

Assignment 8

1. Recall the data set `boston.dat` that you used in Assignment 5. Use the SAS® script `boston2.sas` or the R script `boston2.R` (both on the course web site) to generate your personal random subset of 30 of the communities in the full data set. As before, replace “seed” with your personal number (the same one as before) listed after your name below:

BAI LEI	841	LIU MINZHAO	473
BROWN STEPHEN V	627	LU CUIE	640
CHANG CHE-SHUN	716	LUO XUAN	396
CHEN OU	189	MA LU	206
DONOHUE MICHAEL	792	MALLICK PRANJAL	994
GAO HAIBING	712	MARCUS GABRIEL	113
GARG DIVYA	221	NAMKOONG YOUNG	703
GLUCK MATHEW R	796	NEAL DANIEL W	630
GORDON ROBERT F	976	PETTERSON SONIA	453
GUCI LEDIA	207	PRANO BRIJIDA A	424
HARIHARAN POOJA	326	SHAO ANQI	369
HUANG LEI	380	SINHA AMIT	931
KIM CHANMIN	544	THAYER LAURA K	354
KIRPICH ALEX	834	YE RONGZHONG	172
LEARY EMILY V	395	ZHOU ZHUO	446
LI KE	732	ZHU XIAOYU	180
LIN TONG	100	demo	656

Add commands to the end of your script to fit the linear regression model (with intercept) for the median home value `MEDV` as a function of all other variables in the data set (with each independent variable contributing a single linear term to the regression equation). Add further commands (or options) to do the following:

- Produce a plot of *Studentized* residuals versus the corresponding predicted values. Describe any aspects of the plot that seem to deviate from what would be expected under the usual ordinary least squares assumptions.
- Produce a normal probability plot of the *standardized* residuals.
- For each observation in your data set, obtain the leverage value (v_{ii}), Cook’s D , `DFFITs`, `DFBETAS` (for the intercept and each variable), and `COVRATIO`.
- Use the usual rules of thumb to determine appropriate thresholds for each of the diagnostic statistics of part (c). For each statistic, which observations (if any) appear to be problematic?
- Obtain a partial regression leverage plot for each independent variable. In each case, are there any potentially influential data points? Is there any evidence that a transformation of the variable or addition of higher-degree polynomial terms would substantially improve the fit?

- (f) Obtain the variance inflation factor for the regression coefficient of each independent variable. Is there any evidence that collinearity (not involving the intercept) is severely affecting the precision of the corresponding parameter estimates? If so, which ones?

Attach printouts of your modified script and all relevant output.

- 2. Perform the following exercises from the textbook, Section 11.5:

- (a) Exercise 11.3 (In each case, choose exactly one plot, type of statistic, or test.)
- (b) Exercise 11.4 (Give one aspect of ordinary least squares that the diagnostic tool can be used to assess, and at least one example of a result that would indicate a problem.)