
2. Show that the test for a single outlier in a simple regression setup is not consistent.

3. The data in http://www.math.umbc.edu/~anindya/Stat601/F2004/employee.dat, contains data on 159 employees of a Seattle-based company. The variables are
   - emp: employee number
   - job: rank of employee: 2 management, 3 professional, 4 technical, 5 skilled worker, 6 clerical, 7 semi-skilled
   - sex: 1 female, 0 male
   - race1: 1 white, 0 other
   - race2: 1 african american, 0 other
   - years: years at company
   - ed: education, 0 no diploma, 1 high school, 2 some school, 3 4-yr college, 4 graduate school
   - salary: converted to full-time for part-time employee

   (a) Before starting to build a model, do exploratory data analysis (EDA) to get a feeling for the data. This includes looking at summary statistics, correlations between variables, pairwise scatterplots and individual histograms, specially for the dependent variable.

   (b) Do you think the assumption of normality holds? If not, do a Box-Cox transformation.

   (c) Repeat the EDA for the transformed model.

   (d) From the findings of the EDA, guess which variables seem to be important in explaining salary. Justify.
(e) Start building a linear model. As a practice use all model selection procedures that you have heard about, e.g. Forward, Backward, Stepwise, MAXR $C_p$ etc. Summarize the results and choose the best candidate model. You should also explore any possibility of interactions.

(f) Do model diagnostics such as the residual analysis to see if the chosen model is a satisfactory one.

(g) Do INFLUENCE diagnostics as well.

(h) Try fitting the full model with all the variables including employee number. (you should not include ‘employee number’ in the list of possible predictors because even though it may turn out significant and even though there maybe intuitive interpretations of why it is important we still do not know how they were assigned.) Are the standard errors of the common coefficient very different from those in the model you chose?

(i) One could argue that job level should really not be included as a predictor of salary, since promotion/rank structure may actually be part of possibly discriminatory practices. Consider the model that includes education, race1, race2, years and sex as predictors, but not job level.

Is race a significant predictor? Give the appropriate test.

Is sex a significant predictor? Give an appropriate test.

Consider the coefficient of ‘years’ in the above model. How should we interpret this coefficient?

(j) The company wants to hire a new employee: An asian female with 4-yr college experience. Predict the salary such a person should be offered. Provide a 95% prediction interval.