Advanced Statistical Computing

~~Computational Statistics~~

Exercises for week 2

Introduction. We are going to work on regression through the origin, using least squares and least (sum of) absolute values (of residuals). The statistical model is

$$Y_i = \beta x_i + e_i,$$

where the errors $e_1, \ldots, e_n$ are i.i.d. We consider two estimators for β. First the ordinary least squares estimator, which is given by the explicit formula

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

Second the minimizer $\tilde{\beta}$ of the sum of the least absolute deviations

$$\beta \mapsto \sum_{i=1}^n |Y_i - \beta x_i|.$$

This estimator can be computed in *R* with the function `rq` in the library `quantreg`. The syntax is the same as for the function `lm` for ordinary least squares.

1. Simulating outliers. The errors ("noise") are not normally distributed, but follow a mixture of two normals. Let the non-outliers come from a normal distribution with zero mean and standard deviation σ. With probability $1-p$ we draw noise from this distribution. With probability p we draw from a normal distribution with zero mean and standard deviation $c\sigma$. Reasonable values are $p = 0.1$ and $c = 2$, or $c = 5$, or $c = 10$. Do such a simulation for the case that $\sigma = 1$. Check your work with histograms and kernel density estimates.

2. Generate data sets. Use $n = 20$ and x from a uniform distribution. Set $\beta = 1$, add noise with outliers (as described above) and generate a few data sets. Make a few plots to check your results.

3. Least squares fit. Generate many data sets, estimate β and σ for them. Show the distributions of the estimates.

4. Summaries Compute mean and standard deviation for the estimates of β and σ. Do this for the three values (2, 5, 10) of c. Keep a record of the results.

5. Quantile regression. Do the same for the quantile regression estimate of β. Compare with the previous results.

6. Quantile regression and standard deviation. To estimate the standard deviation σ when using quantile regression, consider $s = \sum_i^n |y_i - \tilde{\beta} x_i|/n$. Study this estimator for the case of no outliers. Is it biased?

7. False positives. Suppose we reject the null hypothesis that $\beta = 0$ when $\tilde{\beta} > 2s\sqrt{\sum_i^n w_i^2}$, with $w_i = x_i / \sum_j^n x_j^2$. What is the empirical type I error rate?

8. Power. If we accept the above procedure for hypothesis testing, what is its power if β ranges from 0.2 to 1 (in steps of 0.2, or smaller)?