Problem set 1

Your answers are due on September 28, 2023 at noon. You should submit them through Gradescope.

1. Math SAT scores (Y) are normally distributed with a mean of 500 and a standard deviation of 100. An evening school advertises that it can improve students' scores by roughly a third of a standard deviation, or 30 points, if they attend a course which runs over several weeks. The statistician for a consumer protection agency suspects that the courses are not effective. She views the situation as follows: $H_0 := 500$ vs. $H_1 := 530$.

   (a) Sketch the two distributions under the null hypothesis and the alternative hypothesis.

   (b) The consumer protection agency wants to evaluate this claim by sending 50 students to attend classes. One of the students becomes sick during the course and drops out. What is the distribution of the average score of the remaining 49 students under the null, and under the alternative hypothesis?

   (c) Assume that after graduating from the course, the 49 participants take the SAT test and score an average of 520. Is this convincing evidence that the school has fallen short of its claim? What is the p-value for such a score under the null hypothesis?

   (d) What would be the critical value under the null hypothesis if the size of your test were 5% ?

   (e) Given this critical value, what is the power of the test? What options does the statistician have for increasing the power in this situation?

2. In the following two problems you will investigate whether money influences elections, and learn how to work in STATA.

   The file VOTE1.dta (available on Canvas) contains data on election outcomes and campaign expenditures for 173 two-party races for the U.S. House of Representatives in 1988. There are two candidates in each race, A and B. Let *voteA* be the percentage of the votes received by Candidate A and *shareA* be the percentage of total campaign expenditures accounted for by Candidate A. Many factors other than *shareA* affect the election outcome (including the quality of the candidates and possibly the dollar amount spent by A and B). Nevertheless, we can estimate a simple regression model to find out whether spending more relative to one's challenger implies a higher percentage of the vote. Denote the variable X to be *shareA* and the variable Y to be *voteA*.

Do the following tasks using STATA:

(a) Construct variable X from the data available to you;

(b) Calculate the sample means of X and Y;

(c) Calculate the sample standard deviations of X and Y and the sample correlation coefficient between X and Y;

(d) Produce the OLS estimated regression coefficients from the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$;

(e) Create a new variable $\hat{Y}_i$ $i = 1, ..., n$, containing the predicted vote for each election based on the regression above;

(f) Create a new variable $\hat{u}_i$ containing the OLS residual for each election;

(g) Calculate the sum of $\hat{u}_i$ and explain why it should be zero;

(h) Graph the scatterplot of the data points and the regression line.

3. We continue to work with VOTE1 data set. Estimate a regression of votes on spending share, using the "`robust`" option.

   (a) What is the estimated slope? Explain in words what it means. Is the estimated effect of spending on share large or small? Explain what you mean by "large" or "small".

   (b) Report the 95% confidence interval for β_1, the slope of the population regression line.

   (c) Does spending explain a large fraction of the variance in vote? Explain.

   (d) Look at the correlation coefficient between share and vote computed in the previous problem, and compare its square to the R^2. How are they related? Provide a simple mathematical derivation of this fact.

(e) What is the root mean squared error of the regression? What does this mean?

(f) Based on your graph from 2(h), does the error term appear to be homoskedastic or heteroskedastic?

(g) Run the regression again without the “robust" option. Compare the results to what you obtained with the “robust" option. What is the same and what is different?

STATA HINTS. Note that STATA has on-line help. The following commands will be useful:

`list` lists the data

`summarize` computes sample means and standard deviations (the option “`,detail`" gives additional statistics, including the sample variance)

`correlate` produces correlation coefficients (with the option “`, covariance`" this command produces covariances)

`regress` estimates regression by OLS

`predict` computes OLS predicted values and residuals

`generate` creates a new variable

`scatter` creates a scatter-plot. There is an option to connect the dots by adding the option “`,connect()`"