Homework 6

Advanced Methods for Data Analysis (36-402)

Due Thursday February 28, 2019, at 6:00 PM

See the syllabus for general instructions regarding homework. Note that you should **always show all your work** and submit **exactly two files** (Rmd or R file as *code*, and knitted or scanned/merged pdf or html as *writeup*). Make sure that everything you submit is readable.

Problem 1: Bootstrapping a Cross-Validation

In Homework #5, we compared four models for predicting median house value from various predictors. Two of the models had R^2 and prediction errors that were very close, and it wasn't obvious that one model was better than the other. The bootstrap allows us to measure the uncertainty in the (cross-validation) mean-squared-error calculations to help decide whether one of the models is actually better than the other.

Part a

Read the two data files `housetrain.csv` and `housetest.csv` back into R and combine them into a single object using the `rbind()` function. This time, we are going to bootstrap the 5-fold cross-validation analysis for the models that were called Model 2 and Model 3 in Homework #5:

Model 2: A simple linear regression of `Median_house_value` on `Mean_household_income`.

Model 3: A multiple regression of `Median_house_value` on *both* `Median_household_income` and `Mean_household_income`.

Some of the code for 5-fold cross-validation from Homework #5 (with modifications) can be useful for this problem. Create $B = 200$ bootstrap samples each consisting of $n = 10605$ rows from the combined data set selected at random with replacement.

```
## some sample code
B <- 200; n <- nrow(dat)
boot_indices <- replicate(B, sample(1:n, n, replace=TRUE))
```

We will use the "resampling cases" form of the bootstrap (also called "resampling (X, Y) pairs" or nonparametric bootstrap).

For each bootstrap sample b, randomly divide the n observations into 5 disjoint sets of size 2121 each. Treating each of the 5 folds as test data and the other 4 as training data, calculate prediction error for each model and call the average of the 5 prediction errors for Model 2 $\widehat{MSE2}_b^*$. Call the average for Model 3 $\widehat{MSE3}_b^*$. Draw a histogram of the $\widehat{MSE2}_b^* - \widehat{MSE3}_b^*$ values. What visual evidence is there about whether one model is better?

Part b

Let $T_b^* = \widehat{MSE2}_b^* - \widehat{MSE3}_b^*$ for $b = 1, \ldots, 200$. Draw a normal q-q plot of the T_b^* values and add the `qqline`. Do they look like a sample of normal random variables? Treat $T_1^*, \ldots, T_{200}^*$ as a random sample of normal random variables and test the null hypothesis that $\mathbb{E}(T_j^*) = 0$. Does one of the models look better than the other now?

Problem 2: Kernel Regression and the Bootstrap

This problem uses part of a set of data on abalone fishing in Australia (taken from http://archive.ics.uci.edu/ml/machine-learning-databases/abalone); available as `fishdata.csv` on Canvas with 1528 observations on 2 variables, with header. Each observation corresponds to a particular caught male abalone, and the columns correspond to the following attributes:

Name	Data Type	Meas.	Description
Diameter	continuous	mm	perpendicular to length
Weight	continuous	grams	whole abalone

We will be predicting the `Weight` of male abalones from the diameter measurements. There are enough data values here to clutter a plot unless you use a small plotting symbol such as `pch="."`.

Part a

Plot the `Weight` versus `Diameter`. Does it look like a linear regression is likely to provide a good fit? Say why or why not.

Part b

Fit a kernel regression of `Weight` on `Diameter`. As usual, let $r(x)$ denote the true regression function $\mathbb{E}\{Y|X=x\}$ and $\hat{r}(x)$ its nonparametric estimate. For this part and subsequent parts, use the heuristic choice for the bandwidth used in the previous HWs: the standard deviation of the predictor divided by $n^{1/5}$, where n is the size of dataset.

Part c

In this problem, we will use the bootstrap to compute confidence intervals for $r(x)$ for each x from 100 to 650 in steps of 5 (111 different values of x). Call these x values `x0`. Use $B = 1000$ bootstrap samples of (`Diameter`,`Weight`) pairs drawn together from the empirical distribution of the data. (This corresponds to "resampling (X,Y) pairs" or "resampling cases" or nonparametric bootstrap in the language of bootstrapping regressions.) For each bootstrap sample b, fit a kernel regression as in part b, and predict the response `Weight` at each of the 111 values in `x0`. (We can call these $\hat{r}^*_b(x)$.)

For each x in `x0`, we are going to compute a 95% pivotal bootstrap confidence interval for $r(x)$. As you may recall from previous Statistics classes, a $1-\alpha$ confidence interval for a paremeter t_0 is a *random* interval that contains t_0 with probability $1-\alpha$. In other words, it is an interval that contains all the "plausible" values for t_0, where "plausible" means all values except those for which the probability of seeing the observed sample is less than or equal to α.

For each $x_i \in$ `x0`, you have B estimates $\hat{r}^*_b(x_i)$. Let $q_\tau(x_i)$ be the τ-quantile of the bootstrap distribution of $\hat{r}^*_b(x_i)$. A $1-\alpha$ pivotal bootstrap confidence interval fo $r(x_i)$ can be computed as

$$CI(x_i) = \left[2\hat{r}(x_i) - q_{\alpha/2}(x_i), 2\hat{r}(x_i) - q_{1-\alpha/2(x_i)}\right]$$

where $x_i \in$ `x0`.

Draw a plot with the original data and the estimated regression function $\hat{r}(x)$ for each x in `x0` added as a line. Finally, add to the plot the upper and lower endpoints of all of the pivotal bootstrap confidence intervals, using a different line type than used for $\hat{r}(x)$.

We will discuss in greater detail the pivotal bootstrap confidence interval in lecture on Tuesday. For now, take a look at Section 6.2.2 in Shalizi's book.

Part d (Extra Credit)

There are some x values where the confidence intervals are wider than they are for the other x values. Give a plausible explanation for this based on all of the analysis in the earlier parts of this problem.

Problem 3: Omitted Variables and Causal Regression

Return to the `cats` data that are available in `library(MASS)`. Check the lecture materials for some example calculations using these data. In particular, a linear regression of heart weight `Hwt` on body weight `Bwt` is explored. In this problem, we also consider predicting heart weight Y from body weight X, but we take into account the third variable `Sex` of the cats.

Part a

Fit the linear regression of `Hwt` on `Bwt`. Examine the residuals separately for male and female cats, and comment on what you see.

Examine the empirical distribution of body weight separately for male and female cats, and comment on what you see.

Part b

Treat `Sex` as an omitted variable that could be a potential confounder in studying the relation between `Hwt` and `Bwt`. Go back to the in-class discussion of causal regression (you may find previous HWs helpful as well). Assume that `Sex` is the only confounder, and estimate the causal regression function $\theta(x)$ of Y on X, where the argument x stands for different values of body weight.

Draw a scatter plot of `Hwt` (on vertical) versus `Bwt` with the following four lines added so that we can tell them apart:

(i) the simple linear regression of `Hwt` on `Bwt`,

(ii) the causal regression line,

(iii) the simple regression of `Hwt` on `Bwt` for Females only, and

(iv) the simple regression of `Hwt` on `Bwt` for Males only.

Comment on when, if ever, each line would be most appropriate for making predictions about heart weight from body weight.

Part c

Assume a model of the form

$$Y = \beta_{0,S} + \beta_{1,S} X + \varepsilon,$$

where S stands for `Sex`, and ε has a distribution that might depend on `Sex`, but is otherwise unspecified. You may have already fit this model earlier in this problem if you ran separate regressions for males and females. If not, do it now. For each sex separately, run a bootstrap by resampling residuals so that we can

examine the different distributions of $(\hat{\beta}_{0,\text{Male}}, \hat{\beta}_{1,\text{Male}})$ and $(\hat{\beta}_{0,\text{Female}}, \hat{\beta}_{1,\text{Female}})$. To do this, take a look at the code below (we will talk more about when to use which type of bootstrap in lecture next week):

```
## You will need to store the results and complete the code below
## datm and datf are the datasets partitioned by sex
## modelm and modelf are the linear regression models by sex
## betahat_null is the betahat vector of coefficients under the null hypothesis
resm <- resid(modelm)
resf <- resid(modelf)
B <- 10000
for (bb in 1:B) {
  newym <- cbind(1, cats$Bwt[cats$Sex=="M"])%*%betahat_null +
    sample(resm, length(resm), replace=TRUE)
  newfitm <- lm(Hwt~Bwt, data=data.frame(Hwt=newym,
                                         Bwt=cats$Bwt[cats$Sex == "M"]))
  boot_coeffs <- coef(newfitm)
  ## Similar for modelf
}
```

In particular, we will test the single null hypothesis

$$H_0: \text{ Both } \beta_{0,\text{Male}} = \beta_{0,\text{Female}} \text{ and } \beta_{1,\text{Male}} = \beta_{1,\text{Female}}.$$

The alternative hypothesis is that H_0 is false, i.e., either the slopes differ or the intercepts differ or both.

Use the following form of the test:

- Define the test statistic

$$T = (\hat{\beta}_{0,\text{Male}} - \hat{\beta}_{0,\text{Female}})^2 + (\hat{\beta}_{1,\text{Male}} - \hat{\beta}_{1,\text{Female}})^2.$$

- Reject H_0 at level α if $T > q_\alpha$, where q_α is the $1-\alpha$ quantile of the null distribution of T, that is, the distribution of T when $\beta_{0,\text{Male}} = \beta_{0,\text{Female}}$ and $\beta_{1,\text{Male}} = \beta_{1,\text{Female}}$.

To bootstrap from the null distribution, replace the estimated male and female regression parameters by a common set of parameters, e.g., the marginal regression line from part (a) or the causal regression line from part (b). It won't matter which one you use because of the form of the statistic T. Be sure to replace the ε's by sampled values from the appropriate residuals. Perform the test by bootstrapping the p-value. That is, find $\Pr(\hat{T}^* \geq T)$, where T is the test statistic computed from the observed data and $\hat{T}^*$ is drawn computed from a generic bootstrap sample from the null distribution.