Diana Liang

10/9/2019

Stats 500: HW #3

**PART A: SAT Data**

1. summary(sat\_model)  
   ## Call:  
   ## lm(formula = total ~ takers + ratio + salary, data = sat)  
   ##   
   ## Residuals:  
   ## Min 1Q Median 3Q Max   
   ## -89.244 -21.485 -0.798 17.685 68.262   
   ##   
   ## Coefficients:  
   ## Estimate Std. Error t value Pr(>|t|)   
   ## (Intercept) 1057.8982 44.3287 23.865 <2e-16 \*\*\*  
   ## takers -2.9134 0.2282 -12.764 <2e-16 \*\*\*  
   ## ratio -4.6394 2.1215 -2.187 0.0339 \*   
   ## salary 2.5525 1.0045 2.541 0.0145 \*   
   ## ---  
   ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   ##   
   ## Residual standard error: 32.41 on 46 degrees of freedom  
   ## Multiple R-squared: 0.8239, Adjusted R-squared: 0.8124   
   ## F-statistic: 71.72 on 3 and 46 DF, p-value: < 2.2e-16

The coefficient of determination is 0.8239, meaning that the linear model explains around 82.39% of the variation in the data, which is a good percentage of the data. So this linear model would seem to be a representation of the SAT data.

1. Ho = teacher’s salary has no effect or a negative effect on SAT score

Ha = teacher’s salary has a positive effect on SAT score

## Salary T-Test Statistic: 2.541   
## Salary P-Value: 0.007245436

Since the t-test provided a p-value that is lower than the significance level of 0.01, there is enough evidence to reject the null hypothesis and suggest that salary does have a positive effect on SAT score.

1. Ho = pupil/student ratio has no effect on SAT score

Ha = pupil/student ratio has an effect on SAT score

## Ratio T-Test Statistic: -2.187   
## Ratio P-Value: 0.03386468

Since the t-test provided a p-value that is greater than the significance level of 0.01, there is not enough evidence to suggest that pupil/student ratio has an effect on SAT score.

1. Ho = none of the predictors (takers, ratio, salary) have an effect on SAT score

Ha = at least one of the predictors has an effect on the SAT score

## SAT data F-Test Statistic: 71.72   
## SAT data P-Value: 0

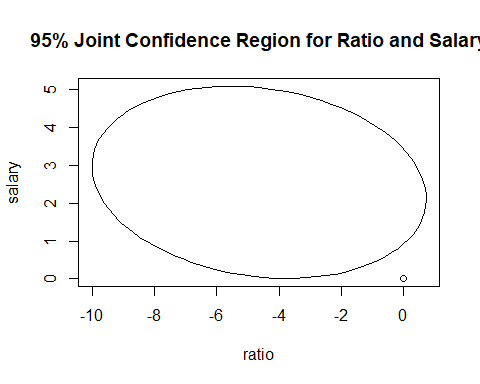
Since the F-test provided a p-value that is lower than the significance level of 0.01, there is enough evidence to reject the null hypothesis and suggest that at least one of the predictors is useful in predicting SAT score.

1. ## 2.5 % 97.5 %  
   ## salary 0.5304797 4.574461 | 95% Confidence Interval

## 0.5 % 99.5 %  
## salary -0.146684 5.251624 | 99% Confidence Interval

0 is not included in the 95% CI for the predictor salary but is included in the 99% CI. This confirms the p-value of 0.0145 for salary, since it is significant at a significance level of 0.05, a confidence interval of 95%, but not at the significance level of 0.01, a confidence interval of 99%.

1. The joint 95% CI for ratio and salary does not include the point (0,0), so at least one of the predictors is useful to modelling SAT scores. This region is mirrored by the F-test on a pair of predictors. The null hypothesis (Ho) would be that predictors ratio and salary have no effect on the SAT score while the alternative hypothesis (Ha) would be that at least one of these predictors has an effect on the SAT score. Since the p-value of the F-test is less than the significance level of 0.05, the equivalent of a 95% CI, there is enough evidence to reject the null hypothesis.



## Analysis of Variance Table  
##   
## Model 1: total ~ takers  
## Model 2: total ~ takers + ratio + salary  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 48 58433   
## 2 46 48315 2 10118 4.8165 0.01261 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1. summary(sat\_model7)

##   
## Call:  
## lm(formula = total ~ takers + ratio + salary + expend, data = sat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -90.531 -20.855 -1.746 15.979 66.571   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1045.9715 52.8698 19.784 < 2e-16 \*\*\*  
## takers -2.9045 0.2313 -12.559 2.61e-16 \*\*\*  
## ratio -3.6242 3.2154 -1.127 0.266   
## salary 1.6379 2.3872 0.686 0.496   
## expend 4.4626 10.5465 0.423 0.674   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.7 on 45 degrees of freedom  
## Multiple R-squared: 0.8246, Adjusted R-squared: 0.809   
## F-statistic: 52.88 on 4 and 45 DF, p-value: < 2.2e-16

The new model including the predictor ‘expend’ still has a similar coefficient of determination to the original model, meaning that the model explains a similar degree of the data variation. The beta estimate for takers is about the same as is the p-value, so this predictor was not greatly affected by the additional predictor. The other two original predictors, ratio and salary, have smaller beta estimates and greater standard errors, making their p-values much greater, to the point where there is a good possibility that these predictors have no effect on the SAT score. The predictor expend, meanwhile, has a large standard error and p-value, meaning that this predictor most likely has no effect on the SAT score. Overall, while the coefficient of determination suggests that the new model is just as good of a representation of the data as the original model, the addition of the predictor expend at the very least did not better represent the data and most likely skewed the predictive power of the other input variables.

1. Ho = the predictors salary, expend, and ratio have no effect on SAT score

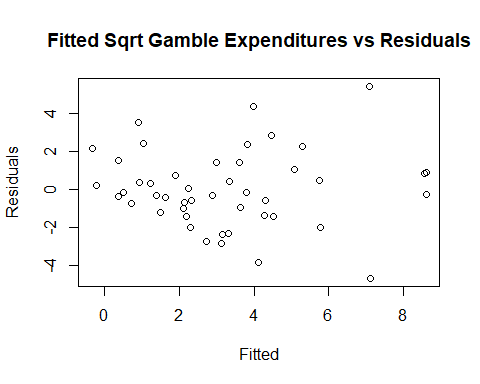
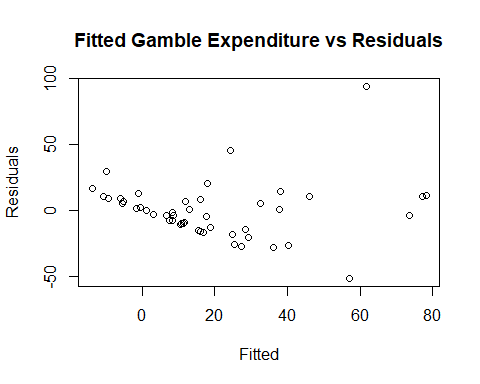
Ha = at least one of these predictors has an effect on the SAT score

## Analysis of Variance Table  
##   
## Model 1: total ~ takers  
## Model 2: total ~ takers + ratio + salary + expend  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 48 58433   
## 2 45 48124 3 10309 3.2133 0.03165 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Since the p-value of the F-test is less than the significance level of 0.05, there is enough evidence to suggest to reject the null hypothesis and suggest that at least one of these predictors has an effect on the SAT score.

**PART B: Teengamb data**

1. Constant variance assumption

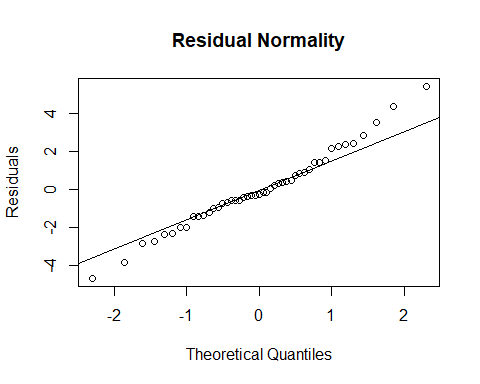


## Original Model Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.3303 2.8789 3.241 0.00224 \*\*  
## gamb\_model$fitted 0.2645 0.0968 2.732 0.00895 \*\*

## Sqrt Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.01136 0.32365 3.125 0.00311 \*\*  
## gamb\_model\_s$fitted 0.14957 0.08242 1.815 0.07623 .

Plotting the fitted values against the residuals of the original model gave a pattern that was confirmed by a beta estimate for fitted values with a p-value less than 0.01. The original model with gamble expenditure as the response variable showed heteroscedasticity, so a new linear model was fit with the square root of gamble expenditure as the response variable. The beta estimate for fitted values in the new model has a p-value greater than 0.01, suggesting that there is a possibility that the two variables do not affect each other. The new model will be used for the rest of this analysis.

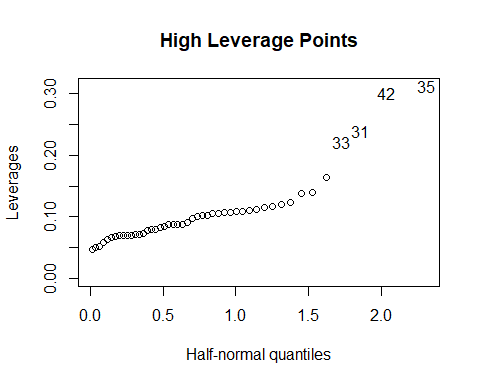
1. Normality assumption



## Shapiro-Wilk normality test  
##   
## data: gamb\_model\_s$residuals  
## W = 0.98321, p-value = 0.7272

The QQ-plot of the residuals against theoretical normal quantiles shows that the residuals generally follow a normal distribution, since no points dramatically deviate from the line representing a perfectly normal distribution. Still the left and right tails diverge the most, meaning that the residual distribution is slightly more long-tailed than normal. While this might slightly invalidate the calculated p-values, there isn’t a dramatic enough difference to discount the current p-values. The Shapiro-Wilks test verifies that the residuals follow a fairly normal distribution, so the p-values from t-tests and F-tests will still have meaning.

1. Large leverage points



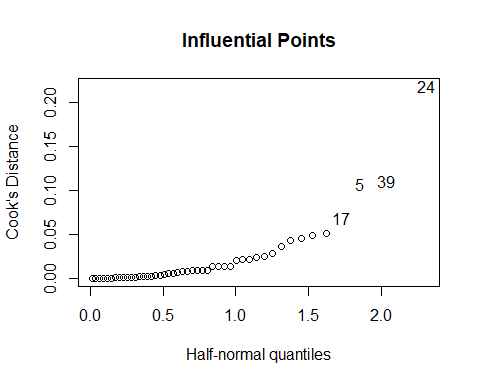
The points 35, 42, 31, and 33 were found with high leverage, which means these points were the most removed from that of the other data points in terms of predictor variables.

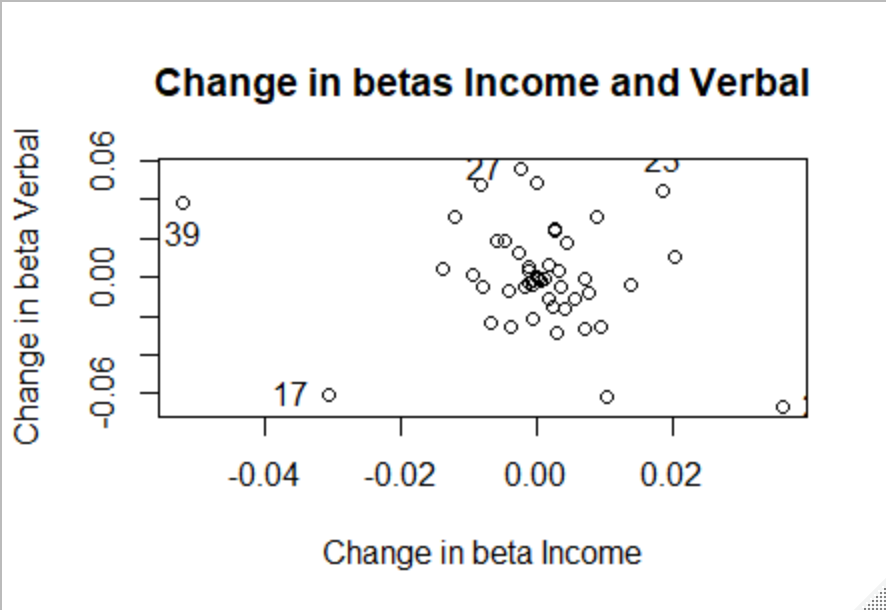
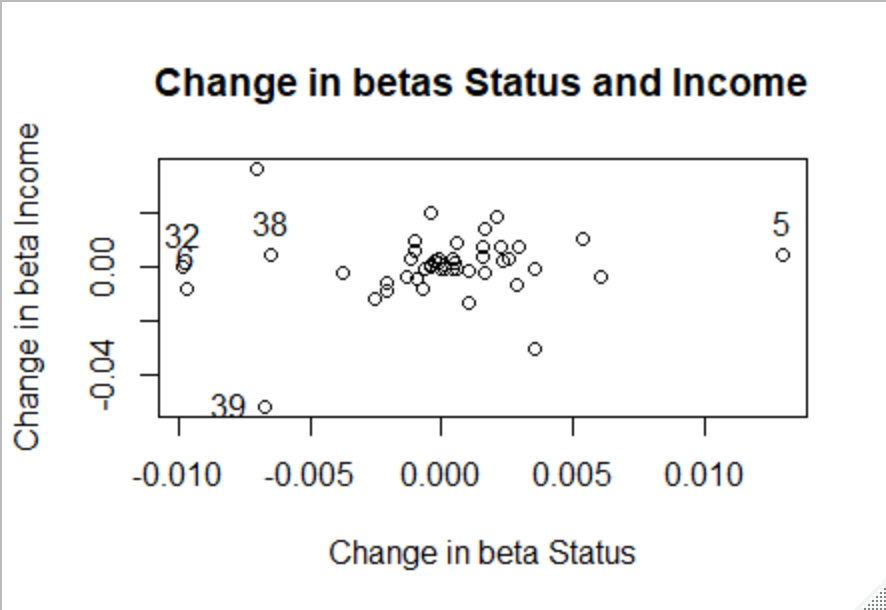
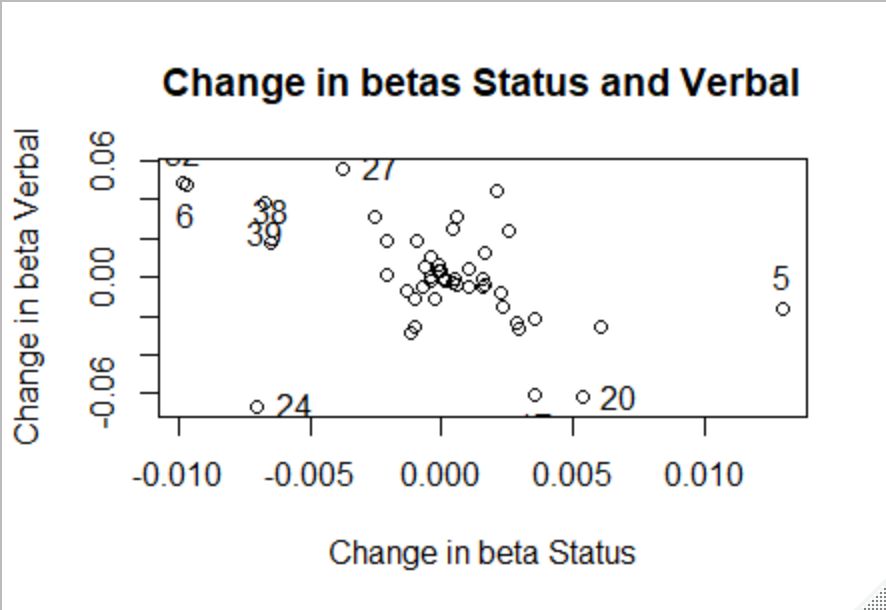
1. Outliers

## Point with Max Studentized RS: 24   
## P-value for Possible Outlier: 0.00414277   
## Corrected Significance Level: 0.00106383

While point 24 had the greatest studentized residual, it did not exceed the Bonferroni corrected significance level, meaning that this point is not an outlier. All other points will have a smaller residual compared to that of point 24, so they cannot be outliers if point 24 is not even an outlier. By these methods, no outliers exist in this data set.

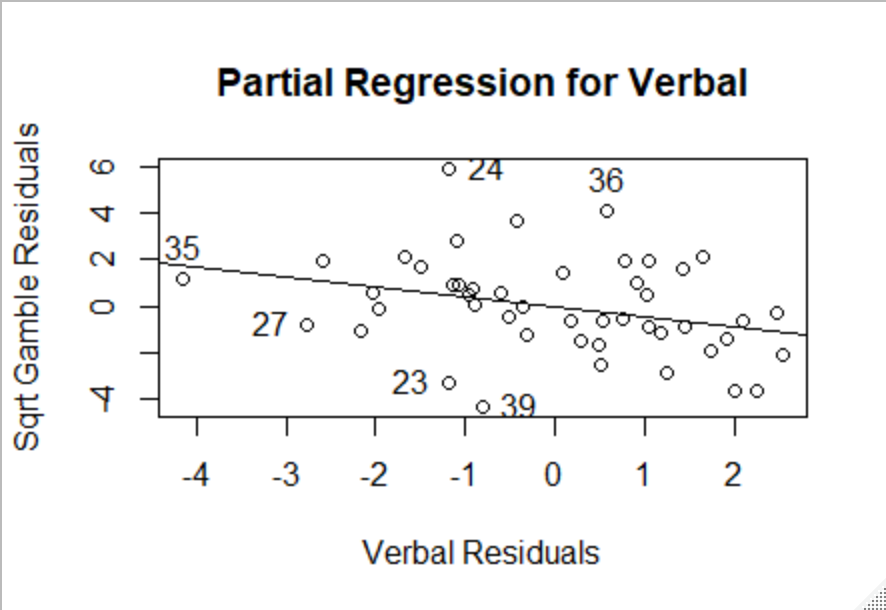
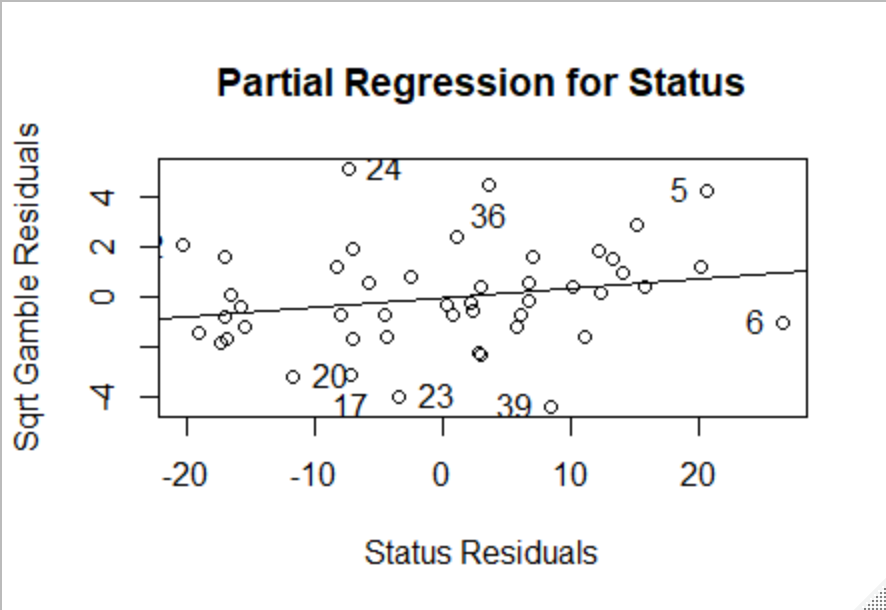
1. Influential points



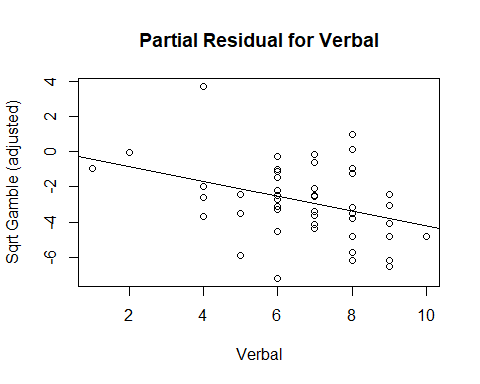
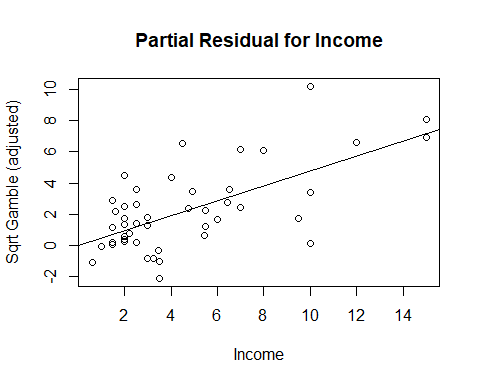
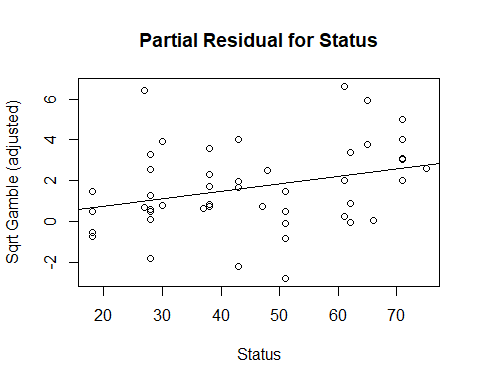


The most influential points are 24, 39, 5, and 17. Point 24, while it is not an outlier by the t-test done above, still has a greater influence on all the beta estimates shown.

1. Structure of relationship between predictors and response



The partial regression plots show points of influence and leverage that we have already identified above, validifying the previous results, and do not have an apparent nonlinear pattern.



The partial residual plots do not have notable patterns or clusters that would hint at heteroscedasticity, meaning that a different non-linear model between the response variable and the predictor variables most likely would not be a proper fit for the data.

**Part C: Confidence Regions**

The confidence region has the estimated beta values in its center. The difference between the coefficients will be less than or equal to the estimated values, so the difference will be “left” of the center. Meanwhile, the sum of the coefficients will be greater than or equal to the estimated values, so the sum will be “right” of the center. If A and B are positively correlated, the difference between the coefficients will have a greater variance than the sum of the two coefficients. This leads to a greater possible area to the left of center, making the confidence region “lean to the left”. On the other hand, if A and B are anti-correlated, the sum of the coefficients will have a greater variance than the difference between them, leading to a greater possible area to the right of center and making the confidence region “lean to the right”.

Graphically, greater variance in the difference of the coefficients would lean the confidence region to a more anti-correlated representation between the two coefficients, making the region “lean to the left”. And greater variance in the sum of the coefficients would lean the confidence region to a more positively correlated representation between the two coefficients, making the region “lean to the right”.