Lab 1: Simple Linear Regression and Correlation

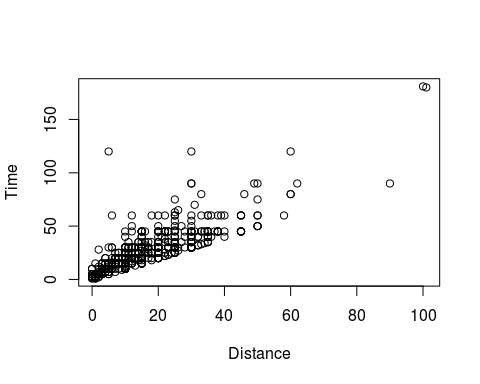
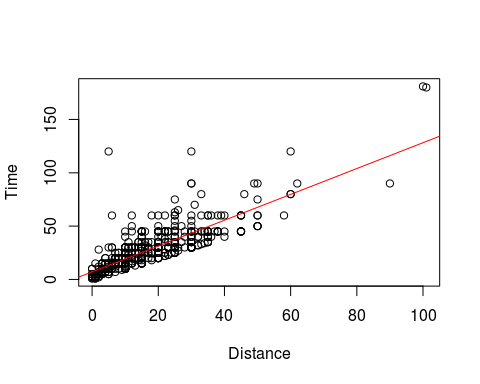
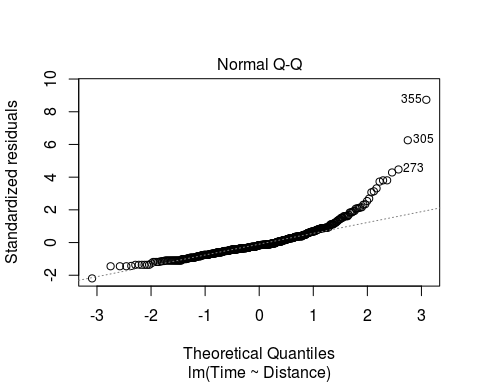
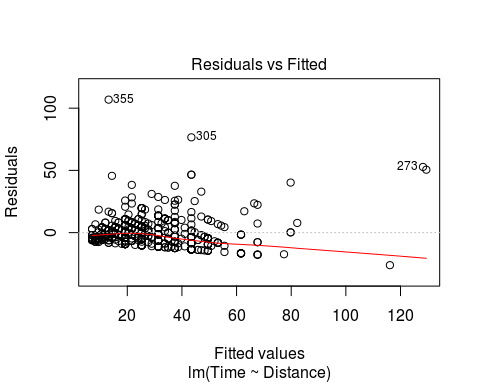
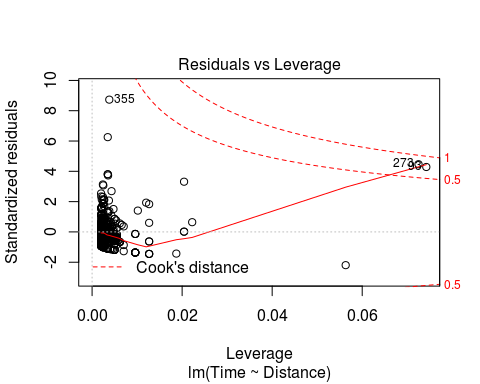
September 15, 2017

### Names:Adolfo Huerta

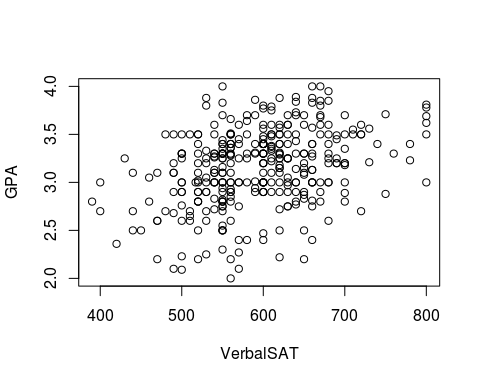
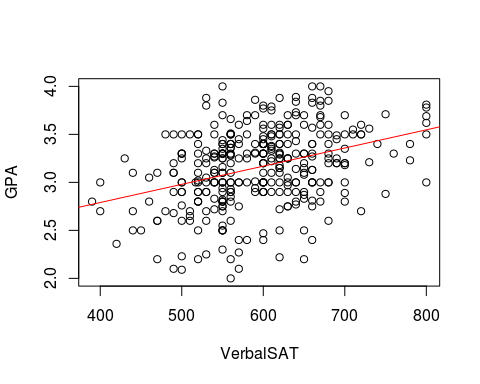
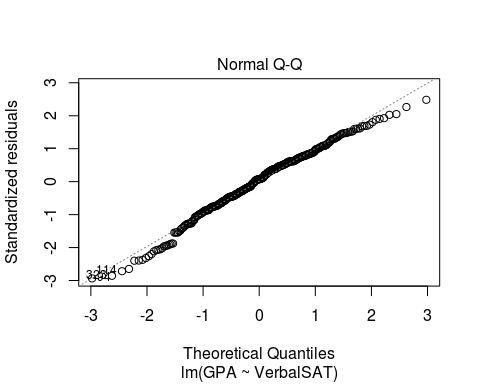
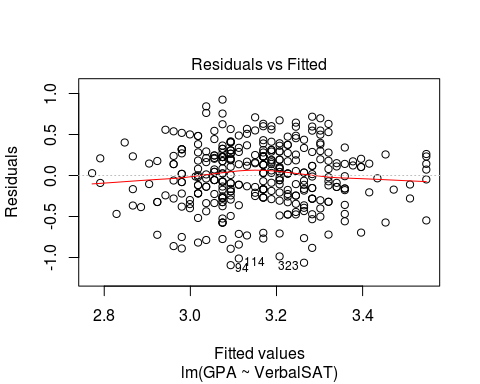
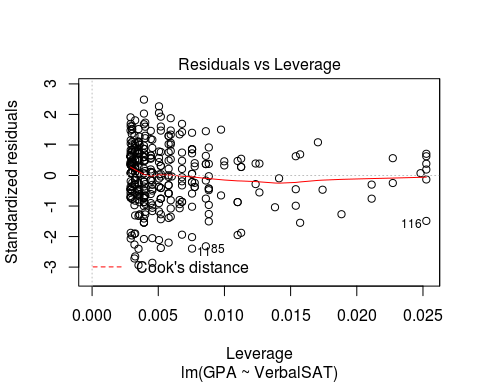
##### What each person contributed to the lab: Worked Independently

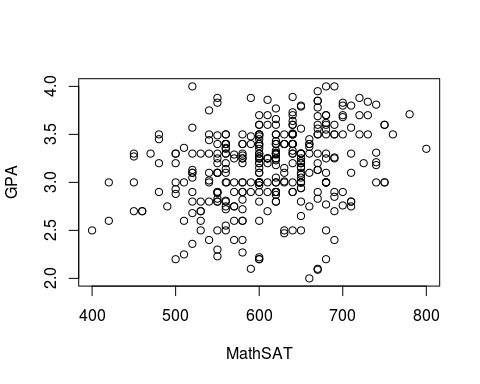
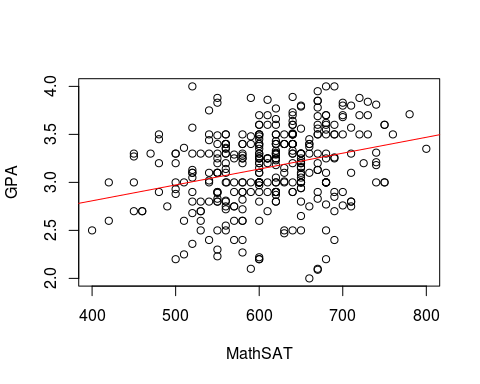
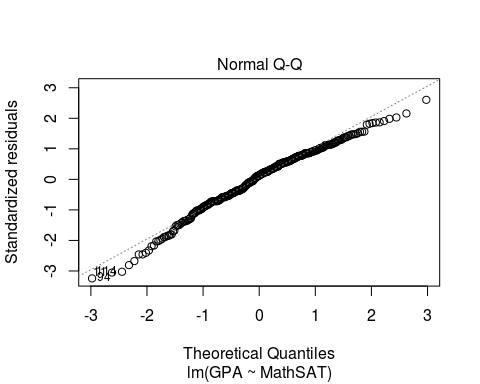
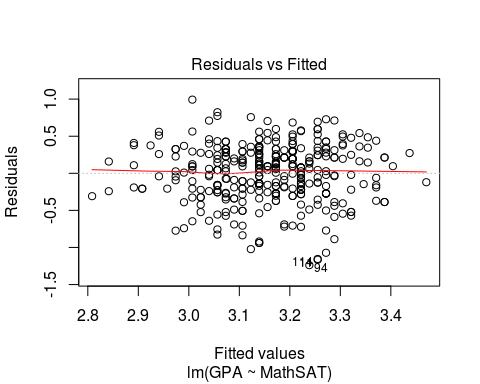
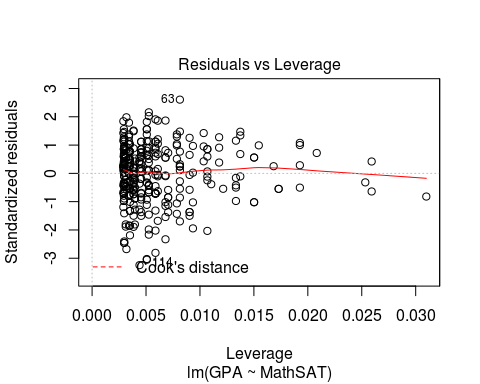
**Please make sure to show all R code and output after each question so that I can see your work.** Write a sentence for each numerical value produced describing its meaning in context with the proper units.

1. We wish to build a model to predict the the time (in minutes) to commute to work in Atlanta, GA based on the distance (in miles). The data for commute time and distance for a random sample of 500 commuters from Atlanta is found in CommuteAtlanta in the Lock5Data library.

* #Note- you might need to install these packages first if you don't already have them.  
  library(Lock5Data)   
  library(mosaic)  
  #what does the attach function do? It lets us easily reference all the variables in the CommuteAtlanta dataset without having to use the notation CommuteAtlanta$variable\_name. Instead, we can just type variable\_name. NOTE: if you are working in multiple R file as once, this can be dangerous because you can overwrite variables in other data sets. Use caution!  
  attach(CommuteAtlanta)
  1. Provide the relevant summary statistics for the study (mean and sd of variables, scatterplot, correlation). Provide and interpretation of the summary statistics (specifically the scatterplot and correlation). Based on summary statistics and plots, is there evidence of a linear relationship?
* #for a code chunk, you can set how big you want the images to be. default is inches  
    
  #you might want to start by investigating what variables are in the CommuteAtlanta dataset  
  View(CommuteAtlanta) #uncomment this and run just this line to print the data table  
  names(CommuteAtlanta)
* ## [1] "City" "Age" "Distance" "Time" "Sex"
* summary(CommuteAtlanta)
* ## City Age Distance Time Sex   
  ## Atlanta:500 Min. :16.00 Min. : 0.00 Min. : 1.00 F:246   
  ## 1st Qu.:31.00 1st Qu.: 8.00 1st Qu.: 15.00 M:254   
  ## Median :40.00 Median : 15.00 Median : 25.00   
  ## Mean :40.24 Mean : 18.16 Mean : 29.11   
  ## 3rd Qu.:49.00 3rd Qu.: 25.00 3rd Qu.: 40.00   
  ## Max. :85.00 Max. :101.00 Max. :181.00
* #here is some more code to get you started  
  plot(Distance,Time)
* 
* mean(Distance)
* ## [1] 18.156
* #Provide more R code and output here  
  sd(Distance)
* ## [1] 13.79828
* sd(Time)
* ## [1] 20.71831
* mean(Time)
* ## [1] 29.11
* cor.test(Distance,Time)
* ##   
  ## Pearson's product-moment correlation  
  ##   
  ## data: x and y  
  ## t = 30.454, df = 498, p-value < 2.2e-16  
  ## alternative hypothesis: true correlation is not equal to 0  
  ## 95 percent confidence interval:  
  ## 0.7736519 0.8352319  
  ## sample estimates:  
  ## cor   
  ## 0.8066198
* Here is an example of how you can use LaTeX commands to write equations. . Here is an example of inline code. The mean of Distance is 18.156. Based on the summary statistics, we can see that there are five variables in this dataset, namely "City", "Age", "Distance", "Time", "Sex". We have a sample size of 500 and they all live in the city of Atlanta. Our age variable ranges from 16 to 85 years old, with an average of 40.24 years old. Our distance variable has a range from 0 to 101 minutes, with an average of 18.16 minutes and a standard deviation of 13.79. Our time variable has a range from 1 to 181 minutes, has an average of 29.11 and a standard deviation of 20.72. We can see from our scatterplot that there appears to be a positive linear relationship between distance and time and we can say that a linear model is suitable for our data. There appears to be two outliers in our data with over 100 miles in distance and over 150 minutes in time, but these would not be too far off if we drew a best fit linear model and we must investigate further before decidiing to keep or remove. Finally, our correlation test is further evidence of a strong linear relationship since our correlation is 0.8066198, which is closer to one.
  1. Fit a linear regression model to predict Commute Time based on Distance in Atlanta.
     1. Find the regression model using R. Write out the model in format, plugging in the estimates. Provide a scatterplot with the regression line added to the plot.
     + #Provide R code and output/graphs   
       model <- lm(Time~Distance)  
       summary(model)
     + ##   
       ## Call:  
       ## lm(formula = Time ~ Distance)  
       ##   
       ## Residuals:  
       ## Min 1Q Median 3Q Max   
       ## -26.124 -6.693 -2.027 4.246 106.824   
       ##   
       ## Coefficients:  
       ## Estimate Std. Error t value Pr(>|t|)   
       ## (Intercept) 7.12034 0.90659 7.854 2.5e-14 \*\*\*  
       ## Distance 1.21115 0.03977 30.454 < 2e-16 \*\*\*  
       ## ---  
       ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
       ##   
       ## Residual standard error: 12.26 on 498 degrees of freedom  
       ## Multiple R-squared: 0.6506, Adjusted R-squared: 0.6499   
       ## F-statistic: 927.4 on 1 and 498 DF, p-value: < 2.2e-16
     + plot(Distance, Time)  
       abline(model, col = "red")
     + 
     + The regression model is
     1. Interpret the slope in the context of the situation.
     + For every additional mile of distance you have between Atlanta and your destination, your time will increase by 1.21 minutes.
     1. Interpret the y-intercept in the context of the situation. Is the interpretation practical in context?
     + The time it takes to arrive from Atlanta to your destiantion is 7.12 minutes, when your distance is 0 miles. This interpretation is not practical since traveling zero miles is not quantifiable and we can say that it takes zero time as well since the person is not moving any distance.
  2. Conduct the test to determine if the linear relationship is statistically significant.
     1. Set up the appropriate null and alternative hypothesis both verbally and symbolically.
     + symbolically: verbally: The slope of the regression line between time and distance is zero ( there is no linear relationship).
     + symbolically: verbally: The slope of the regression line between time and distance is not zero ( there is a linear relationship).
     1. Use R to find the ANOVA table and/or t-test statistics to test the hypothesis test.
     + #Put Code Here and insert output or graphic into word document  
       anova(model)
     + ## Analysis of Variance Table  
       ##   
       ## Response: Time  
       ## Df Sum Sq Mean Sq F value Pr(>F)   
       ## Distance 1 139363 139363 927.45 < 2.2e-16 \*\*\*  
       ## Residuals 498 74832 150   
       ## ---  
       ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
     + summary(model)
     + ##   
       ## Call:  
       ## lm(formula = Time ~ Distance)  
       ##   
       ## Residuals:  
       ## Min 1Q Median 3Q Max   
       ## -26.124 -6.693 -2.027 4.246 106.824   
       ##   
       ## Coefficients:  
       ## Estimate Std. Error t value Pr(>|t|)   
       ## (Intercept) 7.12034 0.90659 7.854 2.5e-14 \*\*\*  
       ## Distance 1.21115 0.03977 30.454 < 2e-16 \*\*\*  
       ## ---  
       ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
       ##   
       ## Residual standard error: 12.26 on 498 degrees of freedom  
       ## Multiple R-squared: 0.6506, Adjusted R-squared: 0.6499   
       ## F-statistic: 927.4 on 1 and 498 DF, p-value: < 2.2e-16
     1. Interpret the meaning of the appropriate p-value in the context of the situation.
     + The appropriate p-value in the context of the situation is 2.2e-16 and tells us there is evidence that the slope of the regression line between distance and time is not zero.
     1. State the conclusion of the hypothesis test in the context of the problem.
     + We reject the null hypothesis, in favor of the alternative hypothesis that the slope of the regression line between distance and time is not zero, since our p-value of 2.2e-16 is less than our significance level of 0.05.
     1. Interpret the (coefficient of determination) value in the context of the model variables.
     + Our coefficient of determination is 0.6506, which tells us the proportion of variability in the time variable that is explained by the regression with the distance variable. Therefore, 65.06% of the variablity in time it takes to commute from Atlanta to a specific destiantion, is explained by the linear regression with the distance in miles to the destination.
  3. We must determine whether the results of the analysis are trustworthy, i.e., the assumptions to perform the Regression Analysis are met by the data.
     1. Provide and interpret the results of the QQPlot and Shapiro-Wilk Test to determine if the normality of residuals is a satisfied assumption.
     + #Put code and output here  
       shapiro.test(residuals(model))
     + ##   
       ## Shapiro-Wilk normality test  
       ##   
       ## data: residuals(model)  
       ## W = 0.80163, p-value < 2.2e-16
     + plot(model, 2)
     + 
     + The Shapiro-Wilk Test determines if the data satisfies the normality assumpiton. In this case the p-value < 2.2e-16 so we reject the null that the residuals are normally distributed, in favor of the alternative hypothesis that the residuals are not normally distributed. We can also see this from the QQ plot, since there is not an approximately straight-line relationship between quantiles of this data and quantiles from the normal distribution.
     1. Provide and interpret the Residual Plot to test Equal Variance of the residuals.
     + #Put code and output here  
       plot(model, 1)
     +  >To test equal variance of the residuals we observe how the fitted values vary with the residuals. We want the variablity in time variable to be the same regardless of the values of the distance variable. In this case, we can see from the residual plot that there is no pattern or fanning/trumpeting to the points about zero, so the constant variance assumption is satisfied.
     1. Interpret the results of the Residual Plot to test the Independence of Residuals assumption.
     + To test the independence of the residuals assumption we want the value of the residuals to be independent of the value of x, the distance variable. Since we do not observe any pattern, such as curvature, in the residual plot and the points appear to be randomly distributed about zero then the assumption of independece is satisfied.
     1. Provide and interpret the Cook's Distance plot to test if there are outliers in the data.
     + #Put code and output here  
       plot(model, 5)
     + 
     + The cooks distance is a measure of each point's importance in determining the regression result. According to the Cook's Distance plot for our data, we can see that there are a few values of concern, since they are between the distance of 0.5 and 1. However, there are no values with distances larger than one which would indicate that they have a high influence on the estimate of the slope of the analysis.
     1. After assessing all assumptions, do you have any concerns about the analysis?
     + After assesing all assumptions, I do have concerns about the analysis because according to our Shapiro-Wilk Test, our data does not appear to be normally distributed and fails that assumption. Another concern is that we have values in our data that could be considered outliers according to our Cook's Distance plot, we would have to investigate these points further to see how much they affect our regression result.

1. We are going to try and predict a student's GPA in College based on their SAT score. The data is found in StudentSurvey in the Lock5Data library. First we are going to look at a model that examines GPA and VerbalSAT

* data(StudentSurvey)  
  attach(StudentSurvey)  
    
  names(StudentSurvey)
* ## [1] "Year" "Gender" "Smoke" "Award" "HigherSAT"   
  ## [6] "Exercise" "TV" "Height" "Weight" "Siblings"   
  ## [11] "BirthOrder" "VerbalSAT" "MathSAT" "SAT" "GPA"   
  ## [16] "Pulse" "Piercings"
  1. Provide the relevant summary statistics for the study (mean and sd of variables, scatterplot, correlation). Provide and interpretation of the summary statistics (specifically the scatterplot and correlation). Is there evidence of a linear relationship? (due to missing data, you may need to include na.rm=T for mean() and sd() and input "pairwise.complete.obs" as an argument in cor()).
  + #Provide R code and output/graphs   
    summary(StudentSurvey, na.rm = T)
  + ## Year Gender Smoke Award HigherSAT   
    ## : 2 F:169 No :319 Academy: 31 : 7   
    ## FirstYear: 94 M:193 Yes: 43 Nobel :149 Math :205   
    ## Junior : 35 Olympic:182 Verbal:150   
    ## Senior : 36   
    ## Sophomore:195   
    ##   
    ##   
    ## Exercise TV Height Weight   
    ## Min. : 0.000 Min. : 0.000 Min. :59.00 Min. : 95.0   
    ## 1st Qu.: 5.000 1st Qu.: 3.000 1st Qu.:65.00 1st Qu.:138.0   
    ## Median : 8.000 Median : 5.000 Median :68.00 Median :155.0   
    ## Mean : 9.054 Mean : 6.504 Mean :68.42 Mean :159.8   
    ## 3rd Qu.:12.000 3rd Qu.: 9.000 3rd Qu.:71.00 3rd Qu.:180.0   
    ## Max. :40.000 Max. :40.000 Max. :83.00 Max. :275.0   
    ## NA's :1 NA's :1 NA's :7 NA's :5   
    ## Siblings BirthOrder VerbalSAT MathSAT   
    ## Min. :0.000 Min. :1.00 Min. :390.0 Min. :400.0   
    ## 1st Qu.:1.000 1st Qu.:1.00 1st Qu.:550.0 1st Qu.:560.0   
    ## Median :1.000 Median :2.00 Median :600.0 Median :610.0   
    ## Mean :1.727 Mean :1.83 Mean :594.2 Mean :609.4   
    ## 3rd Qu.:2.000 3rd Qu.:2.00 3rd Qu.:640.0 3rd Qu.:650.0   
    ## Max. :8.000 Max. :8.00 Max. :800.0 Max. :800.0   
    ## NA's :3   
    ## SAT GPA Pulse Piercings   
    ## Min. : 800 Min. :2.000 Min. : 35.00 Min. : 0.000   
    ## 1st Qu.:1130 1st Qu.:2.900 1st Qu.: 62.00 1st Qu.: 0.000   
    ## Median :1200 Median :3.200 Median : 70.00 Median : 0.000   
    ## Mean :1204 Mean :3.158 Mean : 69.57 Mean : 1.673   
    ## 3rd Qu.:1270 3rd Qu.:3.400 3rd Qu.: 77.75 3rd Qu.: 3.000   
    ## Max. :1550 Max. :4.000 Max. :130.00 Max. :10.000   
    ## NA's :17 NA's :1
  + sd(VerbalSAT, na.rm = T)
  + ## [1] 74.1764
  + sd(MathSAT, na.rm = T)
  + ## [1] 68.49007
  + sd(GPA, na.rm = T)
  + ## [1] 0.3983207
  + sd(SAT, na.rm = T)
  + ## [1] 121.2852
  + View(StudentSurvey)   
    names(StudentSurvey)
  + ## [1] "Year" "Gender" "Smoke" "Award" "HigherSAT"   
    ## [6] "Exercise" "TV" "Height" "Weight" "Siblings"   
    ## [11] "BirthOrder" "VerbalSAT" "MathSAT" "SAT" "GPA"   
    ## [16] "Pulse" "Piercings"
  + plot(VerbalSAT, GPA)
  + 
  + cor.test(VerbalSAT, GPA, use="pairwise.complete.obs")
  + ##   
    ## Pearson's product-moment correlation  
    ##   
    ## data: x and y  
    ## t = 6.9886, df = 343, p-value = 1.454e-11  
    ## alternative hypothesis: true correlation is not equal to 0  
    ## 95 percent confidence interval:  
    ## 0.2570427 0.4421537  
    ## sample estimates:  
    ## cor   
    ## 0.3530485
  + According to our stated problem, our relevant summary statistics are that we have 205 and students with Math being their higher SAT score and 150 students with Verbal being their higher SAT score. The Verbal SAT score ranges from a score of 390 to 800, with a mean of 594.2 and a standard deviation of 74.176. The Math SAT ranges from a score of 400 to 800, with a mean of 609.4 and a standard deviation of 68.49. The combined regular SAT scores range from 800 to 1550 points, with a mean score of 1204 and a standard deviation of 121.285. The GPA ranges from a 2.00 to a 4.00, with a mean of 3.158 and a standard deviation of 0.398. From our scatterplot we can see that there appears to evidence of a positive linear relationship between the Verbal SAT scores and the GPA. Finally, from our correlation test, we can see that although it is not a high correlation close to 1, the correlation of 0.353 tells us there is a small positive linear relationship between Verbal SAT scores and GPA in college.
  1. Fit a linear regression model to predict College GPA based on Verbal SAT scores.
     1. Find the regression model. Write out the model in format. Provide a scatterplot with the regression line added to the plot.
     + #Provide R code and output/graphs   
       model2 <- lm(GPA~VerbalSAT)  
       summary(model2)
     + ##   
       ## Call:  
       ## lm(formula = GPA ~ VerbalSAT)  
       ##   
       ## Residuals:  
       ## Min 1Q Median 3Q Max   
       ## -1.0936 -0.2422 0.0282 0.2550 0.9253   
       ##   
       ## Coefficients:  
       ## Estimate Std. Error t value Pr(>|t|)   
       ## (Intercept) 2.0335506 0.1621397 12.542 < 2e-16 \*\*\*  
       ## VerbalSAT 0.0018929 0.0002709 6.989 1.45e-11 \*\*\*  
       ## ---  
       ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
       ##   
       ## Residual standard error: 0.3732 on 343 degrees of freedom  
       ## (17 observations deleted due to missingness)  
       ## Multiple R-squared: 0.1246, Adjusted R-squared: 0.1221   
       ## F-statistic: 48.84 on 1 and 343 DF, p-value: 1.454e-11
     + plot(VerbalSAT, GPA)  
       abline(model2, col = "red")
     + 
     + The regression model is
     1. Interpret the slope in the context of the situation.
     + For every additional point in your Verbal SAT score, your GPA in college will increase by 0.00189.
     1. Interpret the y-intercept in the context of the situation. Is its interpretation practical in context?
     + A students GPA in college will be 2.033, when the student received a score of zero in the Verbal SAT test. This interpretation is not practical in the context of the problem because it would be almost impossible or rare for a student to receive a zero in the Verbal SAT. The student would have to not take the test or not try at all to get this low score.
  2. Now we will conduct the test to determine if the linear relationship is statistically significant.
     1. Set up the appropriate null and alternative hypothesis verbally and symbolically.
     + Null Hypothesis: The slope of the regression line between the student's college GPA and their score in the Verbal SAT is zero. Alternative Hypothesis: The slope of the regression line between the student's college GPA and their score in the Verbal SAT is not zero.
     1. Use R to find the ANOVA table and/or t-test statistics to test the hypothesis test.
     + #Put Code Here and insert output or graphic into word document  
       anova(model2)
     + ## Analysis of Variance Table  
       ##   
       ## Response: GPA  
       ## Df Sum Sq Mean Sq F value Pr(>F)   
       ## VerbalSAT 1 6.803 6.8029 48.84 1.454e-11 \*\*\*  
       ## Residuals 343 47.776 0.1393   
       ## ---  
       ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
     1. Interpret the meaning of the p-value in the context of the situation.
     + The p-value in the context of the situation is 1.454e-11, which is really small and tells us there is evidence that the slope of regression line is not zero.
     1. State the conclusion of the hypothesis test in the context of the problem.
     + We will reject the null hypothesis, in favor of the alternative hypothesis that the slope of the regression line between the student's college GPA and their score in the Verbal SAT is not zero with a p-vlue of 1.454e-11.
     1. Interpret the value in the context of the model variables.
     + The coefficient of determination tells us that 12.46% of the variablity in college GPA, is explained by the linear regression with the score received in the Verbal SAT test.
     1. Given the value, is the relationship observed practically significant? What might explain apparent "disagreement" between the practical and statistical significance of the linear relationship?
     + Given our coefficient of determination, the relationship observed is not practically significant, since only a small portion of the variability of the college GPA is explained by the regression line with the Verbal SAT scores. This disagreement with the statistical significance might be explained by the collinearity with the SAT variable.
  3. We have to determine whether the results of the analysis are trustworthy, i.e., the assumptions to perform the Regression Analysis are met by the data.
     1. Provide and interpret the results of the QQPlot and Shapiro-Wilk Test to determine if the normality of residuals is a satisfied assumption.
     + #Provide R code and output/graphs   
       shapiro.test(residuals(model2))
     + ##   
       ## Shapiro-Wilk normality test  
       ##   
       ## data: residuals(model2)  
       ## W = 0.98764, p-value = 0.00493
     + plot(model2, 2)
     + 
     + According to the results of the QQPlot and the Shapiro-Wilk test, we conclude that the assumption of normality of residuals is satisfied. The QQ plot looks good and in the middle but is not approximately a straight line relationship between quantiles of our data and quantiles from the normal distribution. In the shapiro-wilk test we can see the p-value is 0.00493, which is still smaller than 0.05, therefore we reject the null hypothesis that the residuals are normally distributed.
     1. Provide and interpret the Residual Plot to test Equal Variance of the residuals.
     + #Provide R code and output/graphs   
       plot(model2, 1)
     +  >According to our residuals plot, we can se that there is equal variance of the residuals since there is no obvious pattern of trumpeting to the points about zero, they appear to be equally distributed about zero.
     1. Interpret the results of the Residual Plot to test the Independence of Residuals assumption.
     + According to the result of our residual plot, we can see that there is no obvious pattern, such as curvature, in the residuals versus the fitted values. Since the points are randomly distributed about zero we conclude that the assumpiton of independece of residuals in satisfied.
     1. Provide and interpret the Cook's Distance plot to test if there are outliers in the data.
     + #Provide R code and output/graphs   
       plot(model2, 5)
     + 
     + We can clearly see from our Cook's Distance plot that there are no outliers. The data points are all clustered together and the cook's distance contour lines are not even visible in the graph. This tells us that there are no points with a distance bigger or equal to 0.5. Since we have small distances from the central cluseter of observations it means that removing these observations has little effect on the regression result.
     1. After assessing all assumptions, do you have an concerns about the analysis?
     + After assessing all assumptions, the main concern that we have is that the data might not be normally distributed, although it does appear close to it from the QQ plot. The p-value of 0.00493 given to us by the Shapiro-wilk test, tells us that we can reject the null hypothesis that states that the residuals are normally distributed.
  4. Let's use the model to make predictions about performance in college.
     1. Predict an individuals College GPA given a Verbal SAT score of 600 with a 95% prediction interval. Interpret your interval in the context of the problem.
     + #Put your code and output here   
       predict(model2, interval = "predict",newdata = data.frame(VerbalSAT = 600))
     + ## fit lwr upr  
       ## 1 3.169316 2.43417 3.904462
     + We are 95% confident that the College GPA for a Verbal SAT score of 600 is between 2.434 and 3.904.
     1. Predict the average College GPA for students with a Verbal SAT score of 550 with a 95% confidence interval. Interpret your interval in the context of the problem.
     + #Put your code and output here   
       predict(model2, interval = "confidence",newdata = data.frame(VerbalSAT = 600))
     + ## fit lwr upr  
       ## 1 3.169316 3.129665 3.208967
     + We are 95% confident that the true mean College GPA for a Verbal SAT score of 550 is between 3.129 and 3.208.
     1. Notice the width of the intervals for i. and ii. Why is the prediction interval so much wider than the confidence interval?
     + The prediction interval is so much wider because since we are predicting an individual response and not a mean response, our interval must be larger to account for the error due to extrapolation. As we move further away from the range of x, VerbalSAT, we must account for an increase in the error of prediction due to extrapolation.

1. Now we are going to look at a model that examines GPA and MathSAT instead of verbal SAT.
   1. Provide the relevant summary statistics for the study (mean and sd of variables, scatterplot, correlation). Provide and interpretation of the summary statistics (specifically the scatterplot and correlation). Is there evidence of a linear relationship? (you may need to include na.rm=T again for mean() and sd() and input "pairwise.complete.obs" as an argument in cor()).
   * #Provide R code and output/graphs   
     plot(MathSAT, GPA)
   * 
   * mean(MathSAT, na.rm = T)
   * ## [1] 609.4365
   * sd(MathSAT, na.rm = T)
   * ## [1] 68.49007
   * mean(GPA, na.rm = T)
   * ## [1] 3.157942
   * sd(GPA, na.rm = T)
   * ## [1] 0.3983207
   * cor.test(MathSAT, GPA, use="pairwise.complete.obs")
   * ##   
     ## Pearson's product-moment correlation  
     ##   
     ## data: x and y  
     ## t = 5.4493, df = 343, p-value = 9.677e-08  
     ## alternative hypothesis: true correlation is not equal to 0  
     ## 95 percent confidence interval:  
     ## 0.1821075 0.3766303  
     ## sample estimates:  
     ## cor   
     ## 0.2822677
   * The Math SAT has a mean of 609.4 and a standard deviation of 68.49. The GPA has a mean of 3.158 and a standard deviation of 0.398. From our scatterplot we can see that there does not appear to be evidence of a linear relationship between the Math SAT scores and the GPA. From our correlation test we can see that with a correlation of 0.282267, there does not appear to be a strong linear relationship between Math SAT scores and a students college GPA.
   1. Fit a linear regression model to predict College GPA based on Math SAT scores.
      1. Find the regression model. Write out the model in format. Provide a scatterplot with the regression line added to the plot.
      * #Provide R code and output/graphs   
        model3 <- lm(GPA~MathSAT)  
        summary(model)
      * ##   
        ## Call:  
        ## lm(formula = Time ~ Distance)  
        ##   
        ## Residuals:  
        ## Min 1Q Median 3Q Max   
        ## -26.124 -6.693 -2.027 4.246 106.824   
        ##   
        ## Coefficients:  
        ## Estimate Std. Error t value Pr(>|t|)   
        ## (Intercept) 7.12034 0.90659 7.854 2.5e-14 \*\*\*  
        ## Distance 1.21115 0.03977 30.454 < 2e-16 \*\*\*  
        ## ---  
        ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
        ##   
        ## Residual standard error: 12.26 on 498 degrees of freedom  
        ## Multiple R-squared: 0.6506, Adjusted R-squared: 0.6499   
        ## F-statistic: 927.4 on 1 and 498 DF, p-value: < 2.2e-16
      * plot(MathSAT, GPA)  
        abline(model3, col = "red")
      * 
      * The model is
      1. Interpret the slope in the context of the situation.
      * For every additional point in your Math SAT score, your GPA in college will increase by 0.00165.
      1. Interpret the y-intercept in the context of the situation. Is its interpretation practical in context?
      * A students GPA in college will be 2.147, when the student received a score of zero in the Math SAT test. This interpretation is not practical in the context of the problem because it would be really rare for a student to receive a zero in the Math SAT test.
   2. Now we will conduct the test to determine if the linear relationship is statistically significant.
      1. Set up the appropriate null and alternative hypothesis verbally and symbolically.
      * Null Hypothesis: The slope of the regression line between the student's college GPA and their score in the Math SAT is zero. Alternative Hypothesis: The slope of the regression line between the student's college GPA and their score in the Math SAT is not zero.
      1. Use R to find the ANOVA table and/or t-test statistics to test the hypothesis test.
      * #Put Code Here and insert output or graphic into word document  
        anova(model3)
      * ## Analysis of Variance Table  
        ##   
        ## Response: GPA  
        ## Df Sum Sq Mean Sq F value Pr(>F)   
        ## MathSAT 1 4.349 4.3486 29.694 9.677e-08 \*\*\*  
        ## Residuals 343 50.230 0.1464   
        ## ---  
        ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
      * summary(model3)
      * ##   
        ## Call:  
        ## lm(formula = GPA ~ MathSAT)  
        ##   
        ## Residuals:  
        ## Min 1Q Median 3Q Max   
        ## -1.23862 -0.23935 0.04483 0.27719 0.99300   
        ##   
        ## Coefficients:  
        ## Estimate Std. Error t value Pr(>|t|)   
        ## (Intercept) 2.1466877 0.1867166 11.497 < 2e-16 \*\*\*  
        ## MathSAT 0.0016544 0.0003036 5.449 9.68e-08 \*\*\*  
        ## ---  
        ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
        ##   
        ## Residual standard error: 0.3827 on 343 degrees of freedom  
        ## (17 observations deleted due to missingness)  
        ## Multiple R-squared: 0.07968, Adjusted R-squared: 0.07699   
        ## F-statistic: 29.69 on 1 and 343 DF, p-value: 9.677e-08
      1. Interpret the meaning of the p-value in the context of the situation.
      * the p-value in the context of the situation is 9.677e-08 and it tells us there is enough evidence to say that the slope of the regression line between the student's college GPA and their score in the Math SAT is not zero.
      1. State the conclusion of the hypothesis test in the context of the problem.
      * We reject the null hypothesis, in favor of the alternative hypothesis that he slope of the regression line between the student's college GPA and their score in the Math SAT is not zero, with a p-value of 9.677e-08.
      1. Interpret the value in the context of the model variables.
      * Insert Answer Here
      1. Given the value, is the relationship observed practically significant? What might explain apparent "disagreement" between the practical and statistical significance of the linear relationship?
   3. We have to determine whether the results of the analysis are trustworthy, i.e., the assumptions to perform the Regression Analysis are met by the data.
      1. Provide and interpret the results of the QQPlot and Shapiro-Wilk Test to determine if the nomality of residuals is a satisfied assumption.
      * #Provide R code and output/graphs   
        shapiro.test(residuals(model3))
      * ##   
        ## Shapiro-Wilk normality test  
        ##   
        ## data: residuals(model3)  
        ## W = 0.98263, p-value = 0.0003546
      * plot(model3, 2)
      * 
      * Insert Answer Here
      1. Provide and interpret the Residual Plot to test Equal Variance of the residuals.
      * #Provide R code and output/graphs   
        plot(model3, 1)
      *  >Insert Answer Here
      1. Interpret the results of the Residual Plot to test the Independence of Residuals assumption.
      * Insert Answer Here
      1. Provide and interpret the Cook's Distance plot to test if there are outliers in the data.
      * #Provide R code and output/graphs  
        plot(model3, 5)
      * 
      * Insert Answer Here
      1. After assessing all assumptions, do you have an concerns about the analysis?
      * Insert Answer Here
   4. Compare the new model predicting college GPA with Math SAT Scores to the model predicting GPA with Verbal SAT Scores. Which model would be the best (relative to each other) model to predict college GPA? Justify your answer with the statistical analysis.
   * Insert Answer Here
   1. In both models there is a lot of unexplained variability in College GPA. What might explain that leftover variability?
   * Insert Answer Here