

ST810A – Assignment 2 – Simulation study

Due: 2/27/09

Objective: Our objective is to compare the performance of least squares regression and least absolute deviation regression in the presence of outliers. Assuming the usual simple linear regression setup and notation, the least squares estimate minimizes

$$\sum_{i=1}^n (y_i - \alpha - x_i\beta)^2.$$

This solution is computed in R using `lm(y ~ x)$coef`. The least absolute deviation estimate minimizes

$$\sum_{i=1}^n |y_i - \alpha - x_i\beta|.$$

This solution is computed in R by loading the “quantreg” package and using `rq(y ~ x)$coef`.

Data model: For each simulated data set generate n independent realizations from

$$y_i = \alpha + x_i\beta + [(1 - v_i)\varepsilon_i + v_i10], \quad (1)$$

where $\alpha = \beta = 1$, $x_i = (1 - u_i)z_i + u_i10$, $u_i \sim \text{Bern}(\rho_u)$, $v_i \sim \text{Bern}(\rho_v)$, $z_i \sim N(0,1)$, and $\varepsilon_i \sim N(0,1)$ (all independent). In this model, both the predictor and residuals may be contaminated with outliers; ρ_u is the contamination proportion for the predictor and ρ_v is the contamination proportion for the residuals.

Simulation design: We will compare the two models by varying the sample size and the contamination proportions. For each combination of n , ρ_u , and ρ_v specified below, generate 1,000 data sets, compute the least squares and least absolute deviation fit for each simulated data set and compute the bias and mean squared error for α and β . The designs are

1. Contaminated residuals: $n \in \{50, 500\}$, $\rho_u = 0$, and $\rho_v \in \{0.000, 0.001, \dots, 0.020\}$
2. Contaminated predictors: $n \in \{50, 500\}$, $\rho_u \in \{0.000, 0.001, \dots, 0.020\}$, and $\rho_v = 0$

Turn-in: Write a 2–4 page report in \LaTeX that includes a description of the statistical methods and simulation design, tables and/or figures displaying the results, and a discussion of the results. Also turn in your code, well-commented of course, on a separate page (doesn't have to be \LaTeX).