

Extra regression problems

1. Consider again the daily S&P 500 values in the file `snp500.mpj`. The discussions in the handout and CHS Casebook on estimating rates of returns discusses three different ways of estimating an effective rate of return for an investment — the arithmetic mean of the returns, the geometric mean of the returns, and using the slope in a semilog model on the accumulated value of the investment. Construct each of these estimates, treating the S&P 500 as the investment of interest. Describe the assumptions underlying the different estimates, and which one(s) you prefer.
2. Consider again the information on research universities in the file `collsamp.mpj`. Included in that data set is a variable `Academic reputation`, which is a rating of each university from university presidents and deans. Construct a model to predict academic reputation from the other variables available (these are the percent of incoming freshmen who were in the top 10% of their high school class, the acceptance rate of applicants, the expenditure per student, the freshman retention rate, the graduation rate, and whether or not *U.S. News and World Report* rated the school as a top 50 university). Be sure to check assumptions. Give an interval estimate for the academic reputation value for NYU. Its relevant values are 60% of freshmen in the top 10% of their class, 47% acceptance rate, \$24135 expenditure per student, 86% freshman retention rate, 70% graduation rate, and yes, NYU was rated a top 50 school (#35, as a matter of fact).
3. The file `coaster.mpj` are data collected from The Roller Coaster Page (<http://roller.coaster.net>), a web site devoted to North American roller coasters. The file includes information for 211 roller coasters: whether or not the tracks are made of steel (wood is the alternative), the length of the track in feet, the maximum height of the coaster in feet, the number of inversions (where the cars turn upside down), and the top speed of the coaster in miles per hour.
 - (a) Build a model relating the top speed of the coaster to the other variables. Be sure to check assumptions. Take appropriate corrective action if necessary (if we have discussed how to do so), or discuss any violations of assumptions you see (if we have not discussed how to correct them). What are the implications of your model?
 - (b) The second worksheet in the `.mpj` file gives information for 27 more roller coasters, where the top speed of the ride was missing in The Roller Coaster Page. Use your model from (a) to make predictions for the top speed of these 27 rides (including assessments of variability in those predictions). Which coaster would you suggest a daredevil take a ride on? How about people like me, who hate roller coasters? (*Note:* to apply data from the second worksheet to a regression fit using data from the first worksheet, cut and paste the data from worksheet 2 into columns in worksheet 1).

Note: If you are getting the data as `.mtp` files, the data from part (a) are in the file `coaster1.mtp`, while those from part (b) are in the file `coaster2.mtp`. I recommend that you create a project file that contains both of these worksheets in it.

4. The file `manincome.mpj` was gathered by John Iglar based on information from the 1990 decennial census, and gives demographic information for 34 ZIP code areas in Manhattan. The variables given are the percent of the people living in that ZIP code that are female, the percent that are white, the percent that are black, the median age, the percent with a high school degree but no higher, the percent with a college degree but no higher, the percent with a graduate degree, and the median household income. Build a model relating median household income to the other variables. Take appropriate corrective action if necessary (if we have discussed how to do so), or discuss any violations of assumptions you see (if we have not discussed how to correct them). What are the implications of your model?