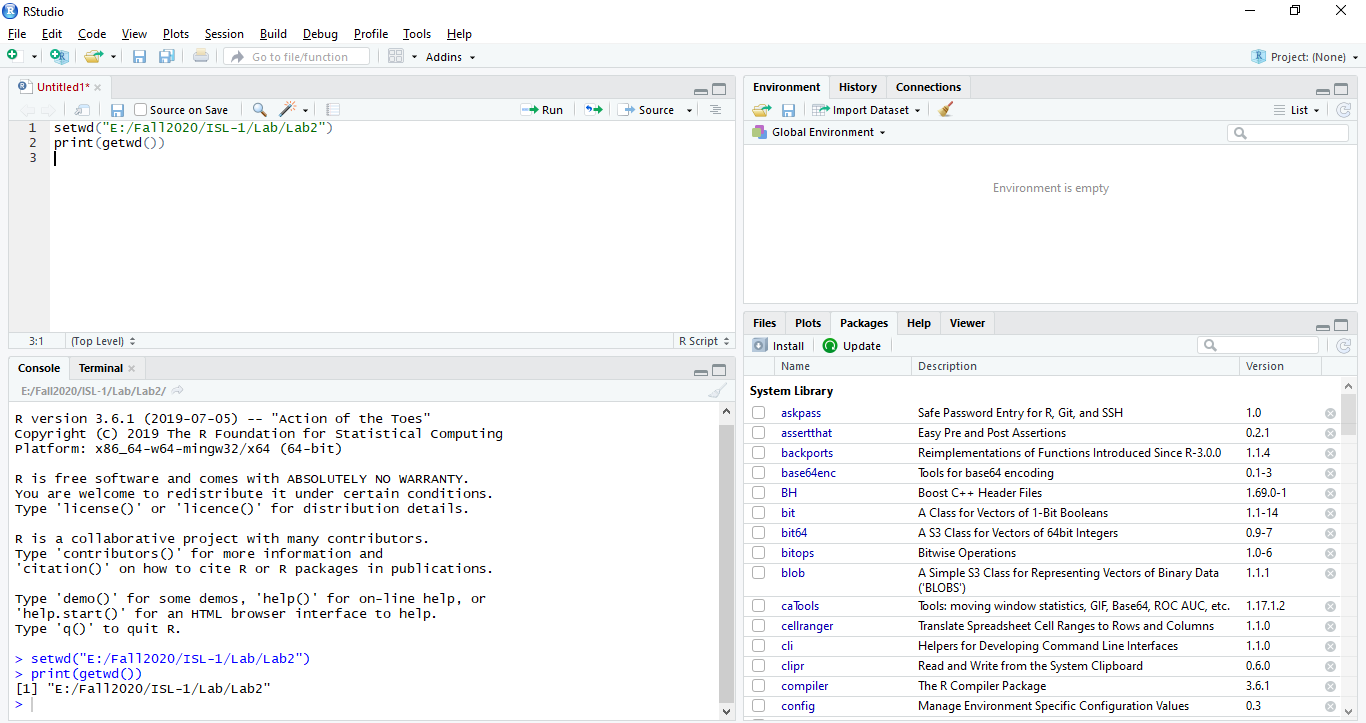
**Introduction to Statistical Learning- Lab2**

**Name: Sandeep Reddy Salkuti**

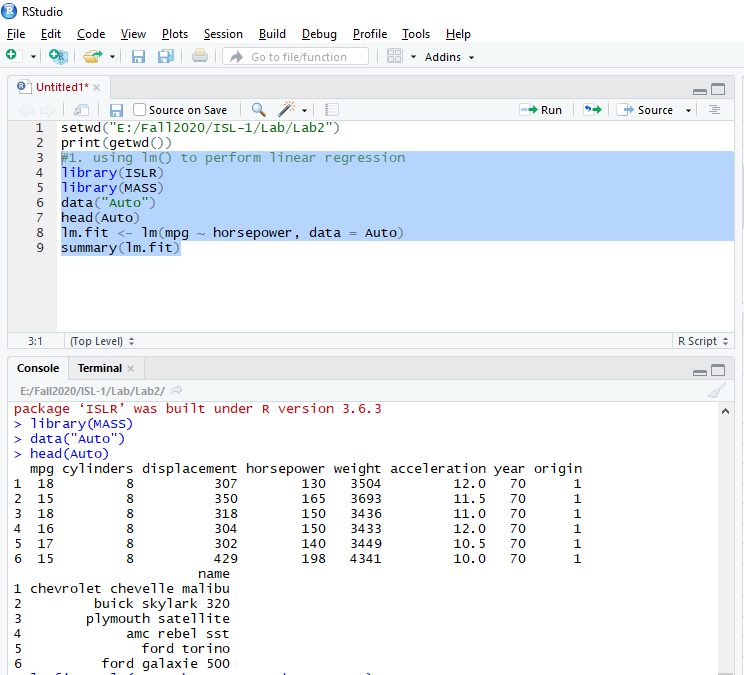
**Student id: 16296868**

**Email:** [**sswf7@umsystem.edu**](mailto:sswf7@umsystem.edu)

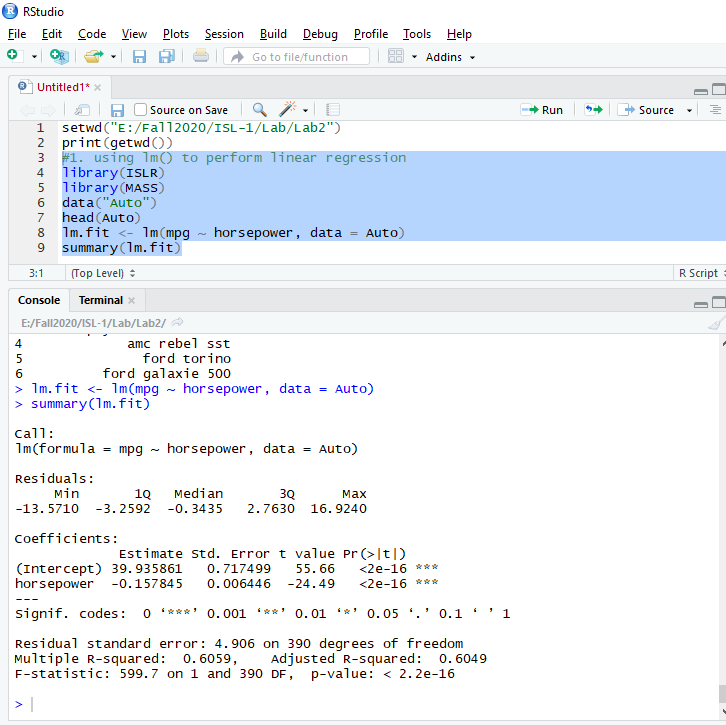
1. **View the video at the following URL and install R https://www.youtube.com/watch?v=5ONFqIk3RFg You may download the R Code for Labs and the Data Sets to use from the textbook website. http://www-bcf.usc.edu/˜gareth/ISL/**



1. **a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results**



**Applying linear regression model**



1. **Is there a relationship between the predictor and the response?**

The p-values from the above for the regression coefficients are nearly zero (2e-16). This shows statistical significance, which in turn means that there exists a relationship.

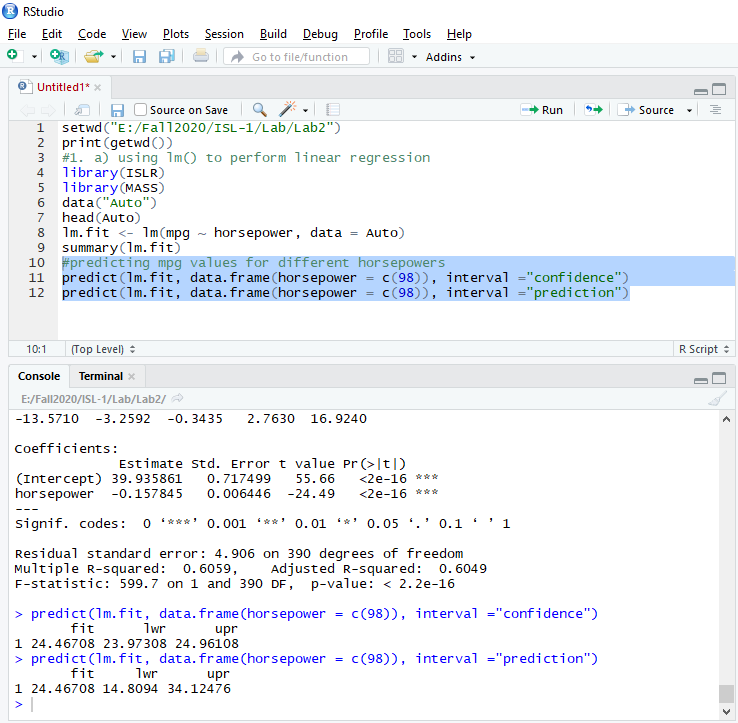
1. **How strong is the relationship between the predictor and the response?**

The R^{2} value indicates that about 61% of the variation in the response variable ( mpg) is due to the predictor variable (horsepower).

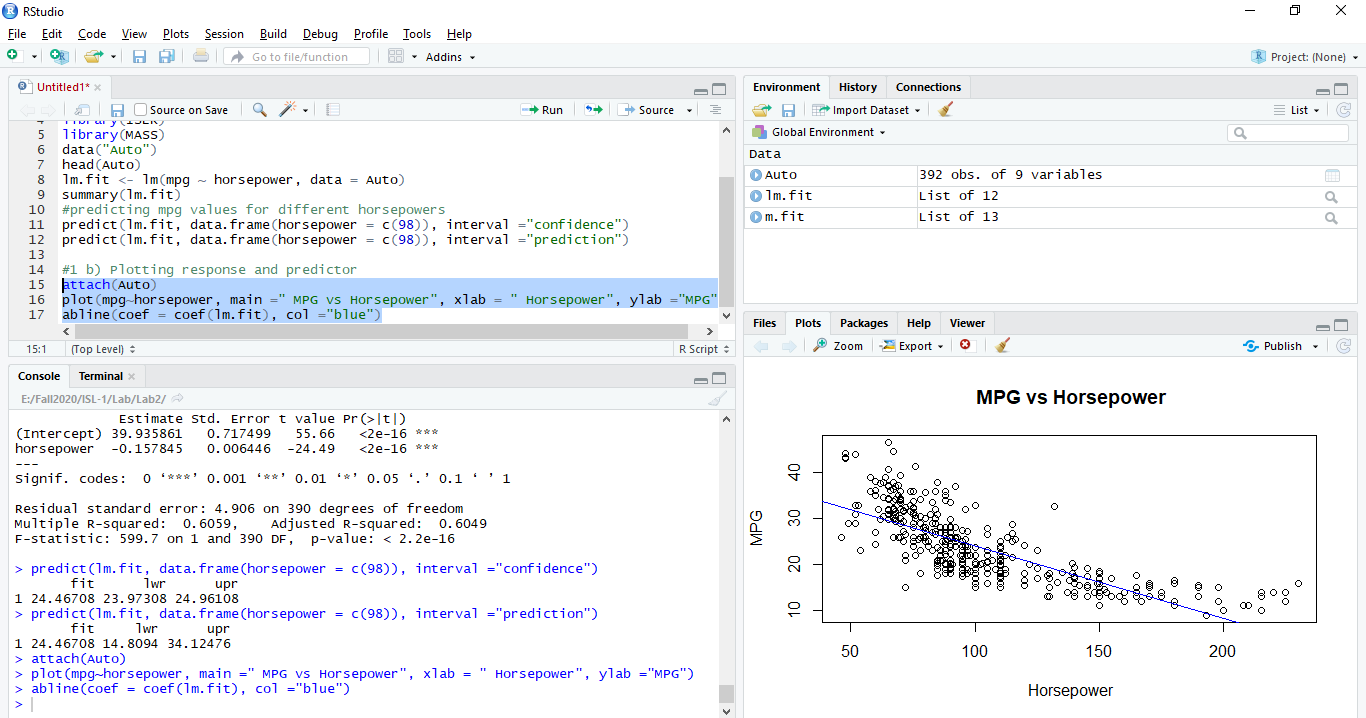
1. **Is the relationship between the predictor and the response positive or negative?**

The regression coefficient for ‘horsepower’ is negative. Hence, the relationship is negative.

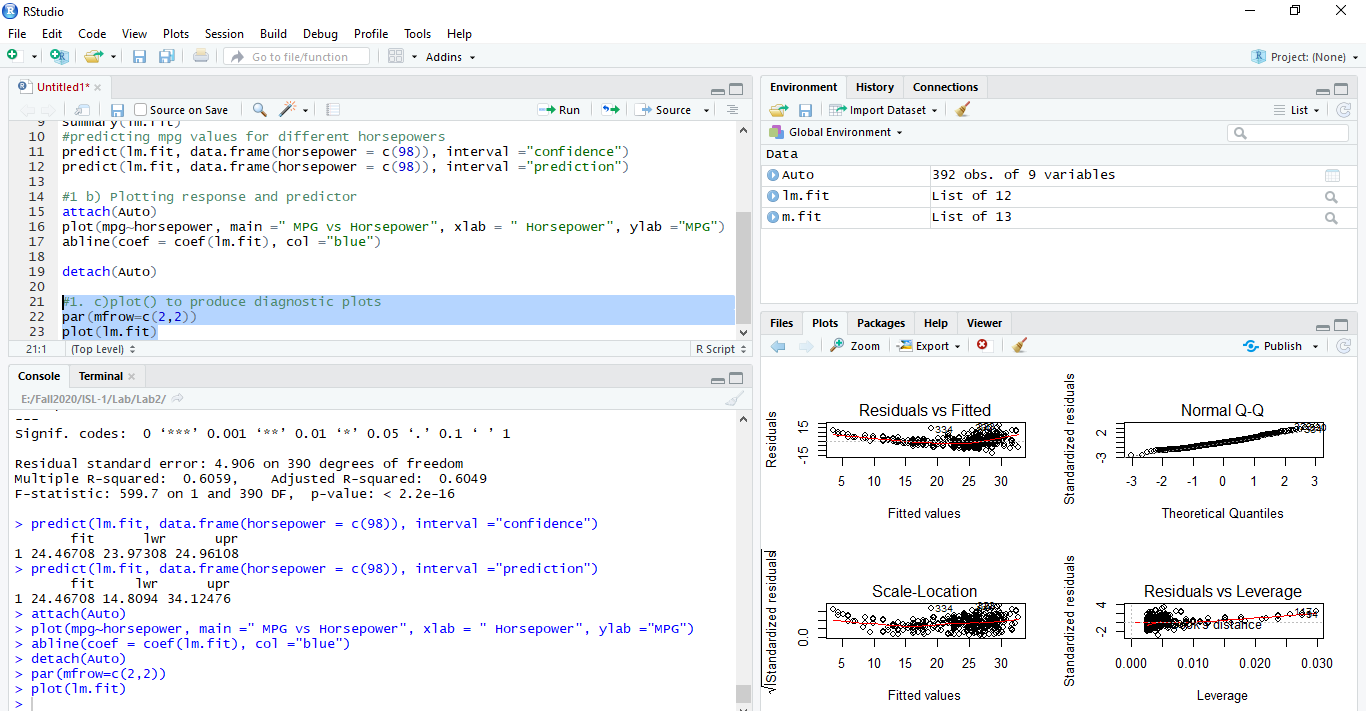
1. **What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?**

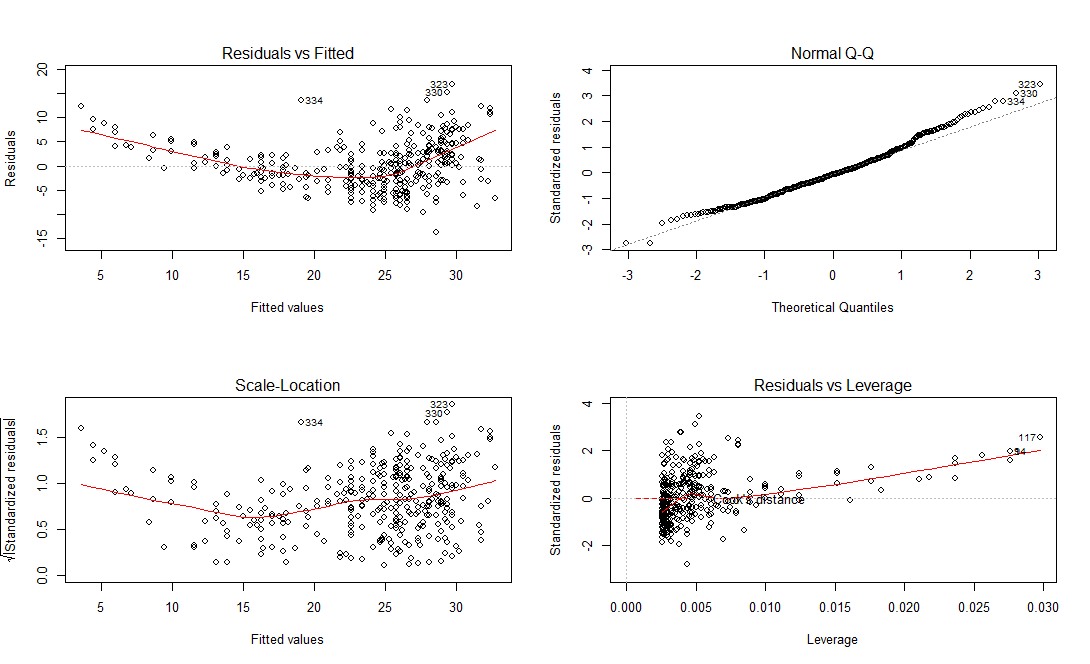


**(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.**



**(c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.**

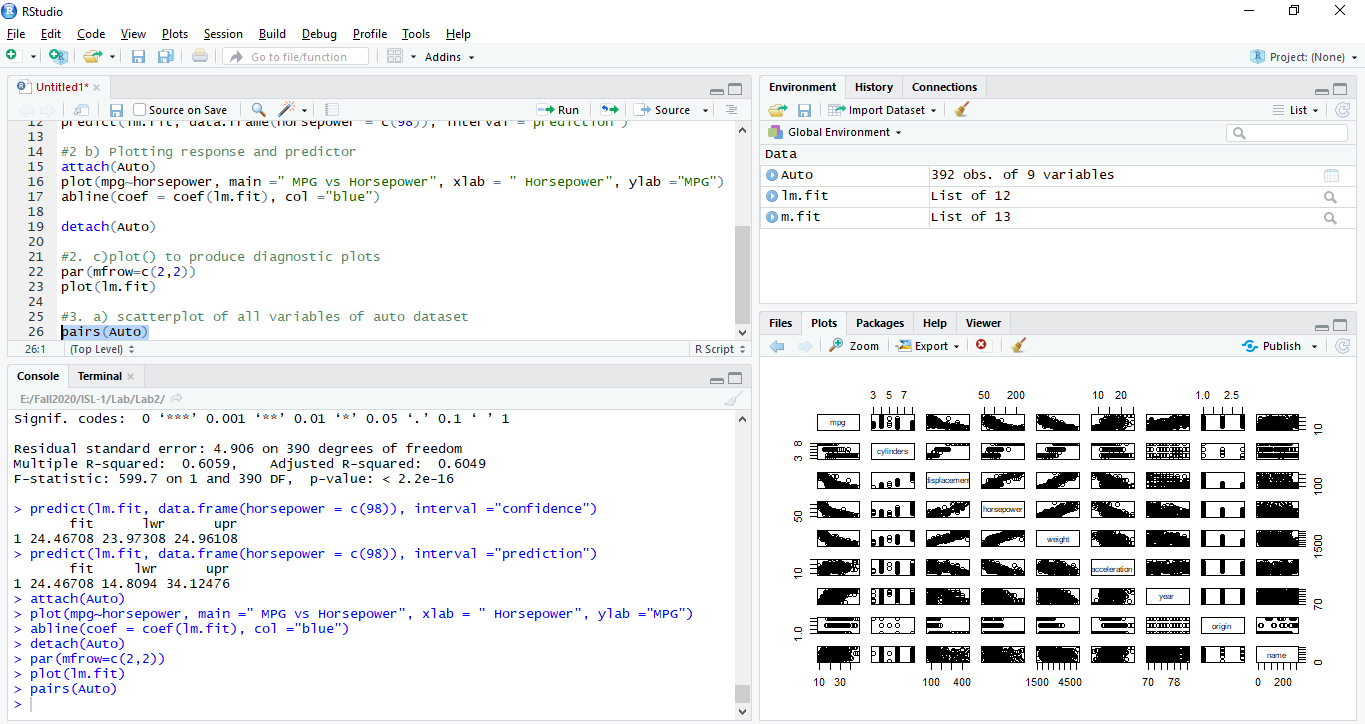


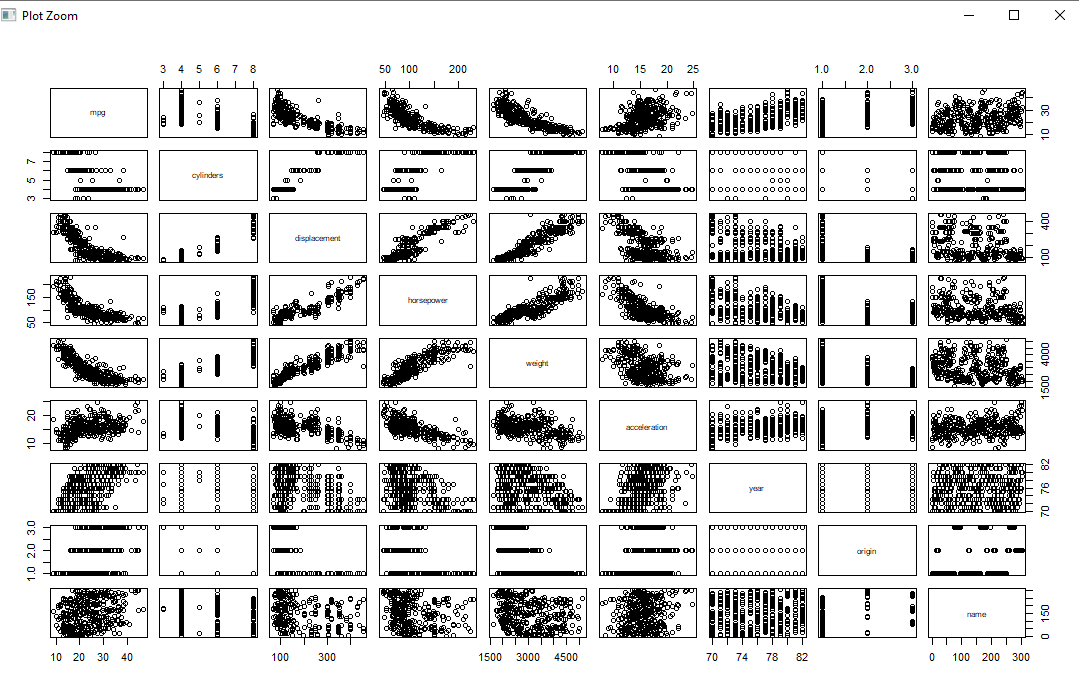


The first plot shows a pattern (U-shaped) between the residuals and the fitted values. This indicates a non-linear relationship between the predictor and response variables. The second plot shows that the residuals are normally distributed. The third plot shows that the variance of the errors is constant. Finally, the fourth plot indicates that there are no leverage points in the data.

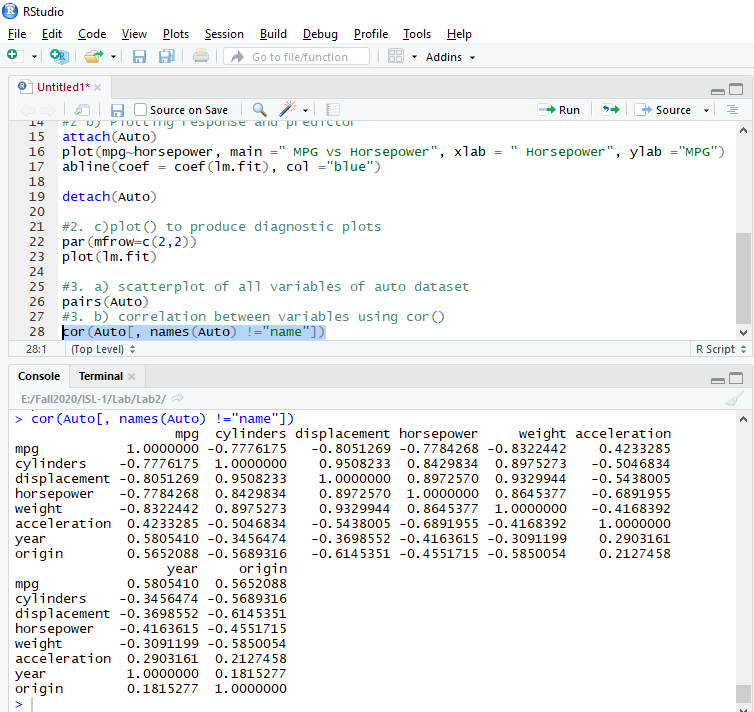
**3. This question involves the use of multiple linear regression on the Auto data set.**

**(a) Produce a scatterplot matrix which includes all of the variables in the data set.**

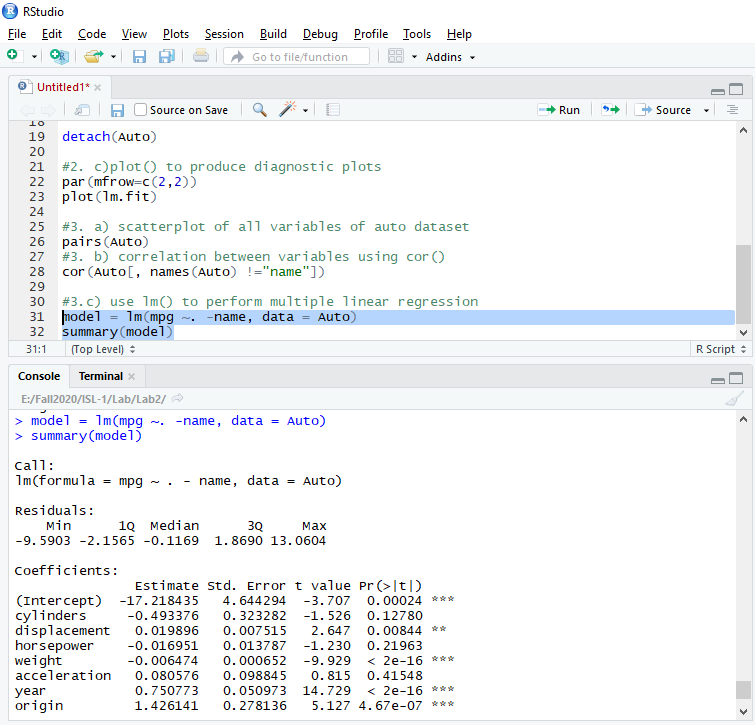




**b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.**



**(c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results.**



1. **Is there a relationship between the predictors and the response?**

Yes, there is. However, some predictors do not have a statistically significant effect on the response. R-squared value implies that 82% of the changes in the response can be explained by the predictors in this regression model.

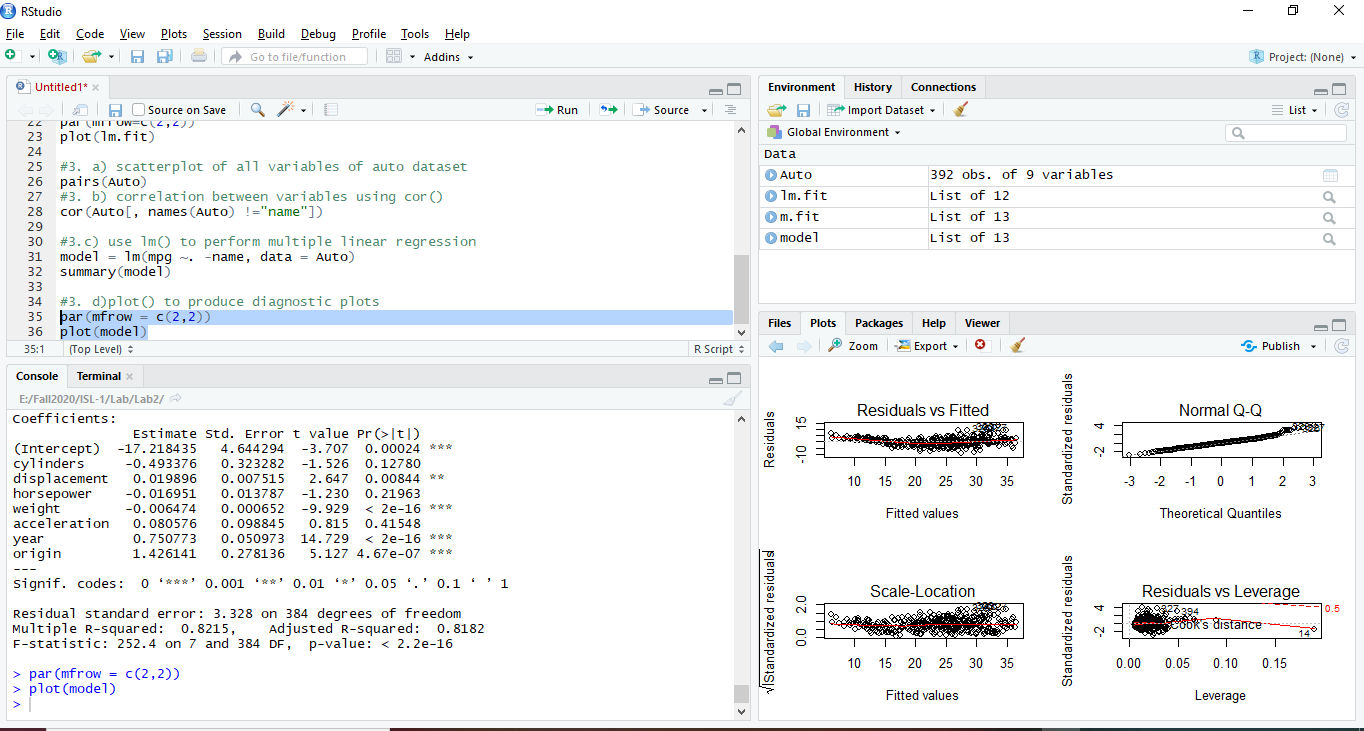
1. **Which predictors appear to have a statistically significant relationship to the response?**

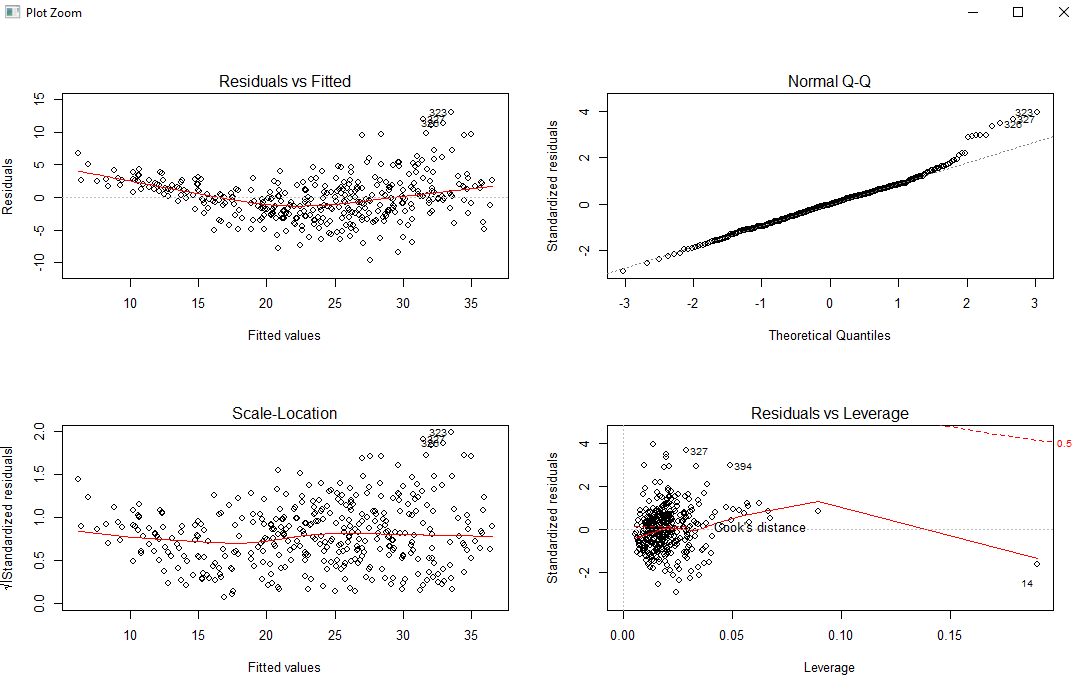
Displacement, weight, year, origin have significant relationship to the response.

1. **What does the coefficient for the year variable suggest?**

When every other predictor held constant, the mpg value increases with each year that passes. Specifically, mpg increase by 1.43 each year.

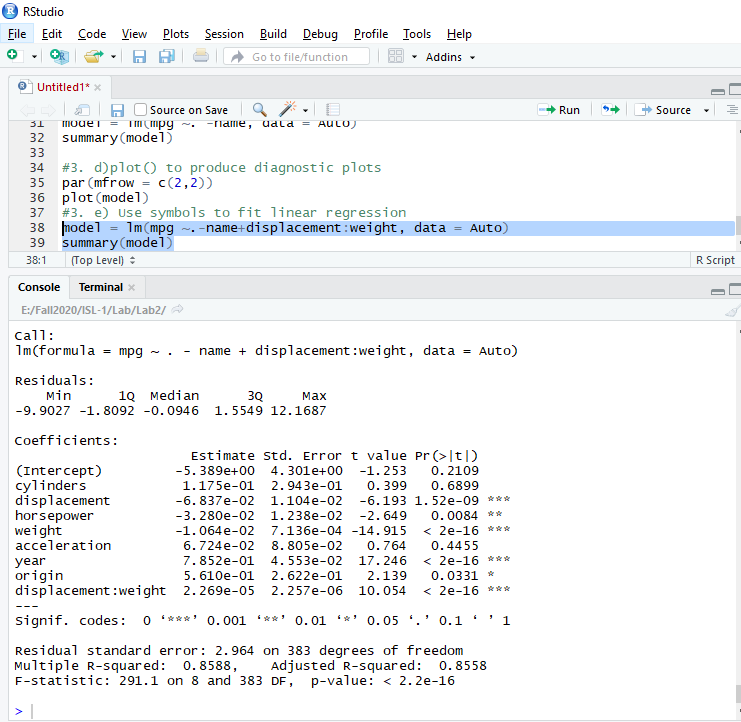
**(d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?**

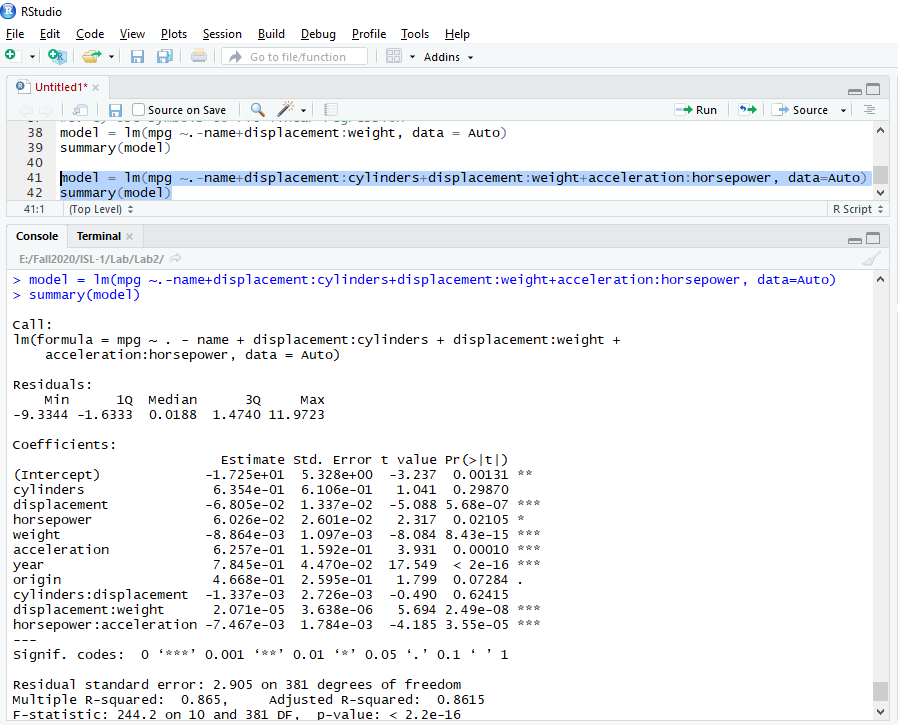


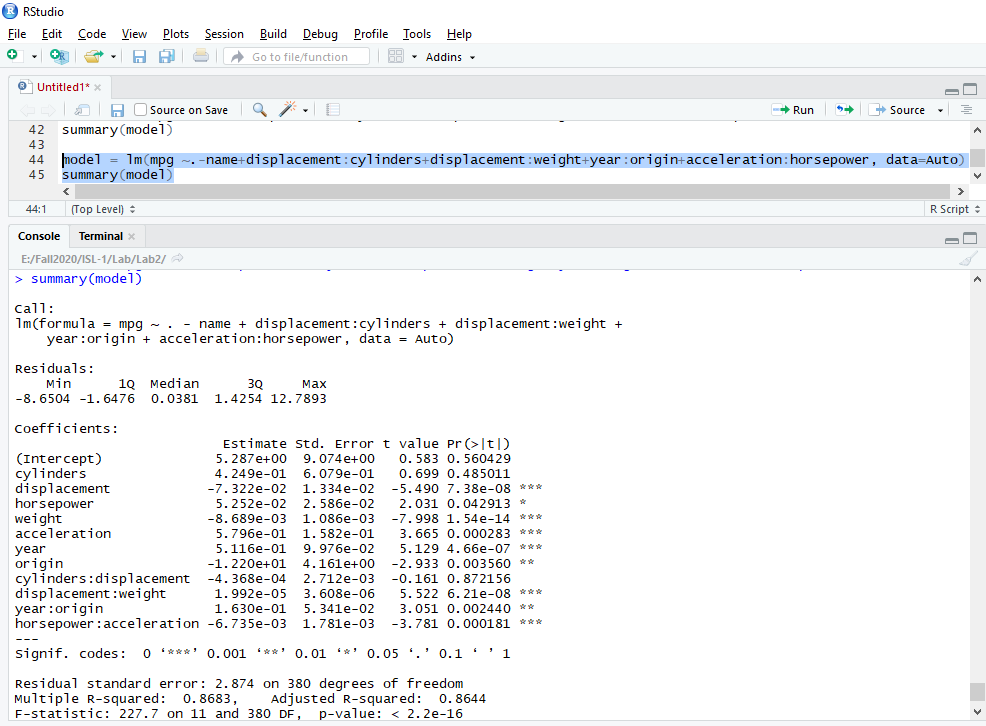


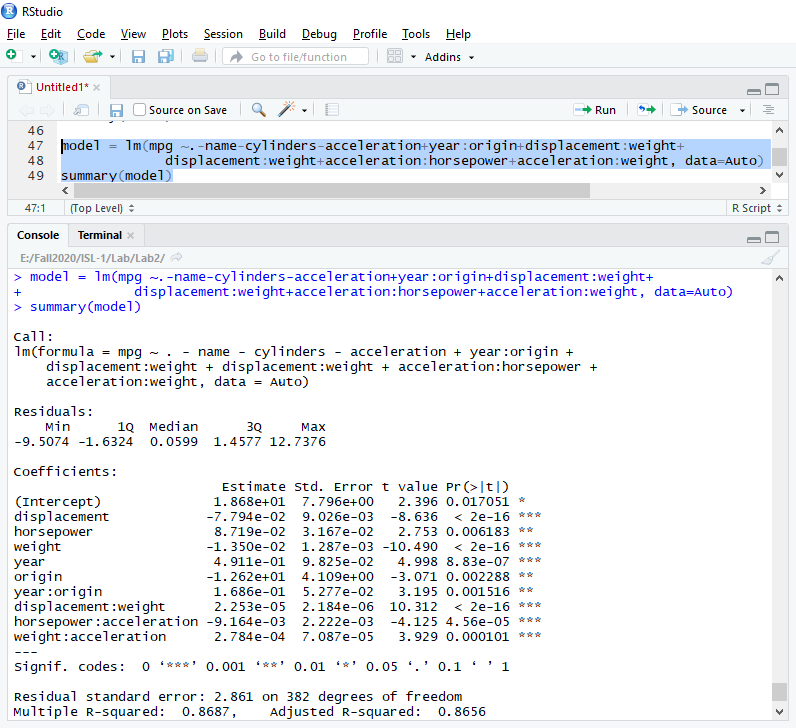
1. The first graph shows that there is a non-linear relationship between the response and the predictors
2. The second graph shows that the residuals are normally distributed and right skewed
3. The third graph shows that the constant variance of error assumption is not true for this model
4. The fourth graphs shows that there are no leverage points. However, there on observation that stands out as a potential leverage point

**(e) Use the ∗ and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?**



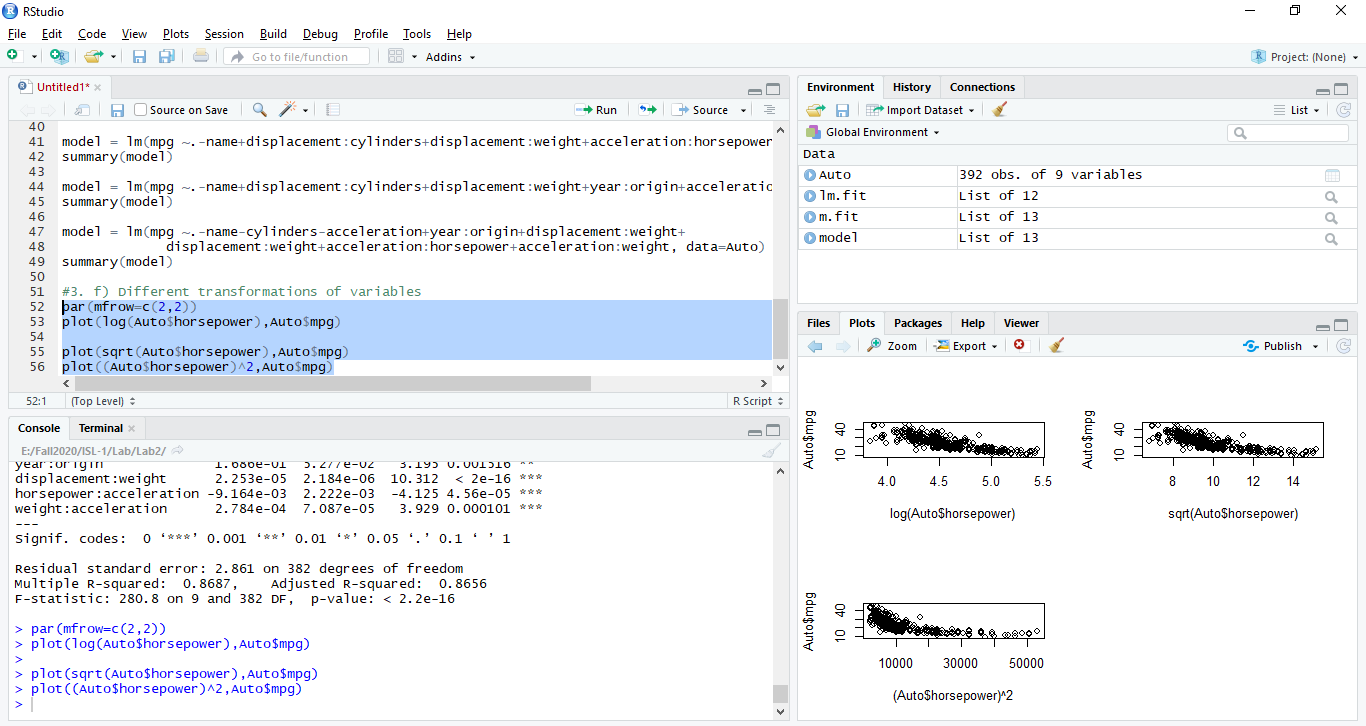


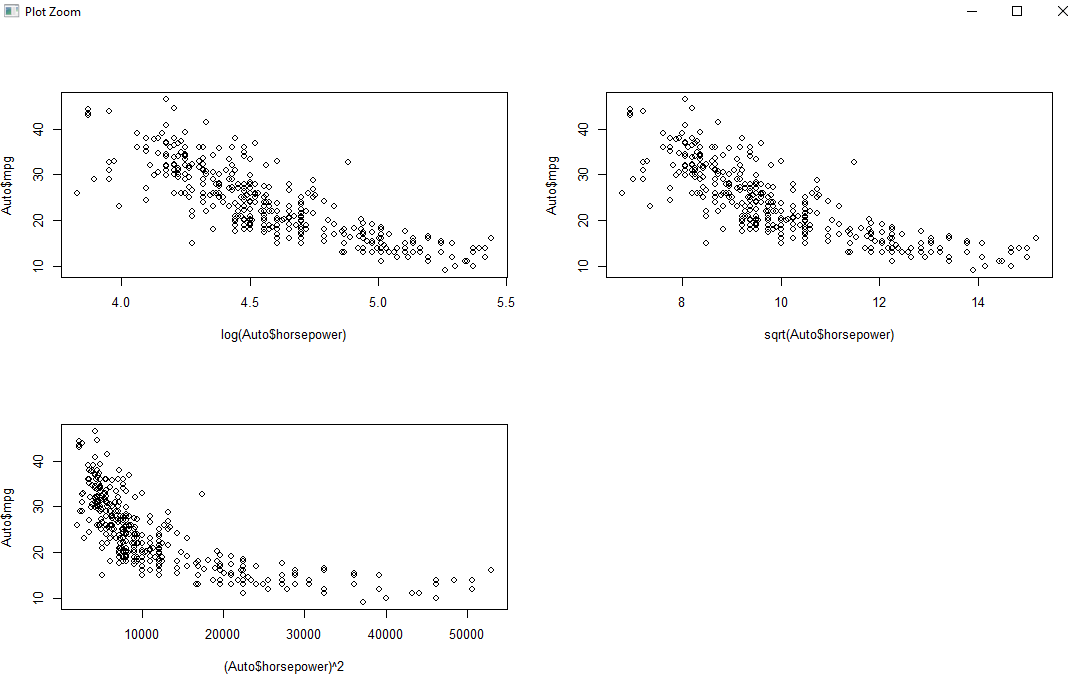




From all the 4 models, the last model is the only one with all variables being significant. And, based on results from a few trials not show here, it is very likely that it is the best combination of predictors and interaction terms. The R-squared statistics estimates that 87% of the changes in the response can be explained by this particular set of predictors (single and interaction.) A higher value was not obtained from the trials.

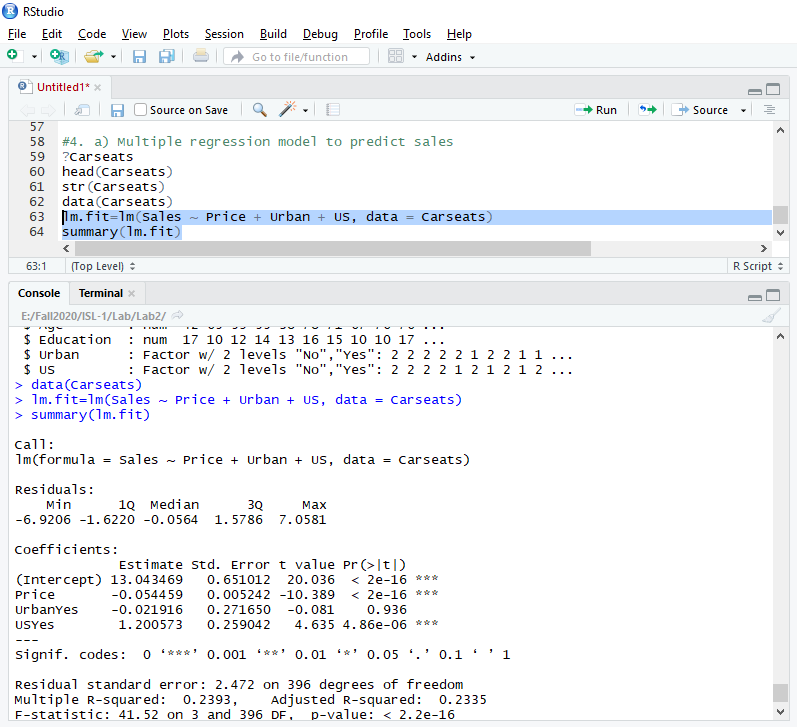
**(f) Try a few different transformations of the variables, such as log(X),√ X, X2 . Comment on your findings.**





**4. This question should be answered using the Carseats data set.**

**(a) Fit a multiple regression model to predict Sales using Price, Urban, and US**



**(b) Provide an interpretation of each coefficient in the model. Be careful some of the variables in the model are qualitative!**

**Price:**

“Price” variable may be interpreted by saying that the average effect of a price increase of 1 dollar is a decrease of 54.4588492 units in sales all other predictors remaining fixed.

**Urban:**

“Urban” variable may be interpreted by saying that on average the unit sales in urban location are 21.9161508 units less than in rural location all other predictors remaining fixed.

**US:**

“US” variable may be interpreted by saying that on average the unit sales in a US store are 1200.5726978 units more than in a non US store all other predictors remaining fixed.

**(c) Write out the model in equation form, being careful to handle the qualitative variables properly.**

Model in equation form can be written as:

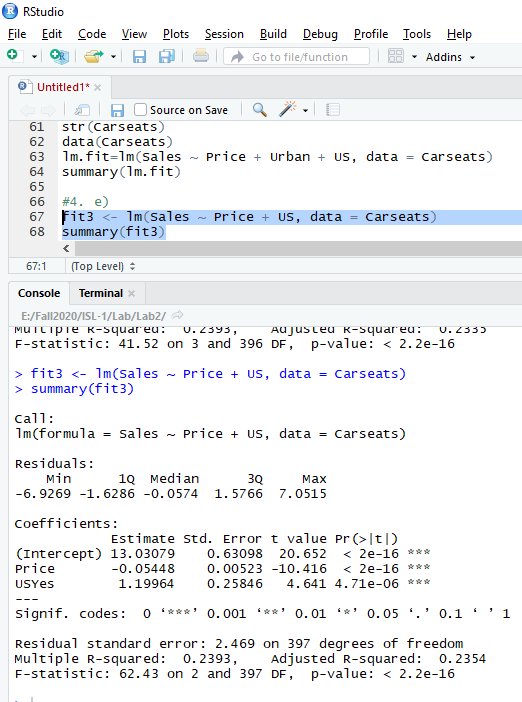
Sales=13.0434689+(−0.0544588)×Price+(−0.0219162)×Urban+(1.2005727)×US+ ε

If store is in urban then urban=1 otherwise urban=0 If store is in us then us=1 otherwise us=0.

**(d) For which of the predictors can you reject the null hypothesis H0 : βj = 0**

For US and Price variables we can reject null hypothesis

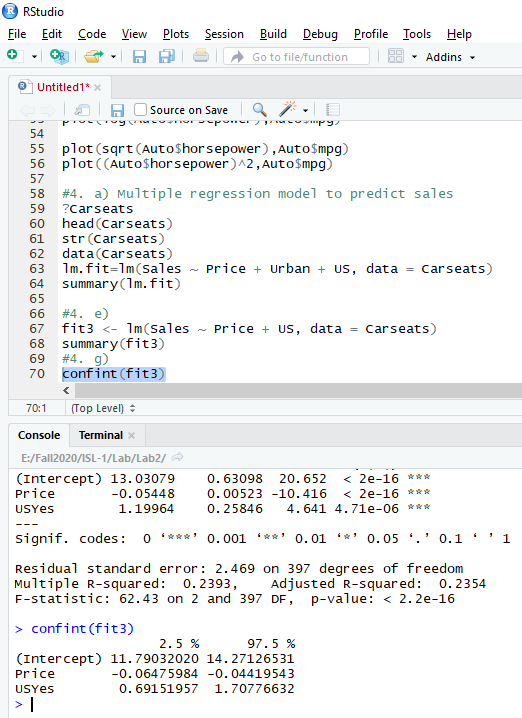
**(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.**



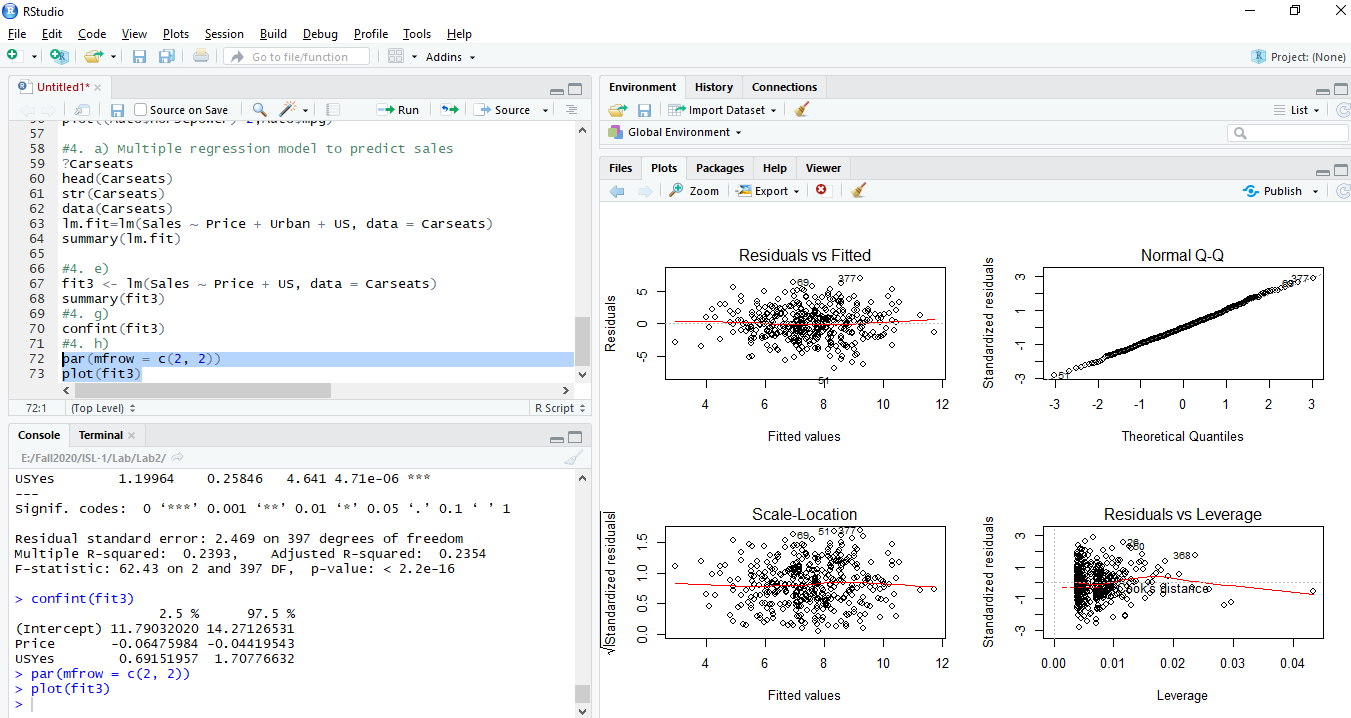
**(f) How well do the models in (a) and (e) fit the data?**

By looking R^2 values for smaller model is better than bigger model. Variability of 24% is shown by the model.

**(g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).**



**(h) Is there evidence of outliers or high leverage observations in the model from (e)?**



The plots of residual vs leverage indicates that there are few outliers and some leverage points.

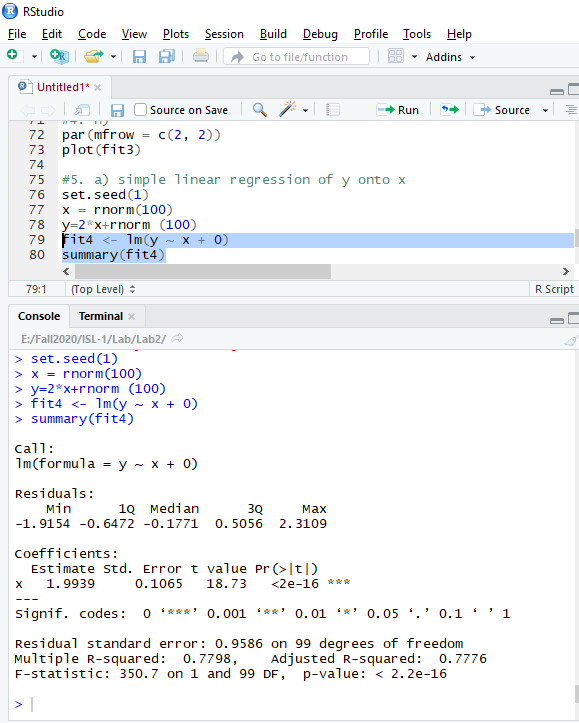
1. **In this problem we will investigate the t-statistic for the null hypothesis H0 : β = 0 in simple linear regression without an intercept. To begin, we generate a predictor x and a response y as follows.**

**> set.seed(1)**

**> x = rnorm(100)**

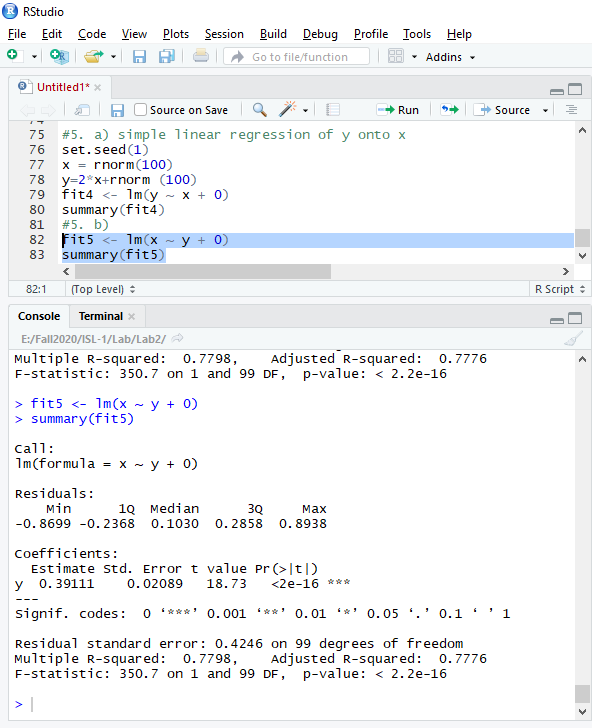
**>y=2\*x+rnorm (100)**

**(a) Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate βˆ, the standard error of this coefficient estimate, and the t-statistic and p-value associated with the null hypothesis H0 : β = 0. Comment on these results. (You can perform regression without an intercept using the command lm(y ∼ x + 0).)**



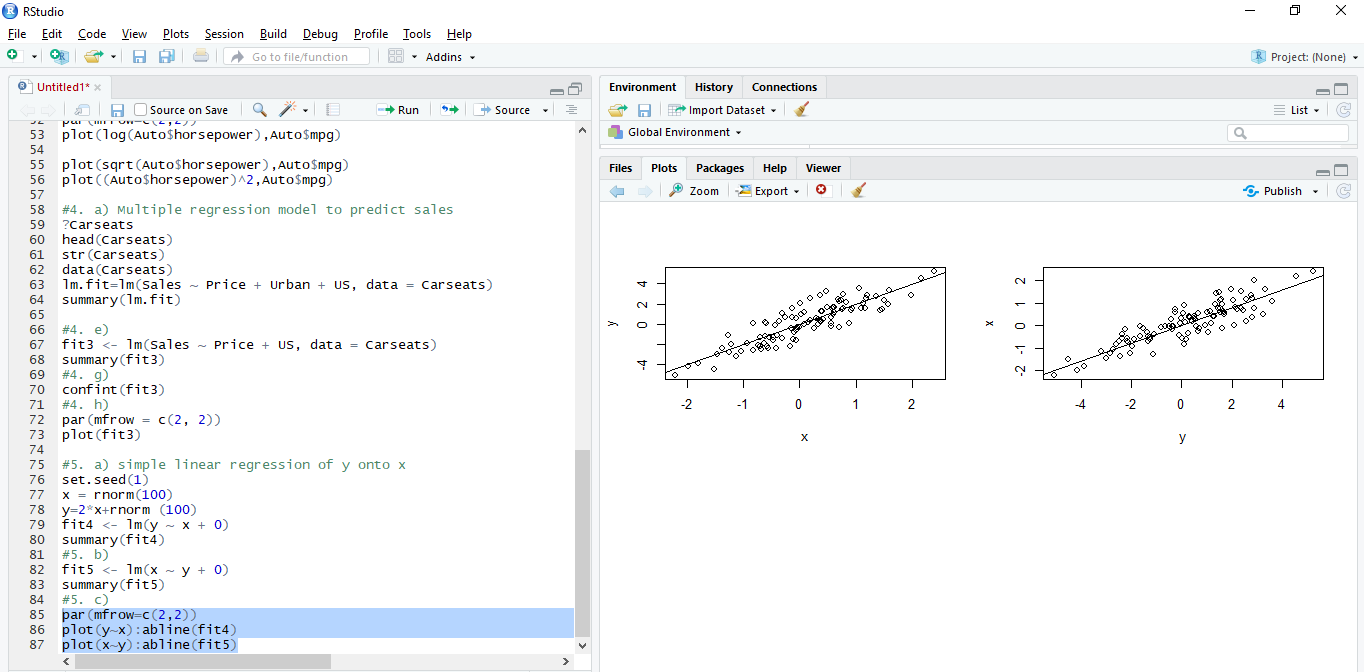
Based on these results, the predictor ‘x’ is statistically significant for estimating the response variable ‘y’.

**(b) Now perform a simple linear regression of x onto y without an intercept, and report the coefficient estimate, its standard error, and the corresponding t-statistic and p-values associated with the null hypothesis H0 : β = 0. Comment on these results.**



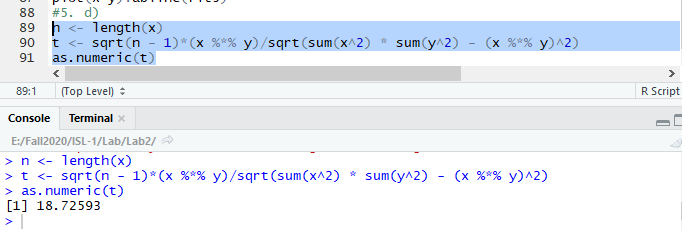
Same observation as the previous model. The predictor variable is significance.

**(c) What is the relationship between the results obtained in (a) and (b)?**



We obtain the same value for the t-statistic and consequently the same value for the corresponding p-value. Both results in (a) and (b) reflect the same line created in (a). In other words, y=2x+εy=2x+ε could also be written x=0.5(y−ε)x=0.5(y−ε).

**(d) For the regression of Y onto X without an intercept, the t-statistic for H0 : β = 0 takes the form β/SE ˆ (βˆ), where βˆ is given by (3.38), and where SE(βˆ) = vuut ∑n i=1(yi − xiβˆ) 2 (n − 1) ∑n i ′=1 x 2 i ′ . (These formulas are slightly different from those given in Sections 3.1.1 and 3.1.2, since here we are performing regression without an intercept.) Show algebraically, and confirm numerically in R, that the t-statistic can be written as ( √ n − 1) ∑n i=1 xiyi √ ( ∑n i=1 x 2 i )(∑n i ′=1 y 2 i ′) − ( ∑n i ′=1 xi ′yi ′)2**

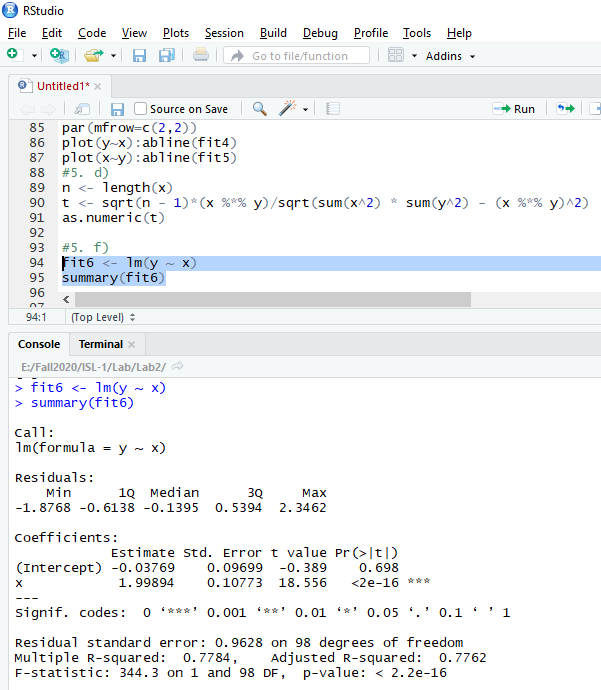


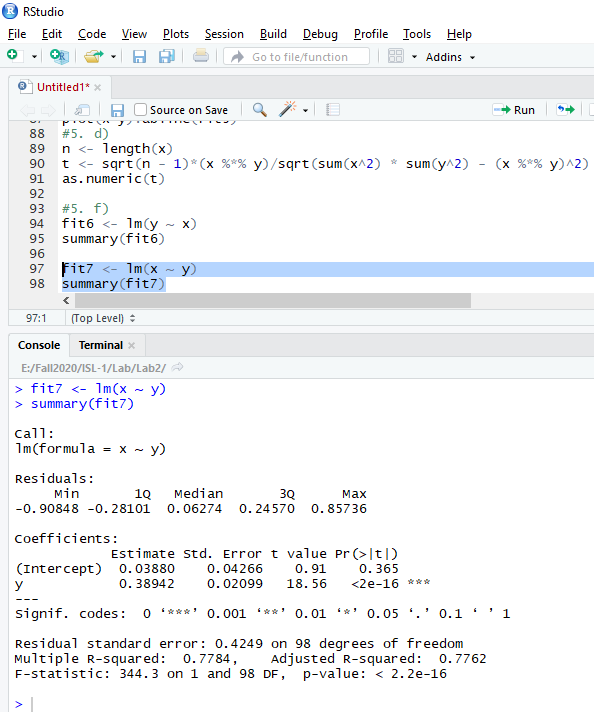
We may see that t-value is same as the t-statistic in above 5b.

**(e) Using the results from (d), argue that the t-statistic for the regression of y onto x is the same as the t-statistic for the regression of x onto y.**

It is clear that if we replace xi by yi in the formula for the t-statistic, the result will be the same.

**(f) In R, show that when regression is performed with an intercept, the t-statistic for H0 : β1 = 0. is the same for the regression of y onto x as it is for the regression of x onto y**





The summary tables shows that the t-value are the same (by approximation).