**HOME WORK TWO Qian Yu**

**ST635 Intermediate Statistical Modeling for Business Fall 2017**

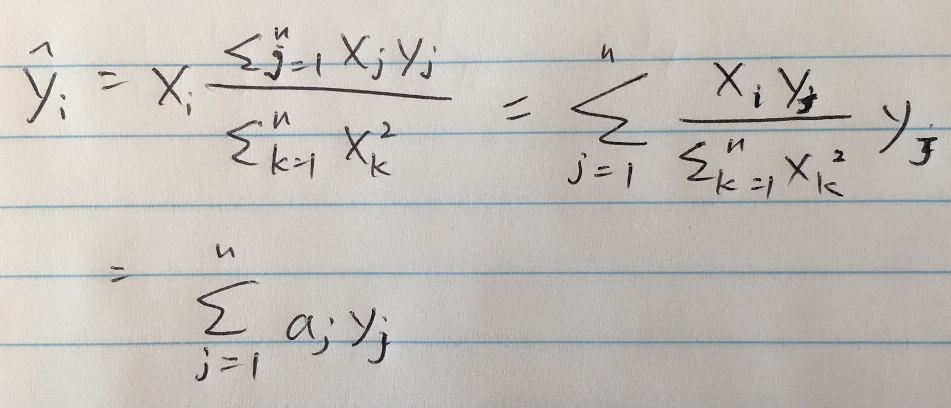
Section 3.7: Problem 5, Problem 7, Problem 9, Problem 11

Section 4.7: Problem 1, Problem 9, Problem 10(a) -- (d), Problem 12

**Section 3.7**

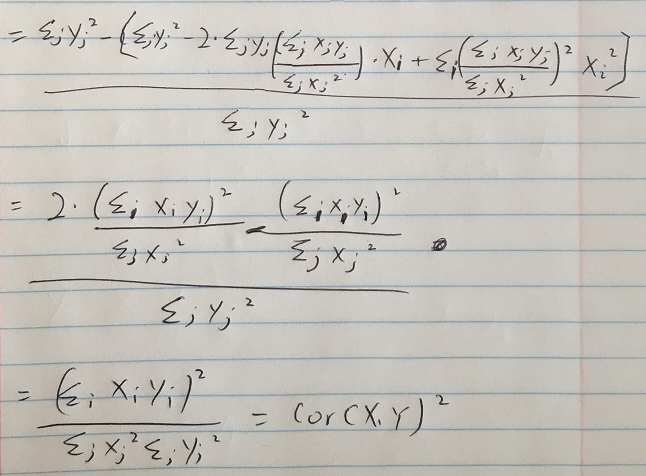
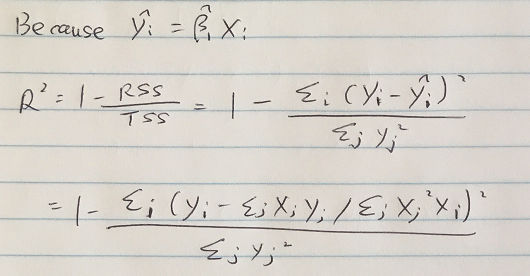
Problem 5

**5.** Consider the fitted values that result from performing linear regression without an intercept.



Problem 7

**7.** It is claimed in the text that in the case of simple linear regression of *Y* onto *X*, the *R*2 statistic (3.17) is equal to the square of the correlation between *X* and *Y* (3.18). Prove that this is the case. For simplicity, you may assume that ¯*x* = ¯*y* = 0.

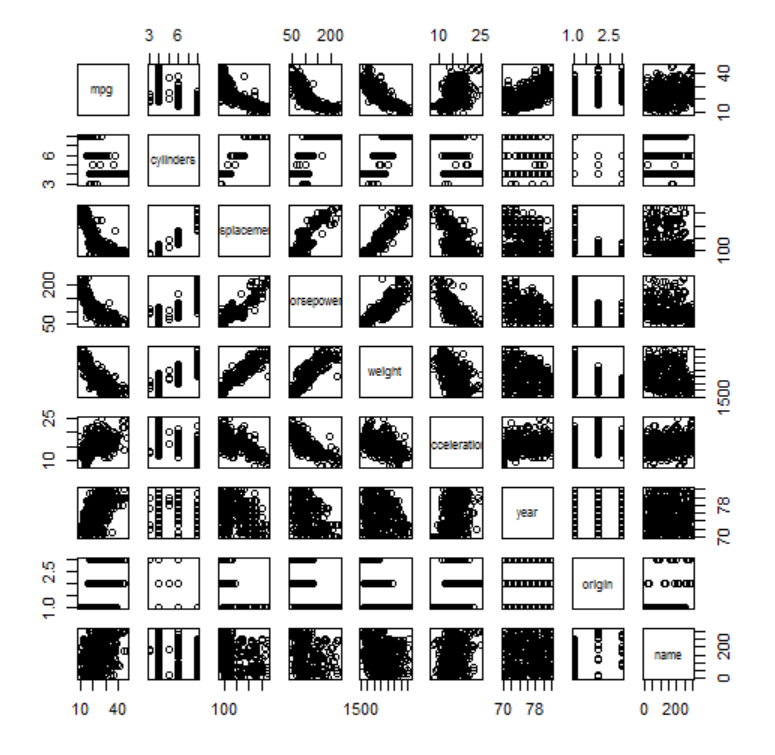


Problem 9

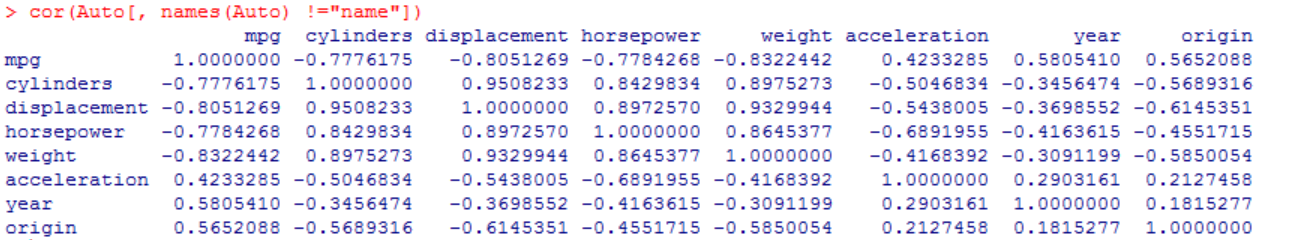
**9.** This question involves the use of multiple linear regression on the Auto data set.

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

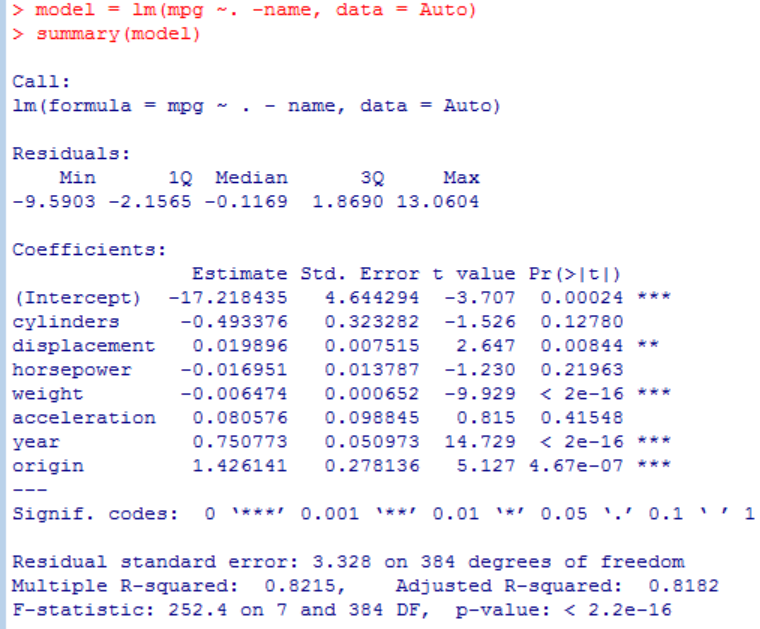




**(b)** Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.



**(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. Comment on the output. For instance:



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

**Yes, there is. R-squared value implies that 82% of the changes in the response can be explained by the predictors in this regression model.**

**Even though, predictors “sylinders”, “horsepower”, and “acceleration” are not statistically significant.**

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

**They are “displacement”, “weight”, “year”, and “origin” predictors.**

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

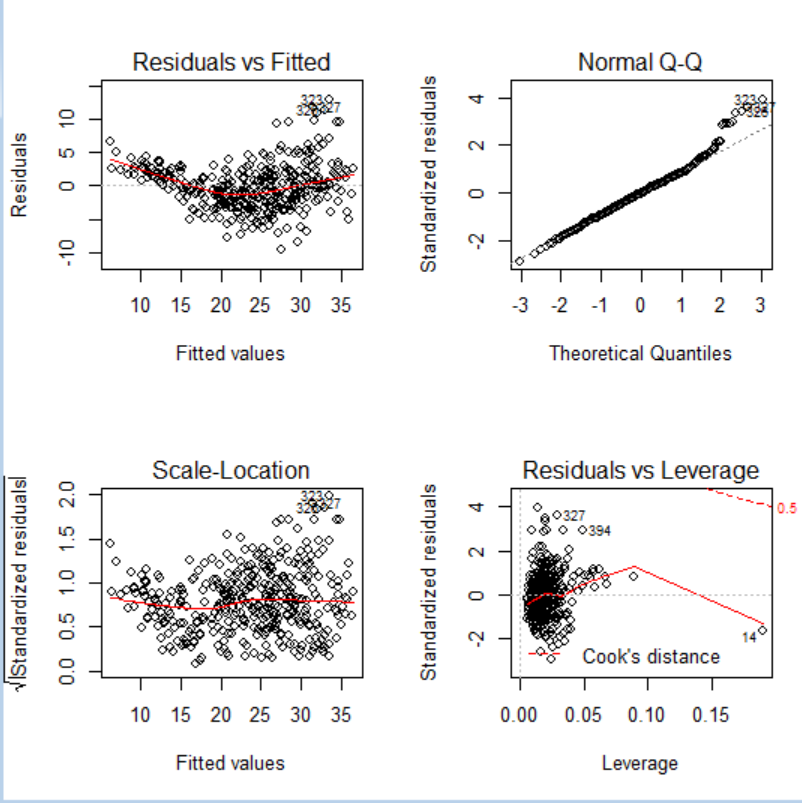
**Hold other factors constant, for every number of year increase, the mpg goes up 0.75 more, and it is valid at 0.001 significant level.**

**(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?

**The plot of residuals versus fitted values indicates the presence of mild non linearity in the data. The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and one high leverage point (point 14).**

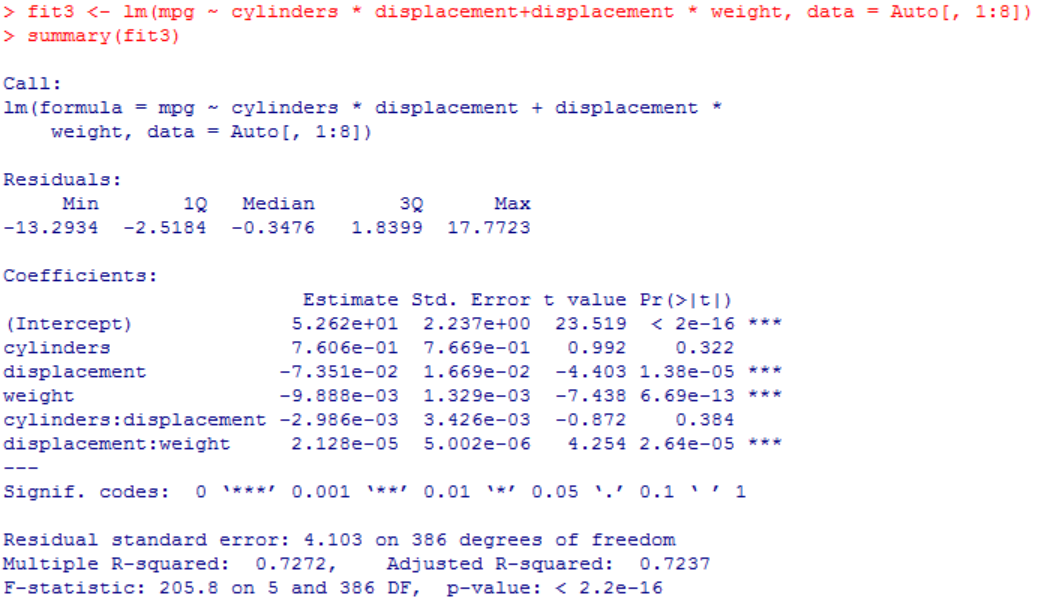




**(e)** Use the \* and: symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

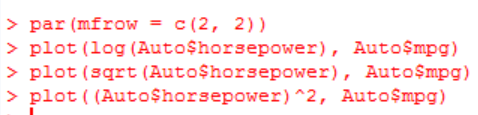
**From the correlation matrix, we obtained the two highest correlated pairs and used them in picking interaction effects.**

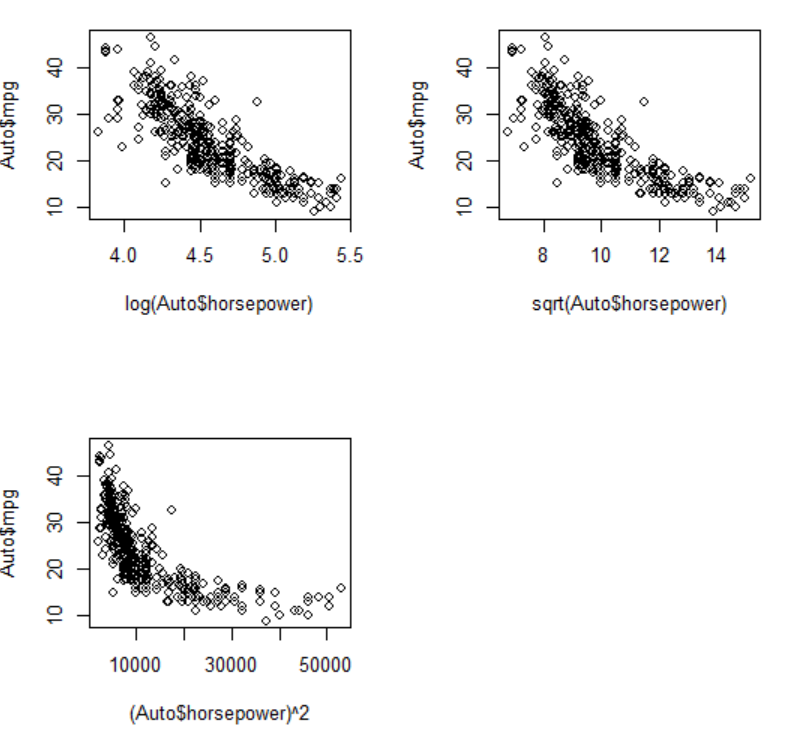
**From the p-values, we can see that the interaction between displacement and weight is statistically signifcant, while the interactiion between cylinders and displacement is not.**



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

**For examping “horsepower”, we found that log transformation gives the most linear from looking at the plot.**





Problem 11

**11.** 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.

**(a)** Perform a simple linear regression of y onto x, without an intercept. Report the coefficient estimate ˆβ, 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

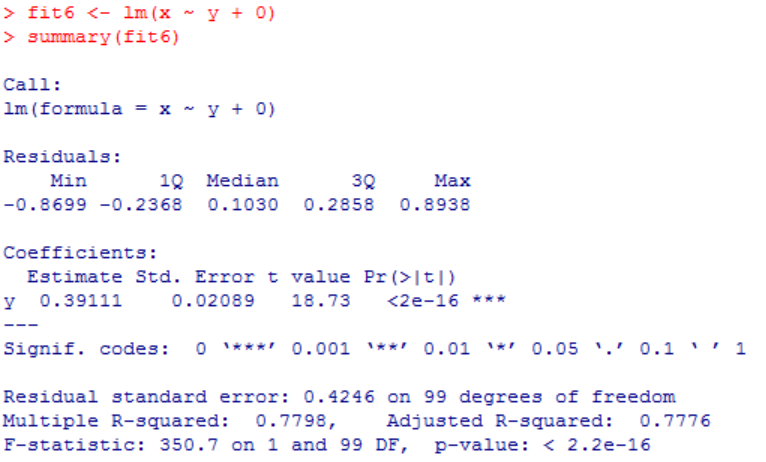
the command lm(y∼x+0).)

**1.9939 for β^, 0.1065 for the standard error, 18.73 for the t-statistic and the p-value < 2e-16. The small p-value allows us to reject H0.**



**(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.

**0.3911 for β^, 0.02089 for the standard error, 18.73 for the t-statistic and p-value** **< 2e-16. The small p-value allows us to reject H0.**



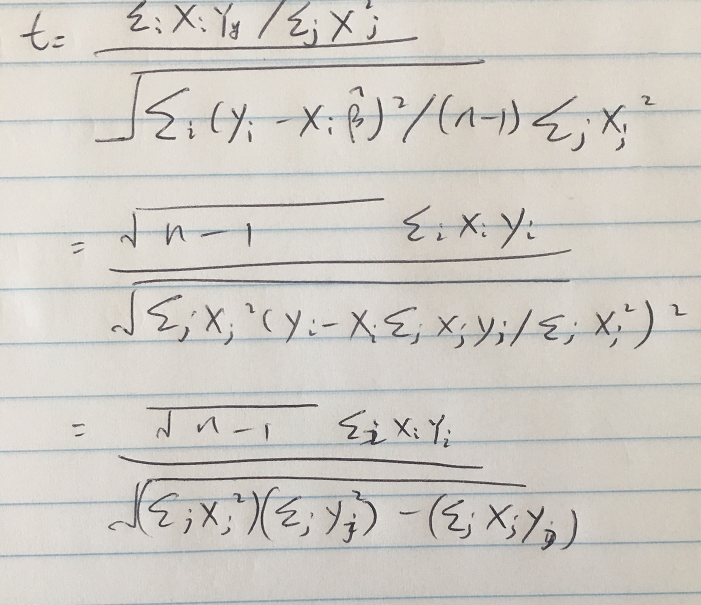
**(c)** What is the relationship between the results obtained in (a) and (b)?

**The t-statistic and p-value are the same. Both results in (a) and (b) reflect the same line created in (a). In other words, y=2x+ε could also be written *x*=0.5(y−ε).**

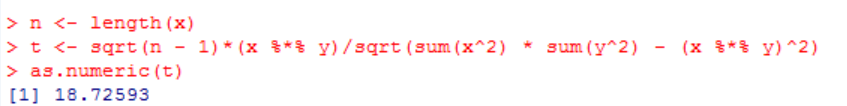
**(d)**

**The result is equal to the value of t from question a & b.**

**Step1:**



**Step2:**

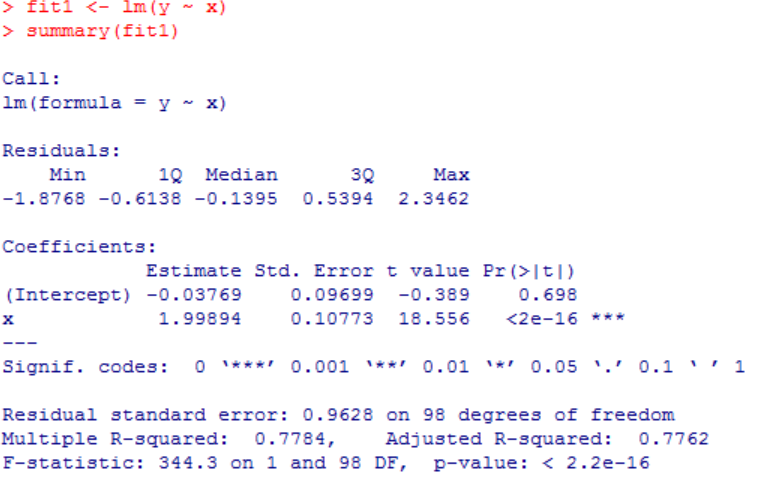


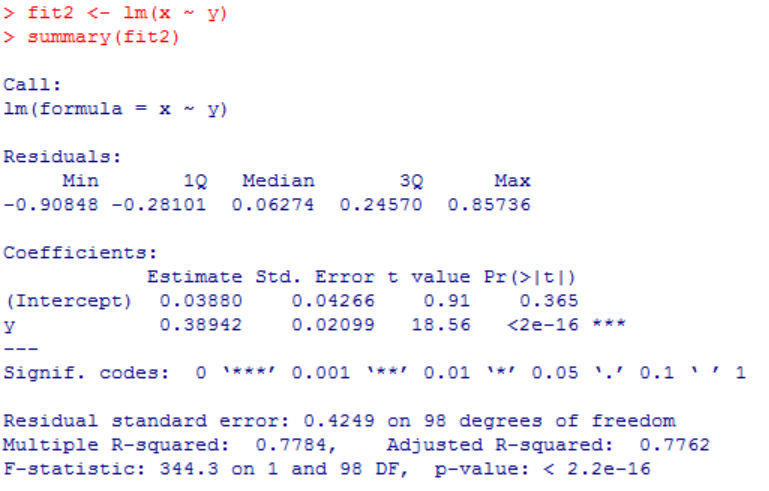
**(e)** Using the results from (d), argue that the t-statistic for the regression of y onto x is the same as the t statistic for the regression of x onto y.

**The result would be the same if we replace xi by yi in the formula for the t-statistic.**

**(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.

**From the result we can see thatare both t of them are equal to 18.56.**

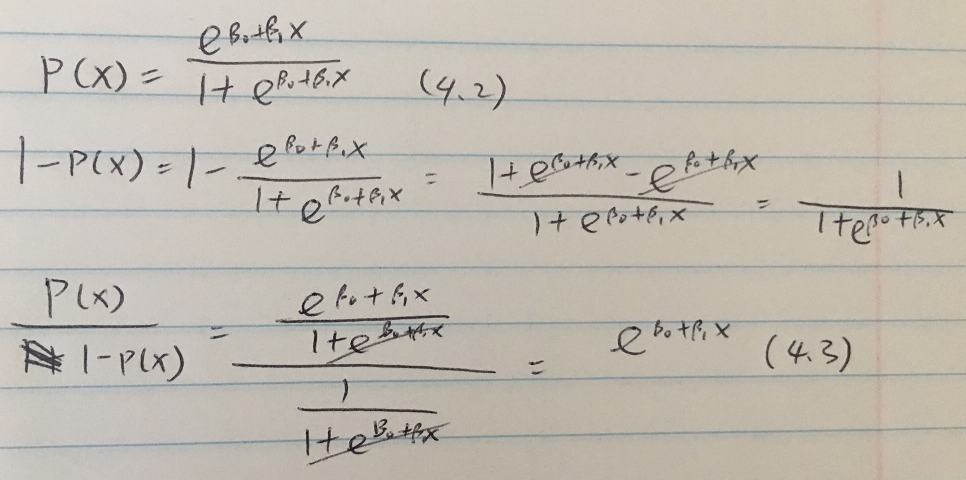




**Section 4.7:**

Problem 1

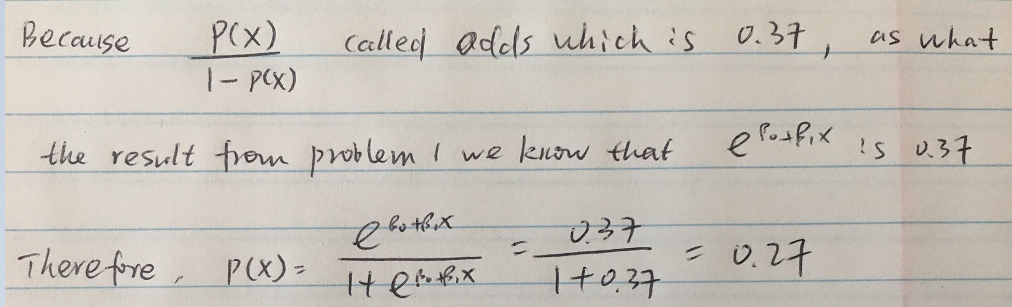
1. Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.



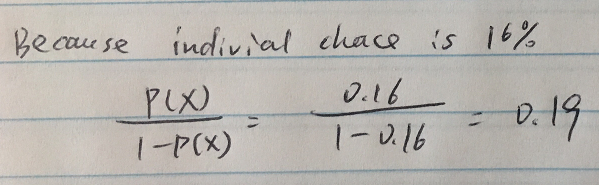
Problem 9

9. This problem has to do with *odds*.

(a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?



(b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

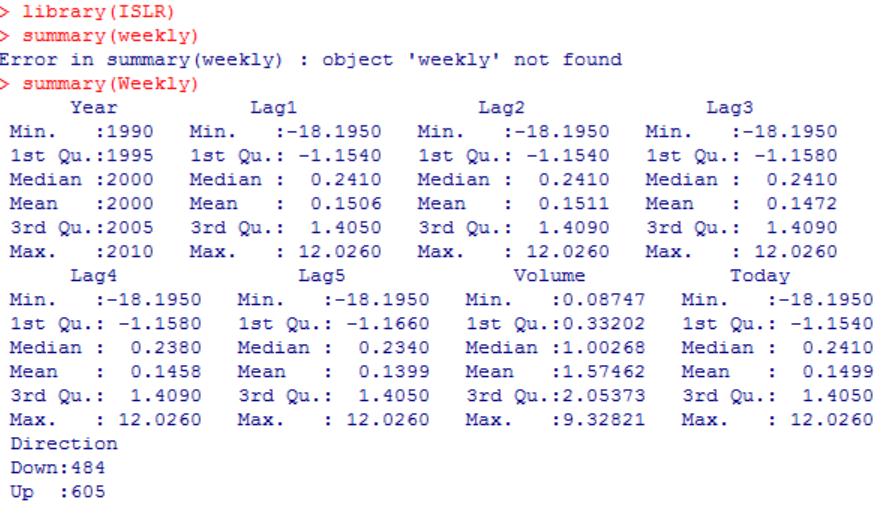


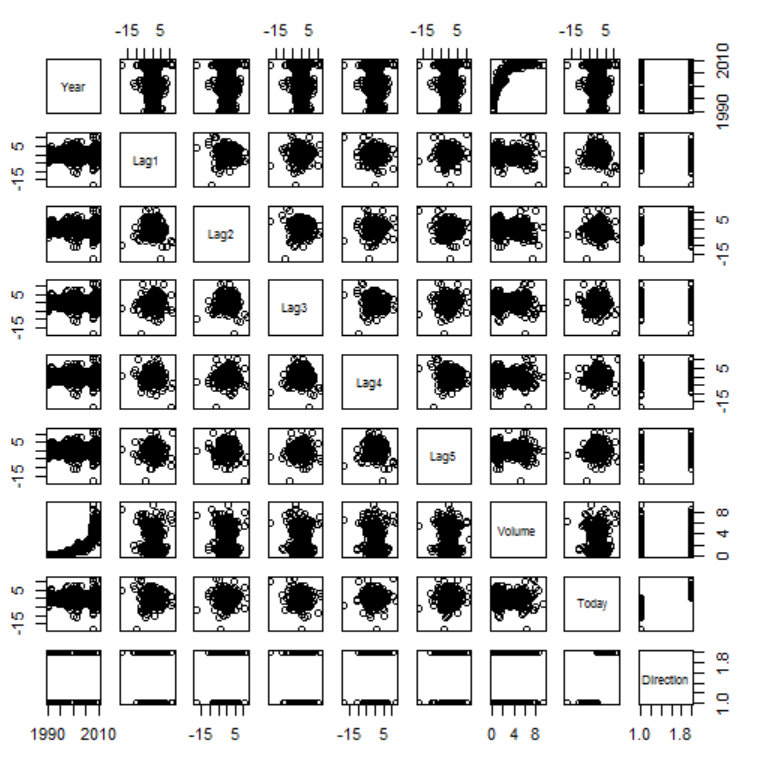
Problem 10(a) -- (d)

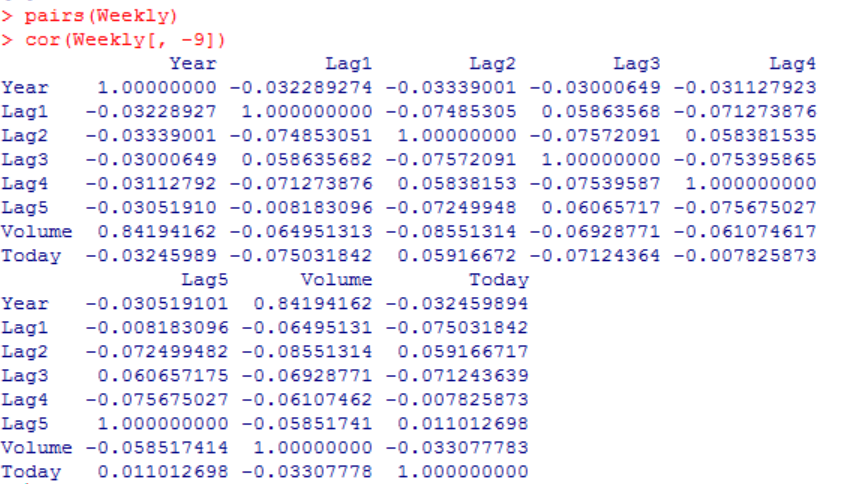
10. This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter’s lab, except that it contains 1, 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

**(a)** Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

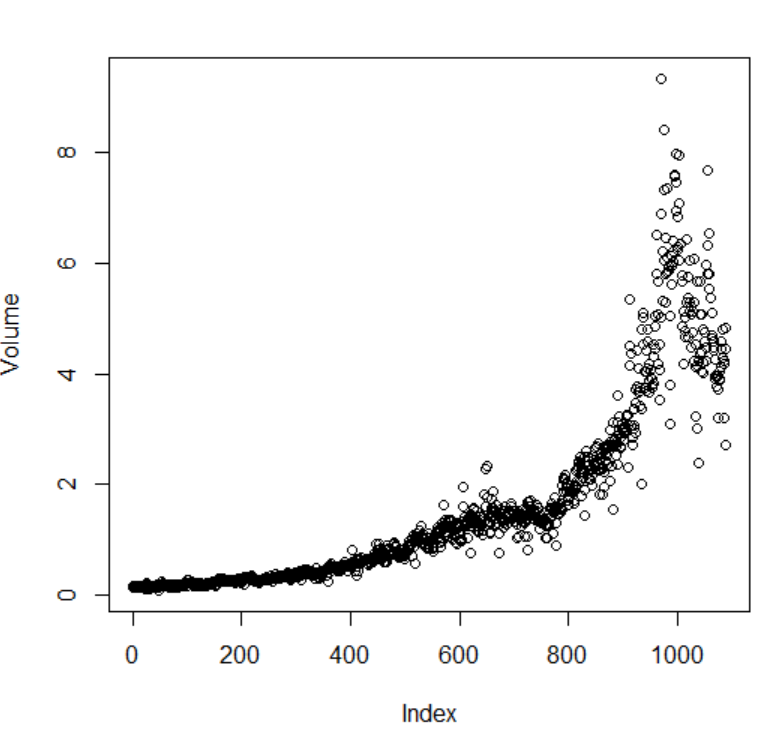
**The correlations between the “lag” variables and today’s returns are close to zero( data2); There is no clear pattern among the “lag” variables data (gragh1), but the volume is increasing over time during year period(gragh2).**

data1

 gragh1

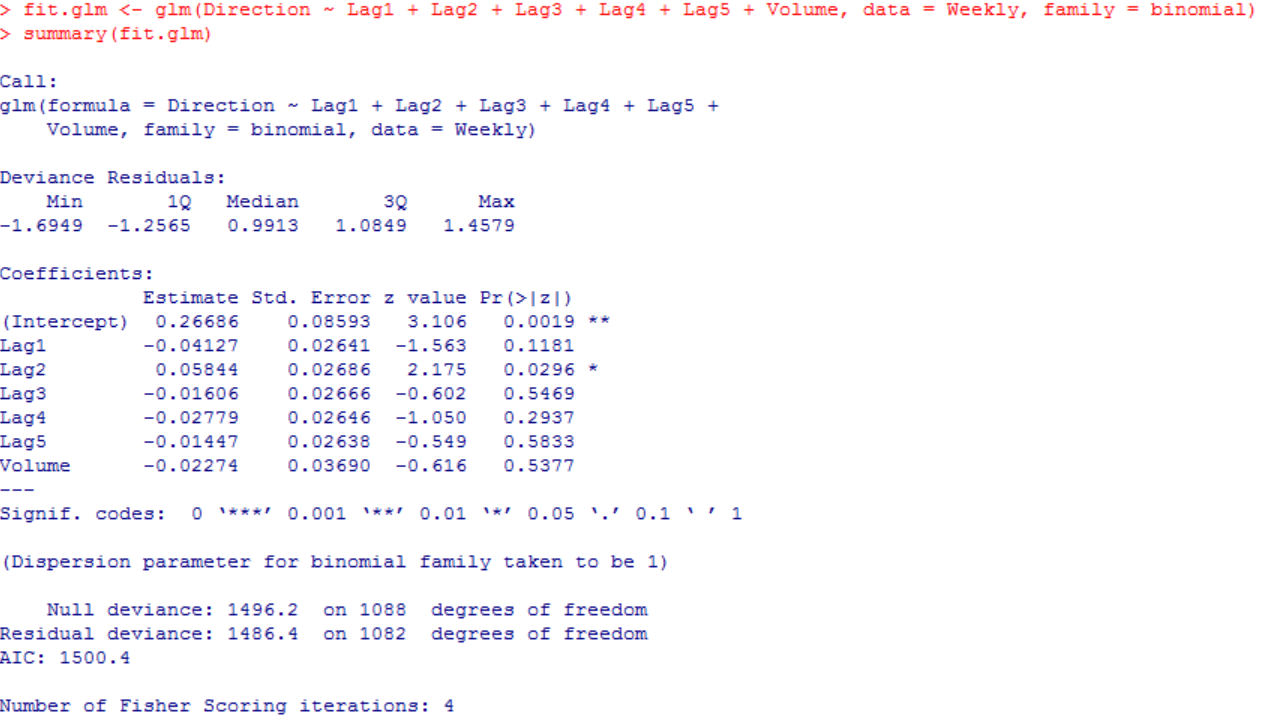
data2



 *gragh2*

**(b)** Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

**Yes, there is. The p-value of predictor Lag2 is 0.0296 which is *statistically* significant in 0.05 level.**



**(c)** Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

**We can know that the percentage of correct predictions on the training data is 56% ((54+557)/1089).**

**The types of mistakes made by logistic regression is being overly optimistic since the conclusion of training error rate is 43.8934803%.**

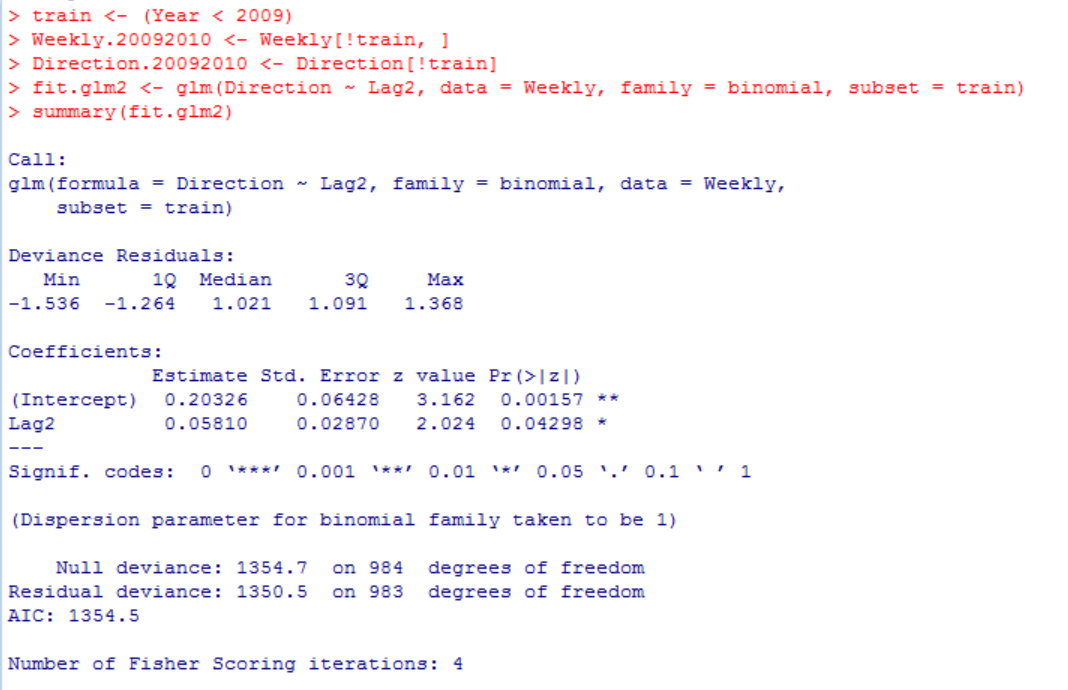


**(d)** Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

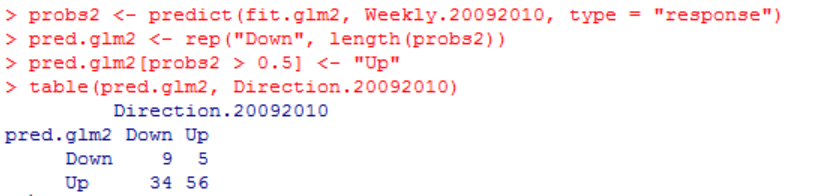
**The percentage of correct predictions on the test data is 62.5% ((9+56)/104), so the test error rate is 37.5%.**

**Also, when the market goes up, the model is right 91.80%; while the market goes down, the model is right only 20.93%.**

**Step1:**



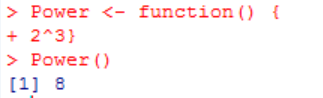
**Step2:**



Problem 12

**12.** This problem involves writing functions.

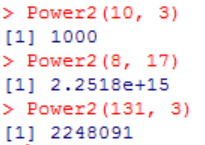
**(a)** Write a function, Power (), that prints out the result of raising 2 to the 3rd power. In other words, your function should compute 23 and print out the results. Hint: Recall that x^a raises x to the power a. Use the print () function to output the result.



**(b)** Create a new function, Power2 (), that allows you to pass any two numbers, x and a, and prints out the value of x^a. You can do this by beginning your function with the line

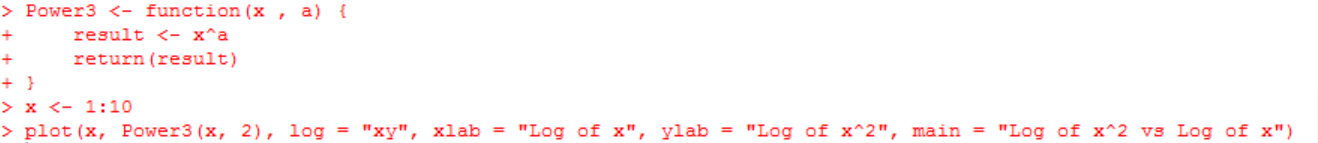


**(c)** Using the Power2 () function that you just wrote, compute 103, 817, and 1313.



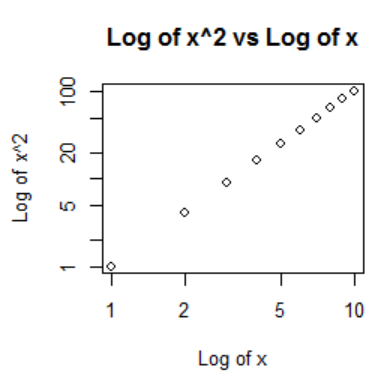
**(d)** Now create a new function, Power3 (), that actually returns the result x^a as an R object, rather than simply printing it to the screen. That is, if you store the value x^a in an object called result within your function, then you can simply return () this result, using the following line:

return (result ) The line above should be the last line in your function, before the } symbol.



**(e)** Now using the Power3 () function, create a plot of f(x) = x2. The x-axis should display a range of integers from 1 to 10, and the y-axis should display x2. Label the axes appropriately, and use an appropriate title for the figure. Consider displaying either the x-axis, the y-axis, or both on the log-scale. You can do this by using log=‘‘x’’, log=‘‘y’’, or log=‘‘xy’’ as arguments to the plot() function.

**Code please see above.**



**(f)** Create a function, PlotPower(), that allows you to create a plot of x against x^a for a fixed a and for a range of values of x. For instance, if you call > PlotPower (1:10 ,3) then a plot should be created with an x-axis taking on values 1, 2, . . . , 10, and a y-axis taking on values 13, 23, . . . , 103.



