BST 764: Applied Statistical Modeling
Breheny

Assignment 6: Robust regression and the bootstrap
Due: Thursday, December 15

Mathematical concepts and derivations

1. We remarked in class that robust regression estimates can be viewed as reweighted least squares estimates. Derive the weight function for the Huber and Tukey estimators as a function of the residual r_i. Plot the three weight functions (Huber, Tukey, and OLS) versus r_i and comment on how they differ and what it means in terms of how estimation is affected. Note that for linear regression, $w(r_i) = 1$ for all r_i.

2. It is sometimes claimed that unlike central limit theorem-based approaches, the bootstrap is a small-sample method that does not require large n. Comment on whether or not you think this is true.

Simulation

3. Conduct a simulation study comparing three regression methods: least squares, Huber's robust regression, and Tukey's biweight. Generate the data in the following manner:

$$X_i \overset{\text{iid}}{\sim} \text{Unif(0,1)}$$
$$Y_i|X_i = 2X_i + E_i,$$

   for $i = 1, 2, \ldots, 50$. The errors, $\{E_i\}$ have the following distribution: with probability $1 - p$, E_i is drawn from the standard normal distribution. With probability p, E_i is drawn from a normal distribution with mean 0 and standard deviation 5. This type of distribution is known as a *mixture* (in this case, a mixture of normals), and can be thought of as arising from a situation in which most of the data is coming from a reasonably reliable source, but with probability p, the data comes from a much messier and less reliable source.

   Examine three different contamination probabilities: $p = \{0, .1, .2\}$. Make a table comparing the mean squared error regarding the estimation of β for each of the three methods and each of the three probabilities (*i.e.*, a 3x3 table). Use at least 500 replications for your simulations. Comment on your results.

4. Conduct a simulation study to determine how accurately the asymptotic and bootstrap approaches estimate the true standard error of a regression coefficient. For one of the scenarios in problem 3 (*i.e.*, choose a value of p and a regression method; it doesn't matter which, but tell me which one you choose), generate 100 additional data sets. For each data set, calculate the asymptotic standard error (available from `summary(fit)$coef`) and the bootstrap standard error (use 200 bootstrap resamples). Calculate the average standard error for each approach and compare it with the actual standard error (which you already have from problem 3). Are the two approaches accurate?

Application

5. One great advantage of the bootstrap is that it can be used to calculate the standard error of arbitrarily complicated statistics for which it would be very difficult to come up with a closed form solution. The course website contains a data set (`testscores.txt`) from a study of student's test scores in various subjects. Eighty-eight students were given 5 tests, on Mechanics, Vectors, Algebra, Analysis, and Statistics. As we would expect, there are large correlations between test scores; for example, the correlation between statistics test scores and algebra test scores is 0.66.

   A natural question about this data is the extent to which these tests measure separate skills vs. general tests of quantitative ability. One way to quantify this is by looking at the eigenvalues of the correlation matrix (*i.e.*, the square of the singular values from principal components analysis); specifically, we calculate:

$$\hat{\theta} = \hat{\lambda}_1 / \sum_{i=1}^{5} \hat{\lambda}_i,$$

   where $\{\hat{\lambda}_i\}$ are the eigenvalues, sorted from largest to smallest.

   (a) What is $\hat{\theta}$ for the test score data? What does that imply about how test scores reflect latent quantitative ability?

   (b) Use the bootstrap to calculate the standard error of $\hat{\theta}$, and calculate a 95% confidence interval.

6. A study done at the University of Iowa investigated the tailgating behavior of young adults. In a driving simulator, subjects were instructed to follow a lead vehicle, which was programmed to vary its speed in an unpredictable fashion. As the lead vehicle does so, more cautious drivers respond by following at a further distance; riskier drivers respond by tailgating. The outcome of interest is the average distance between the driver's car and the lead vehicle over the course of the drive, which we will call the "following distance".

   The study's sample consisted of four groups, classified according to their use of recreational drugs: None (`NODRUG`), alcohol (`ALC`), marijuana (`THC`), and "Ecstasy" (`MDMA`). The primary research question is whether drug use is related to tailgating, with the hypothesis being that individuals with "risk-taking" personalities would be more likely to engage in both.

   (a) Fit a linear regression model (in this case, equivalent to a one-way ANOVA) to the data. Present and comment on the results of the analysis.

   (b) Present a plot (or plots) of the residuals from model (a). Do there appear to be any outliers? Do you think they are affecting the results in part (a)?

   (c) Fit a robust regression model (either the Huber or Tukey approach, or both) to the data. Present the results and comment on how they differ from those in part (a).

   (d) Choose an analysis plan (possibly one of the models from (a) or (c), possibly something else like throwing out outliers) that you think is most appropriate for this data. Briefly describe why you feel this approach is best, and state your final conclusions concerning the primary research question(s).