Assignment 2

PSY5113, Spring 2015

This assignment is due by Feb. 17 in class. Please attach only the relevant syntax and output from using statistical packages.

1. Why are sampling distributions so important in the process of making statistical inferences? Also give a concrete example of a statistical test to demonstrate the role of sampling distribution in hypothesis tests.

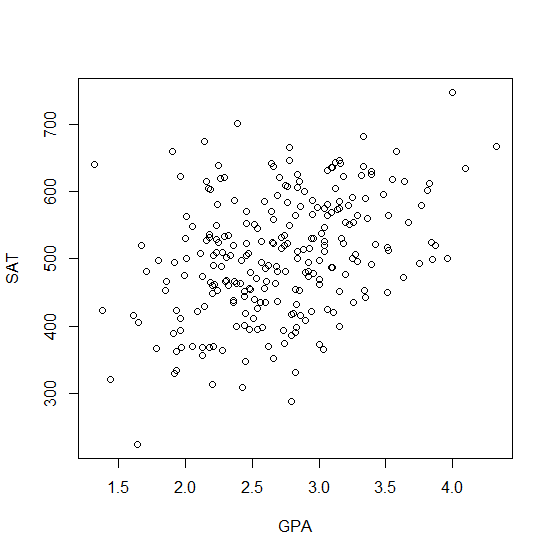
Sampling distributions are important because they inform researchers about the amount of error they can reasonably expect in their estimates of parameters. Moreover, they allow researchers to make hypothesis tests with narrow confidence intervals.

For example, we might assume that the distribution of GPA scores for students with higher than average “previous accomplishment” scores was different from the distribution of GPA scores for students as a whole. To test this, we might perform take a single sample, compute the average and sample standard deviation, and use this data to test the hypothesis that the higher-achieving students represent a different population than lower-achieving students. The average GPA for students scoring above the mean “prevach” score is 2.89 with a standard deviation of 0.53. We can compare this to the expected mean of 2.52, and as this value is easily within a standard deviation of the mean, we may be tempted to conclude that we cannot reject the null hypothesis that these two groups follow the same GPA distribution. However, we can approximate a sampling distribution by dividing the standard deviation by the square root of the sample size and conduct a two-sample t-test. This will yield an estimate of the true population average for High-Prevach students with a much smaller amount of variance. A t-test yields a t-score of -5.8, p<0.001, a highly significant result.

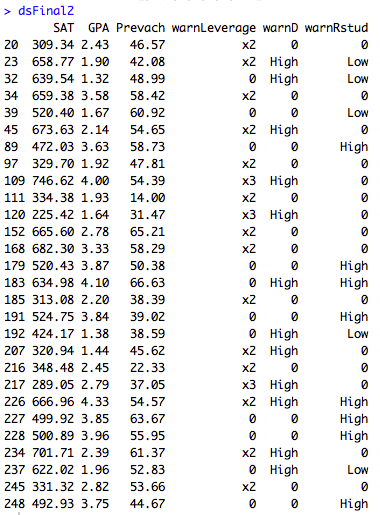
1. The dataset “cgpa.txt” includes three variables: SAT, GPA (College GPA), and PrevAch (previous achievement test score). Your goal is to study the relationship between SAT and college GPA for now. Note that this data set is the modified version of the one used in your first assignment.
2. Check possible outliers when examining the relationship between SAT and GPA. Be sure to detect outliers by (1) using a bivariate scatter plot and (2) looking at Leverage, Cook’s D, and Studentized residuals.

In your answers,

1) show the scatter plot and circle or highlight the potential outliers in the plot,



(2) report cases with above- threshold Leverage, Cook’s D, and Studentized residuals, and (3) identify outliers and list the values of SAT and GPA for the outlier cases.



The above graph represents those observations whose Lever, Cook’s D, and Studentized residuals were above a certain threshold. The threshold for the Leverage was any observation of the independent variable (GPA) that was either twice the average GPA (x2) or three times the average value (x3). The threshold for the Cook’s Distance was 4/N-k-1 or 4/248 or approximately 0.016. The threshold for the residuals was a z-score greater than 1.96 (“High”) or less than -1.96 (“Low”).

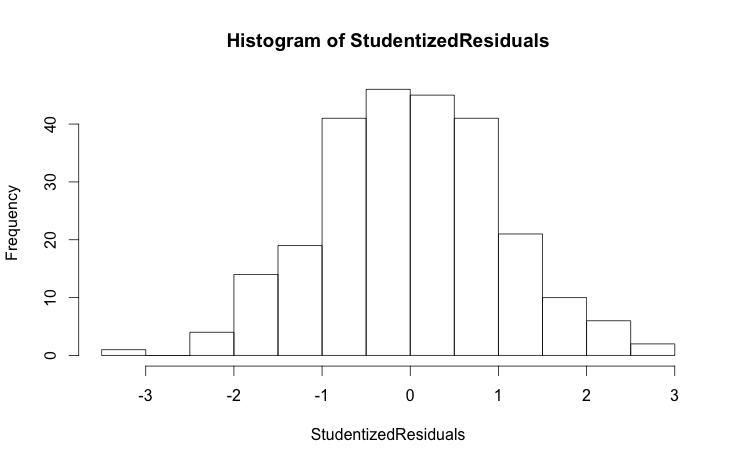
1. Remove the cases that were identified as outliers from the data set. Then conduct a regression analysis predicting GPA from SAT: Write up the results as if you are working on the Results section in a manuscript. Be sure to include the results on overall significant test of the model, the significance tests of individual model parameters, and interpret all these results (APA format is required).

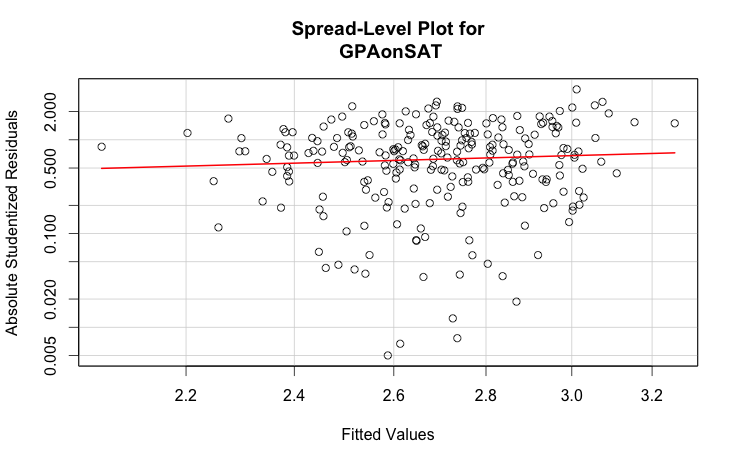
Results suggest that SAT is a significantly predictor of GPA,  = 0.002308, t(248) = 8.302, p<.001. SAT scores also explained a significant, if small, proportion of the variance,  = 0.1415, F(1, 248)=40.89, p<0.001. Several observations show signs of being outliers. Outliers were selected on the basis of having both a high Cook’s Distance and Studentized residuals greater than 1.96 or less than -1.96. This is an intentionally stringent method, selecting only those observations most likely to represent extreme outliers, and excluding as few observations from the analysis as possible. Using these criteria, the new dataset excludes only sets of observations, those from individuals 23, 32, 183, 192, 226, 237, and 248. Running the model with this slightly reduced model yields that explains a significant, and slightly larger percentage of the variance,  = 0.1604, F(1, 242)=46.23, p<0.001. This somewhat increased F-score and correlation coefficient suggests the model (unsurprisingly) fits the reduced set of data slightly better than the full data set. However, the estimate of SAT remains a largely unchanged, if still significant, predictor of GPA,  = 0.002359, t(242) = 6.8, p<0.001. Since removing the most extreme outliers only improved the overall fitness of the model, without affecting the estimates of the predictive ability of SAT on GPA, we conclude that the outliers do not represent extraordinary circumstances, and should remain a part of the analysis. Moreover, the predictive ability of both SAT and the model as a whole remains high, and can be considered robust, even in the presence of outliers.

1. Conduct diagnostics on model assumptions, including normality of errors, homogeneity of residuals, and independence of errors. Be sure to report both graphs and statistical tests for your diagnostics on each assumption.

An analysis of the independence of the residuals yields an autocorrelation coefficient of r=0.03 and a Durbin-Watson statistic of DW(1, 248) = 1.92, p=0.56. This statistic is far from significant, giving no reason to reject the null hypothesis of no autocorrelation among residuals.

Likewise, the assumption of normality seems safe. A graph of the residuals shows no significant deviation from a normal distribution. A Shapiro-Wilk Normality tests test statistic of W = 0.9967, p=0.8827, which gives us no reason to reject the assumption of normality.



However, the residuals do NOT appear to be identically distributed. A White test yields a White’s Chi-Squared value of  (2) = 6.661243, p<0.01. This gives is reason to believe the error variance is not constant across time, and to reject the assumption of homoskedasticity. 

1. Now you are predicting college GPA from SAT and PrevAch using the dataset of “gpa.txt” with outliers removed.
2. Descriptive statistics: report (1) the mean and standard deviation for each of the

three variables, and (2) the correlation matrix for the three variables.

GPA has a mean of 2.7 and standard deviation of 0.543

SAT has a mean of 504.6 and standard deviation of 88.45

Prevach has a mean of 50.529 and standard deviation of 10.572

The correlation matrix of the three variables is:

GPA 1

SAT .376 1

Prevach 0.328 .503 1

1. Compute the estimates for the intercept, slope of SAT, and slope of PrevAch using the formulas provided in the lecture. (show your steps)

 = 0.001734

 = 0.00955



= 1.342

1. Compute the standardized coefficients using the formulas provided in the lecture. (show your steps) State which variable is more important in predicting GPA and why?

 = 0.283

 = 0.186



A one-standard deviation increase in SAT scores has a greater impact than a one-standard deviation increase in Prevach. This suggests a student’s SAT score is more predictive of their college GPA than is their Previous Accomplishment score.

1. Use a computer program to run the regression of GPA from SAT and PrevAch. Report both raw and standardized coefficients and their significance. Interpret the respective effect of SAT and PrevAch on GPA.

SAT has a raw regression coefficient of =0.0017, t(247)=4.206, p<0.001 and weighted regression coefficient of =0.283, t(247)=4.206, p<0.001.

Prevach has a raw regression coefficient of =0.00955, t(247)=2.77, p<0.01 and weighted regression coefficient of =0.186, t(247)=2.77, p<0.01.

Both predictor variables are highly significant predictors of GPA. Of the two, SAT seems to be the better predictor, as a one-standard deviation increase in SAT has a greater impact on predicted GPA than a similar increase in Prevach.

Of the two models, the weighted allows for clearer interpretation. Without standardizing the variables to both have a variance equal to one, the regression coefficient might erroneously lead someone to believe that Prevach was the stronger predictor variable, as it has a higher raw regression coefficient. Standardizing the variables allows a more accurate interpretation.