer (although we have not proved that it's the unique minimizer).

(c) Bonus question (this is completely optional): prove that $\mathbf{0}_p$ is the unique minimizer, i.e. if $\beta \neq \mathbf{0}_p$ then $\text{Loss}(\beta) > \text{Loss}(\mathbf{0}_p)$. Hint: one way to do this is to split into cases, depending on whether $\|X\beta\|_2 = 0$ or > 0.