**Use R to solve the problems below**

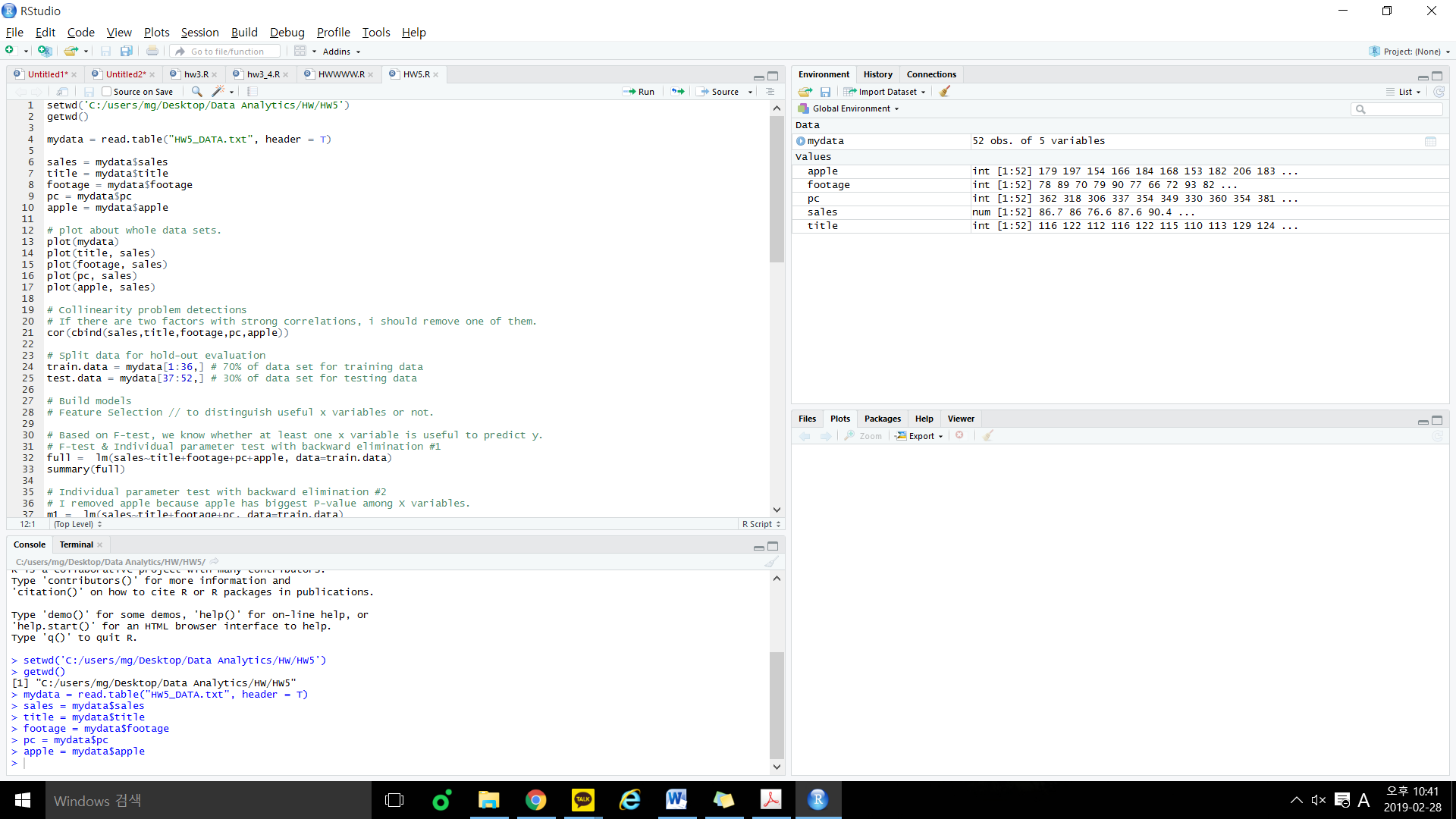
The owner of a rapidly growing computer store tried to explain the increase in biweekly sales of  
computer software, using four explanatory variables: Number of titles displayed (title). Display footage (footage), current customer base of Personal Computers (PC) and current customer base of Apple compatible computers (apple). The data are stored in the data file HW4\_DATA.txt attached to this assignment with SALES in column 1, TITLE in column 2, FOOTAGE in column 3, PC in column 4 and APPLE in column 5.

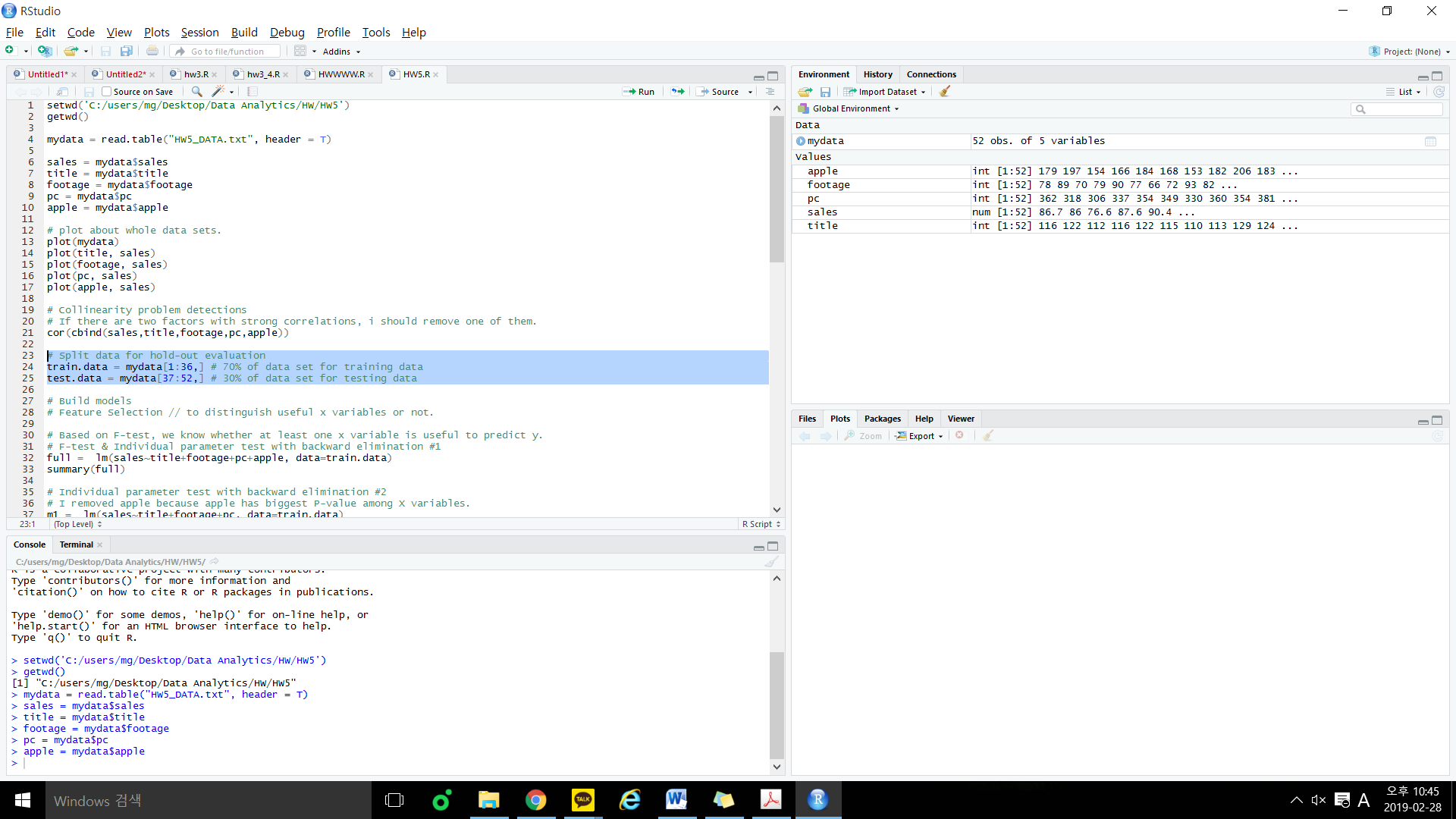
Note:

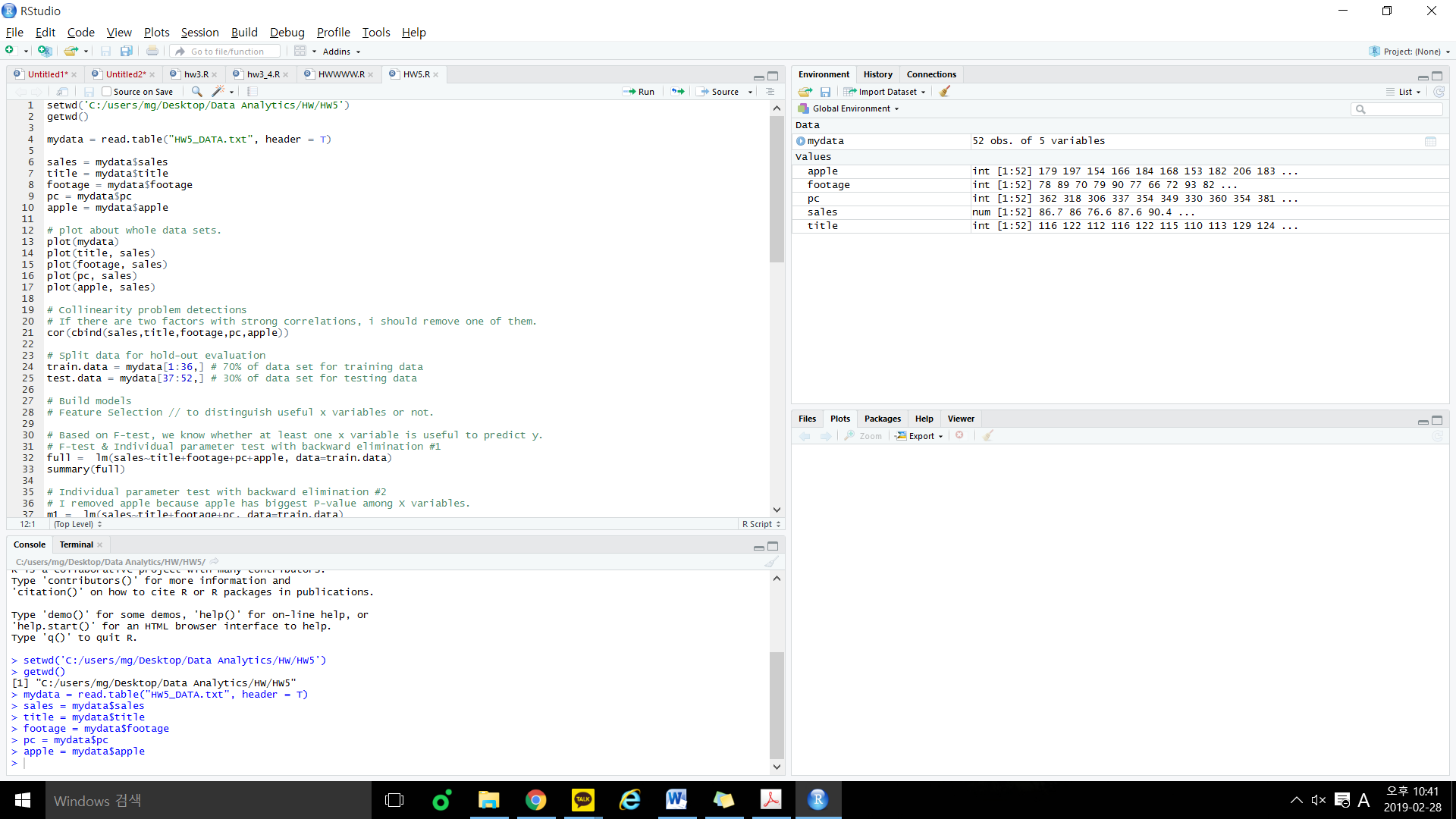
* Every step you use R, you should provide the snapshots of your R commands and R outputs, and paste the plots if it is necessary.
* Use 95% as confidence level for the following questions
* Do NOT shuffle the data, just use the first 0.7 rows as training, and the remaining as testing. In this case, it is easier for TA to grade

Compute the appropriate regression analysis using R and answer the following questions:

1. Import the data in R and define the variables as: sales, title, footage, pc and apple.





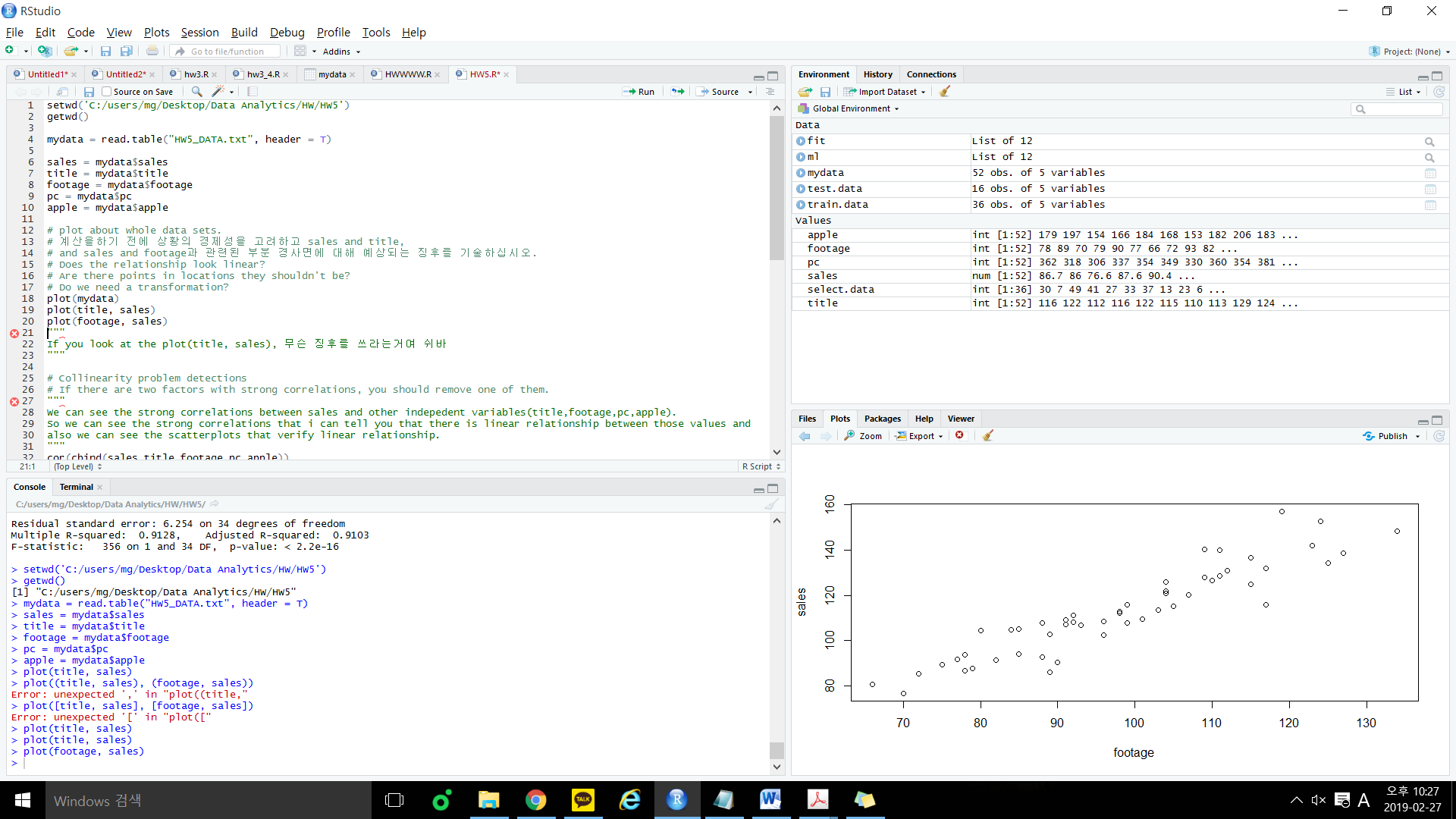
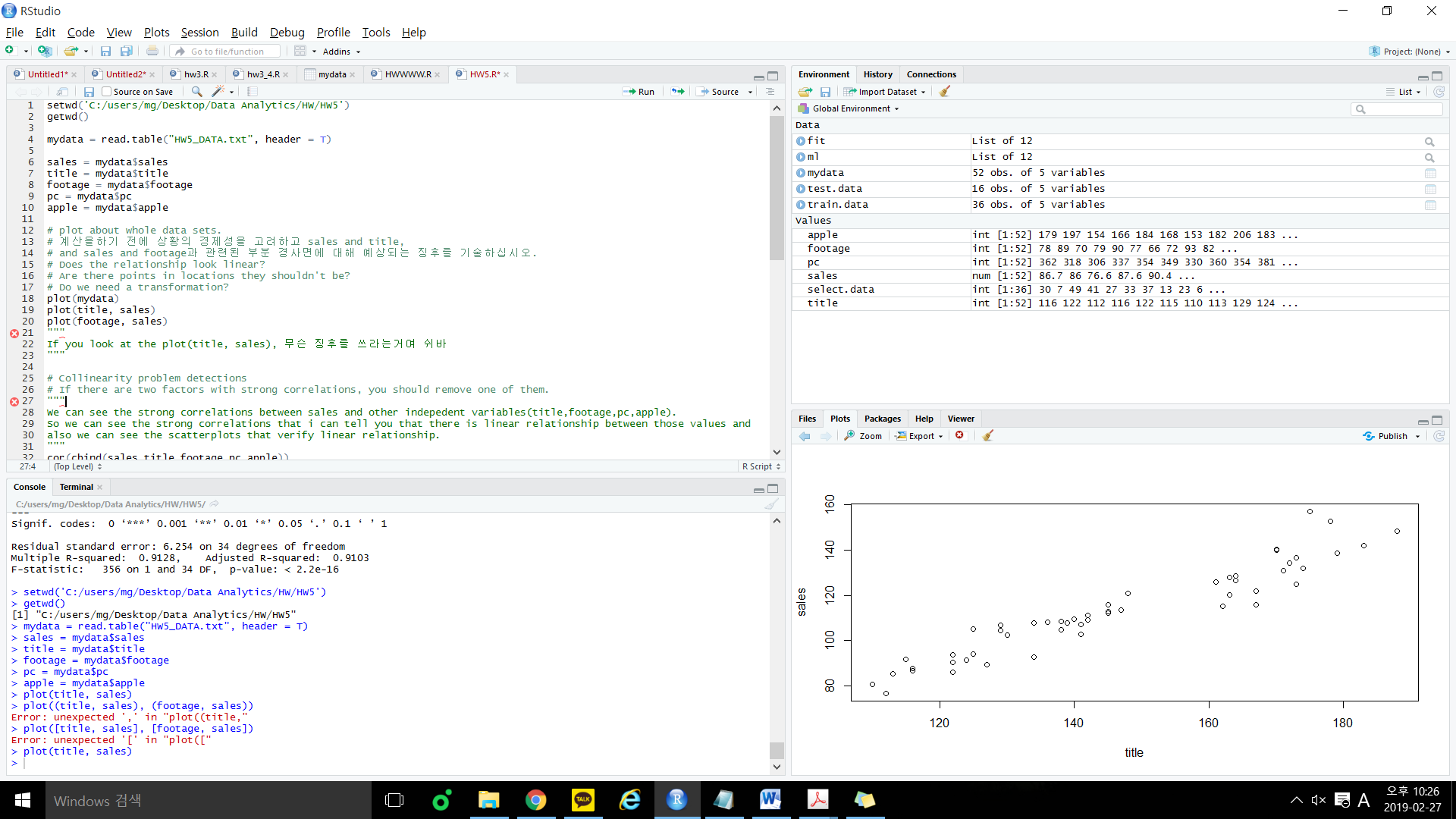


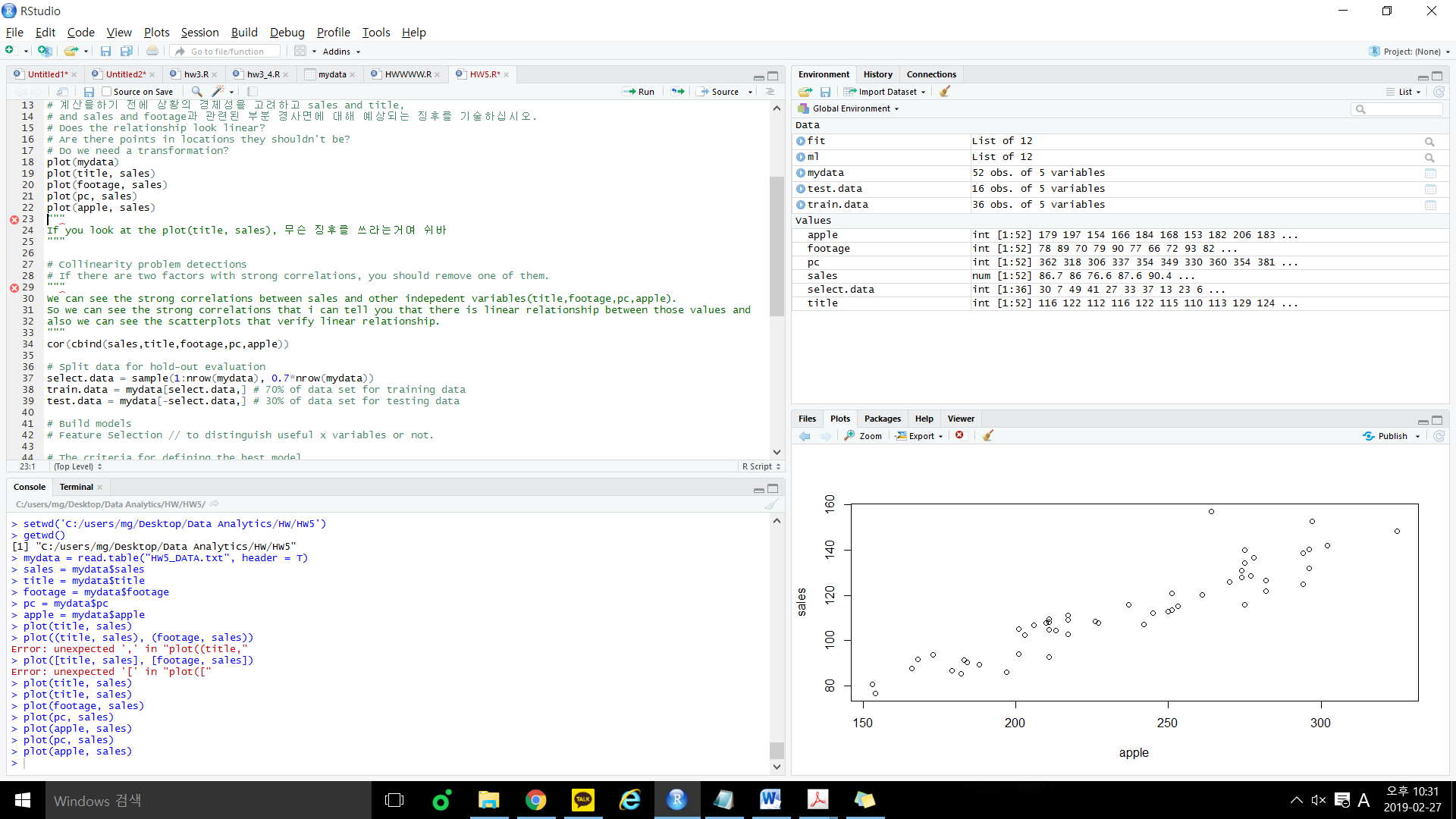
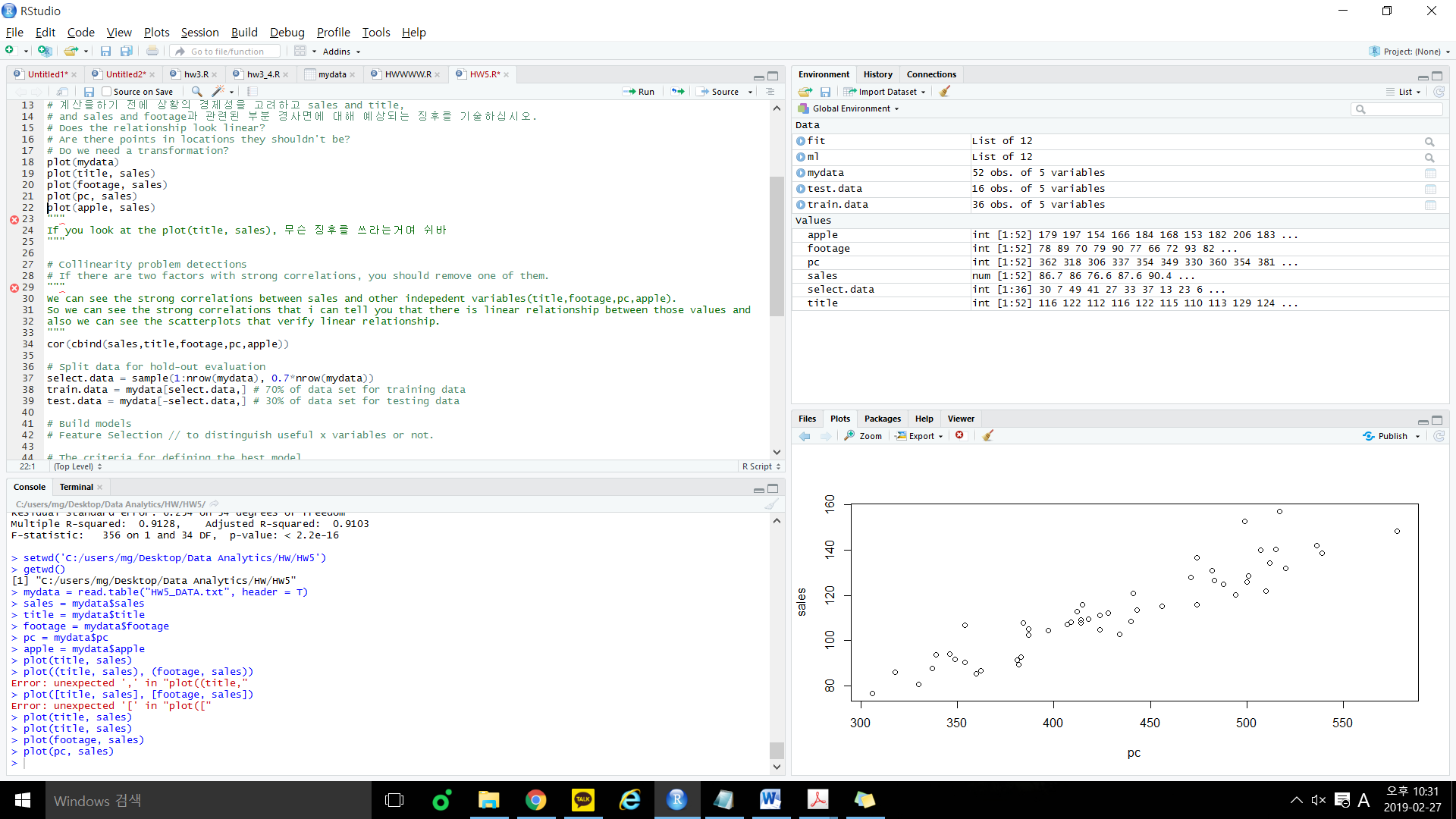
2. [10] Before doing any calculation, consider the economics of the situation and state what sign you would expect for the partial slopes relating sales and title, and sales and footage.

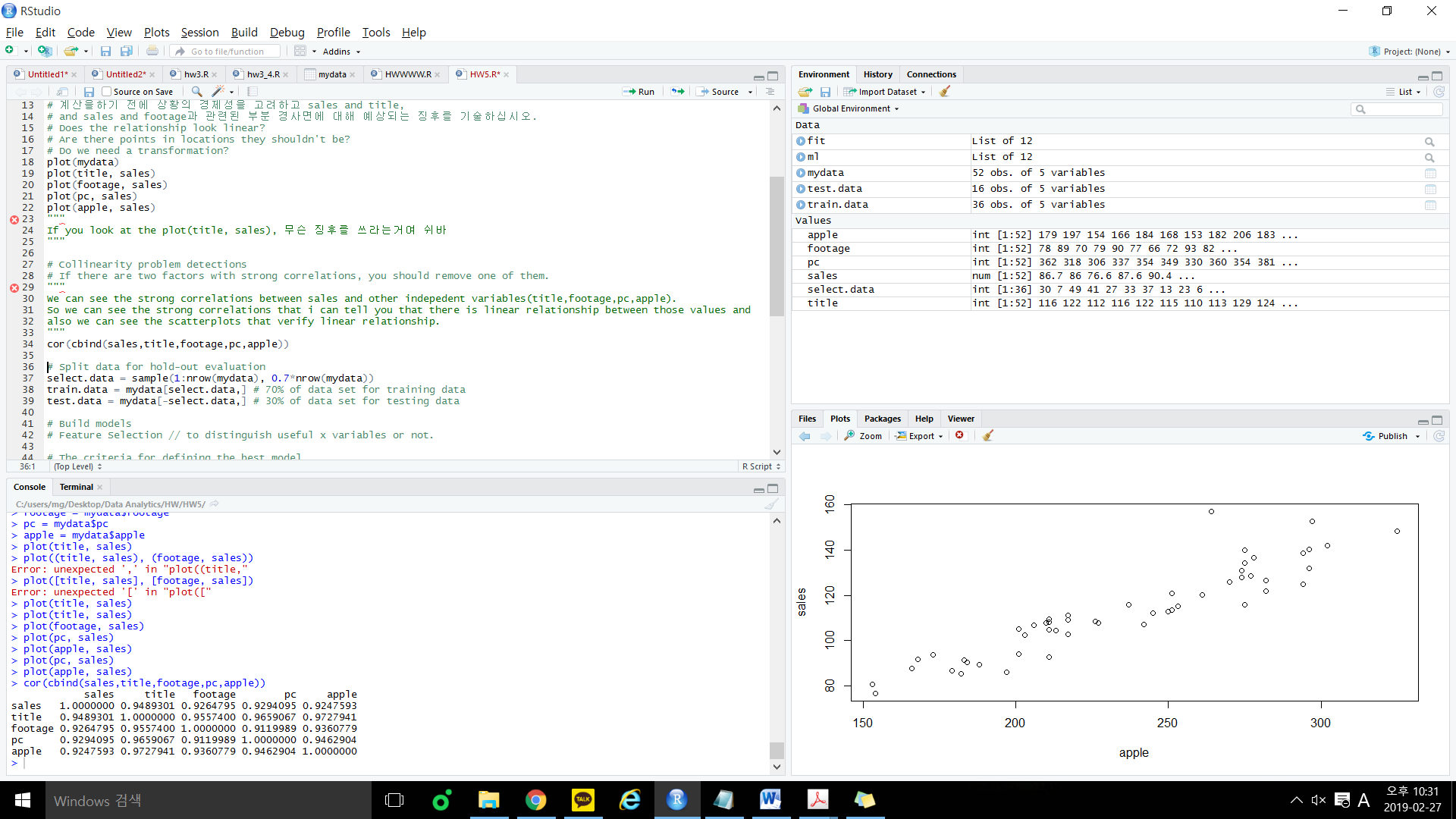
**The increase in biweekly sales of the computer software must be related to some of variables.**

**If it is related to title or footage, slope between sales and each variables relationship will be linearly proportional.**

3. [10] Use scatterplots AND correlations to analyze if there is a linear association among sales and the four predictors. If there is no or weak relationship, just try log and sqrt transformations to see whether the situation can be improved.







In scatterplot, we can see that there is strong linear relationship between sales and other independent variables. Also in correlation result, we can see the strong correlation between and Y.

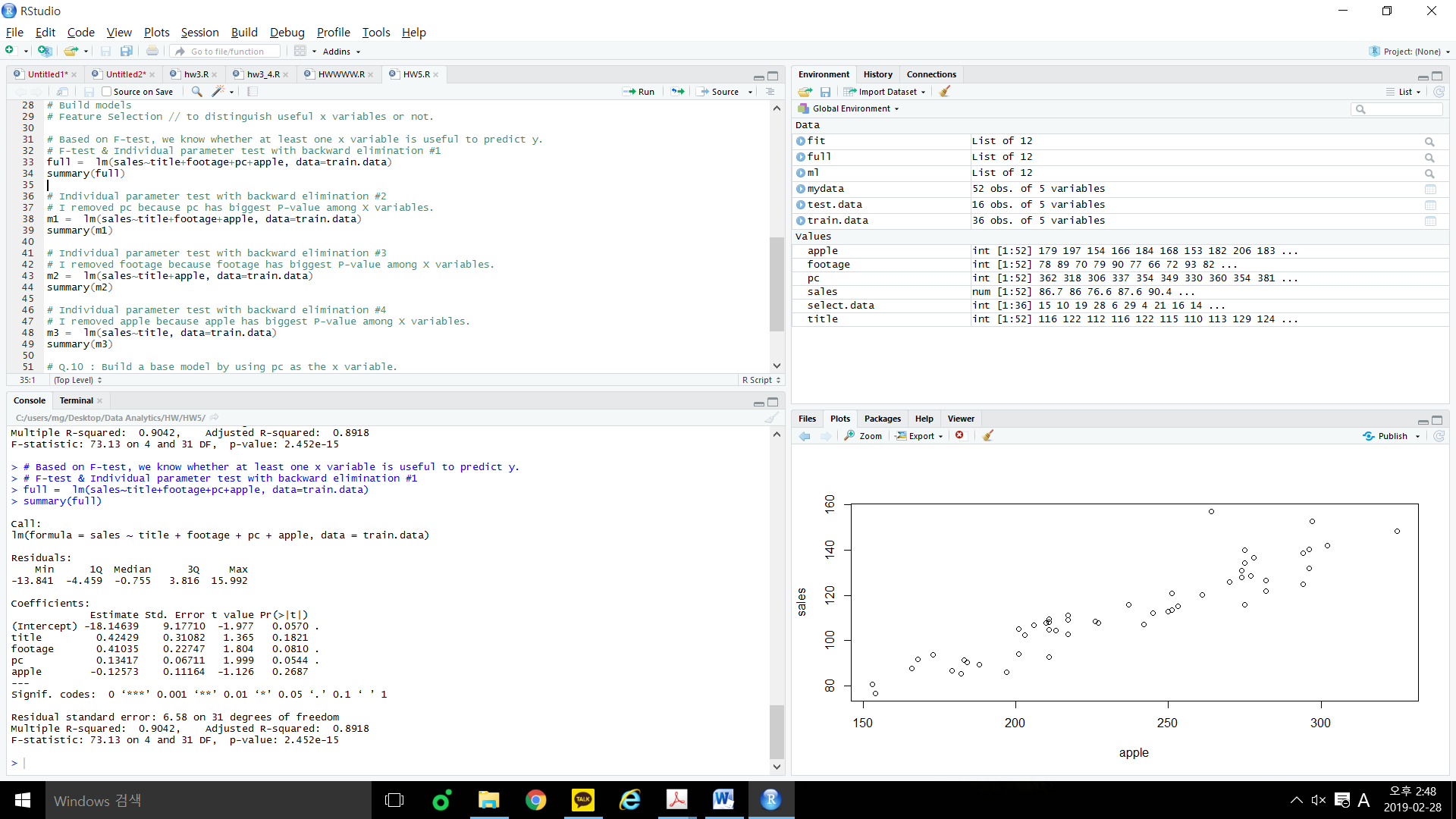
Therefore, we don’t have to use transformations.

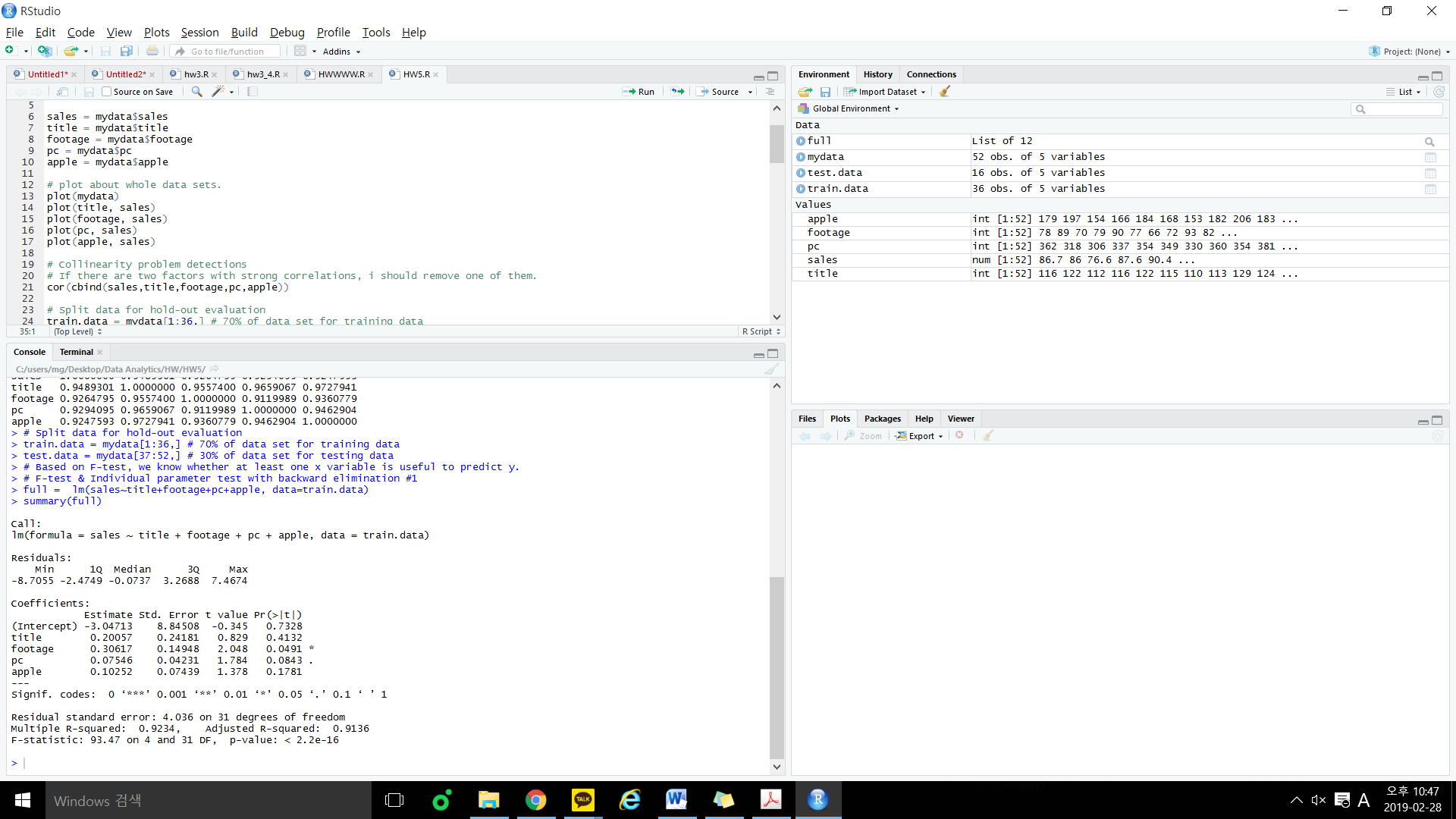
4. [10] Compute in R a multiple regression equation with sales as the dependent variable Y and  
title, footage, pc and apple as the predictors. Write down the expression for the fitted  
regression model. [Run it once, use backward elimination but you do not need to remove X variables in this step]. Interpret the outputs in the F-test and individual parameter tests.

**Fitted regression model)**

=

=





F-test)

If you look at the P-value in this result above,

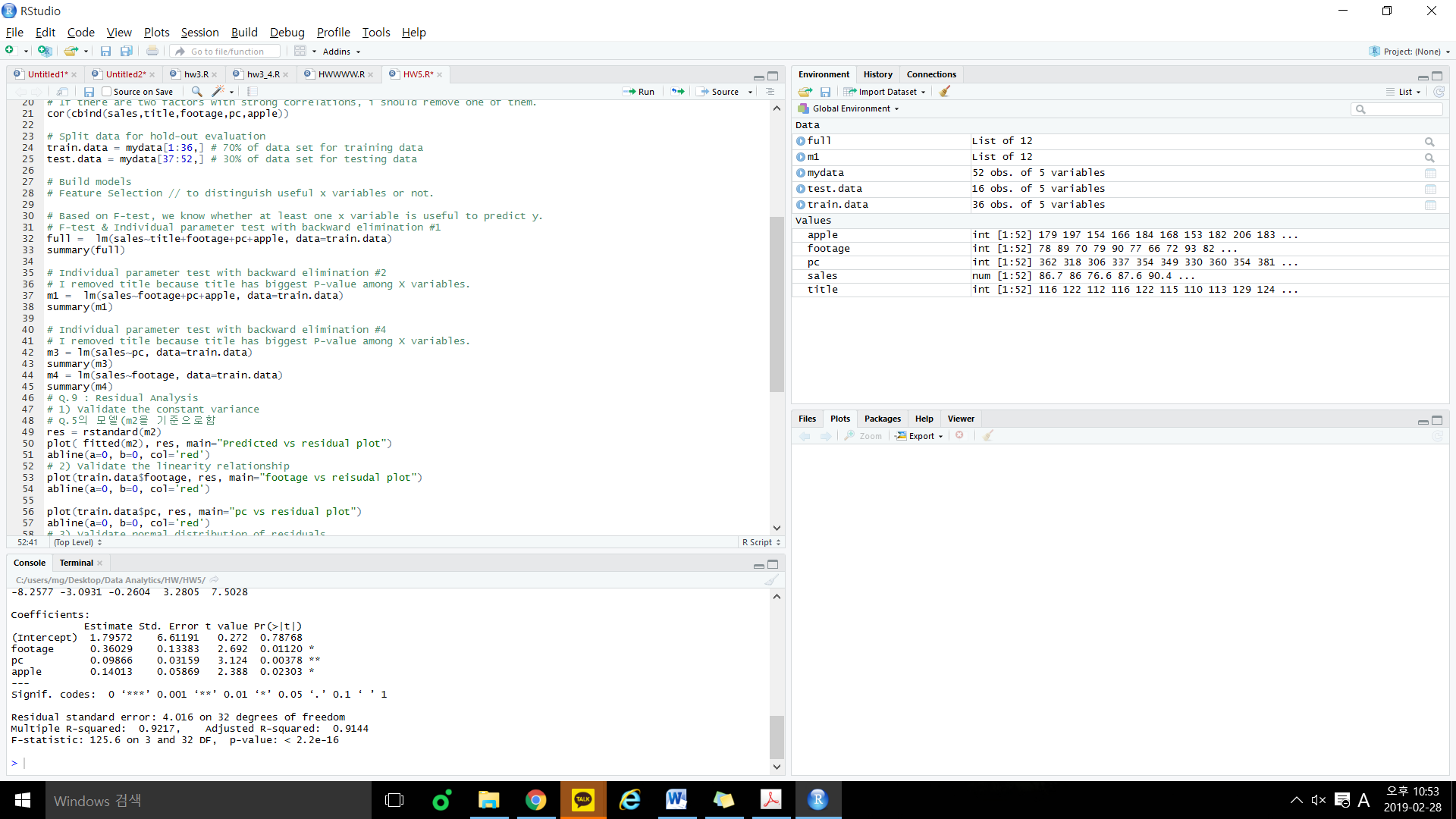
P-value: < 2.2e-16 which means P-value is smaller than .

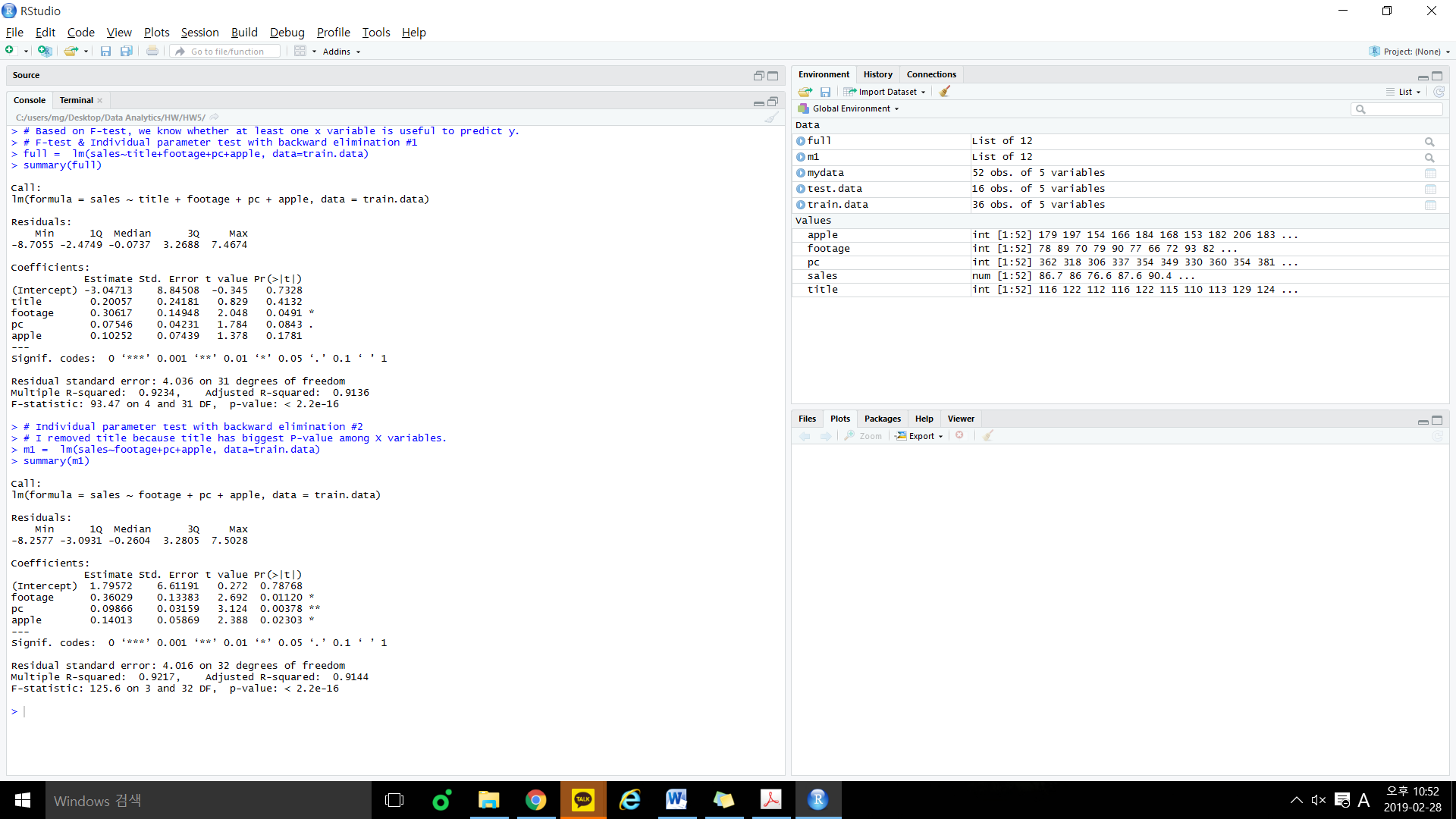
So the conclusion is at 95% confidence level, at least one X variable has significant linear relationship with Y, and it can affect the value of the Y.

Individual parameter test)

Title has largest p-value among x variables, so we should remove this first.

5. [10] By using p-value as the metric and perform backward elimination to get the model, Write down the expression for the final reduced regression model. Based on the test results, what variables have now a significant effect on Y? Does each estimated coefficient have the sign you expected in part 2)?





**Final Reduced Regression Model)**

=

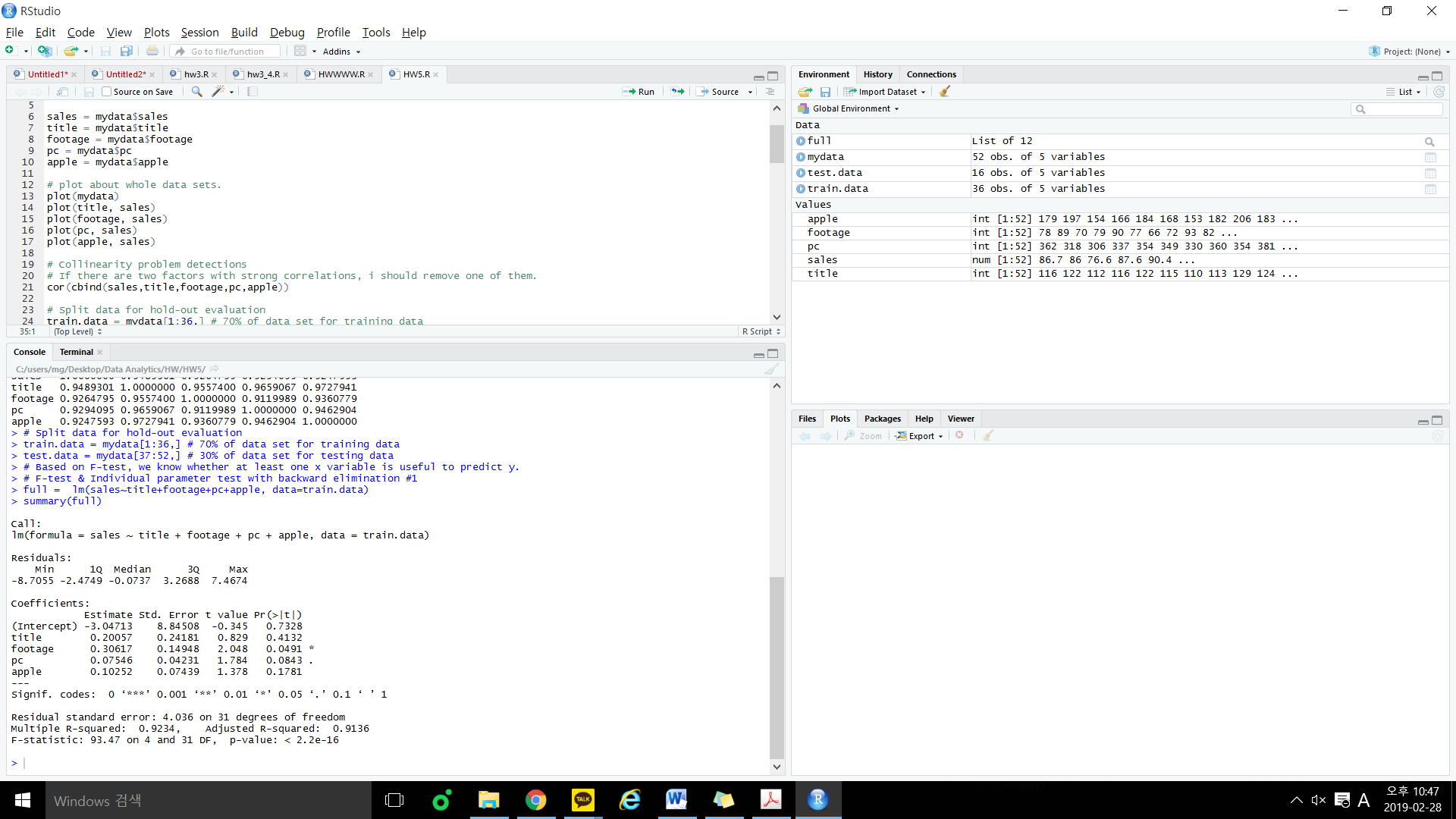
First, I removed title which have biggest P-value in each step and which is greater than 0.05 one by one in R.

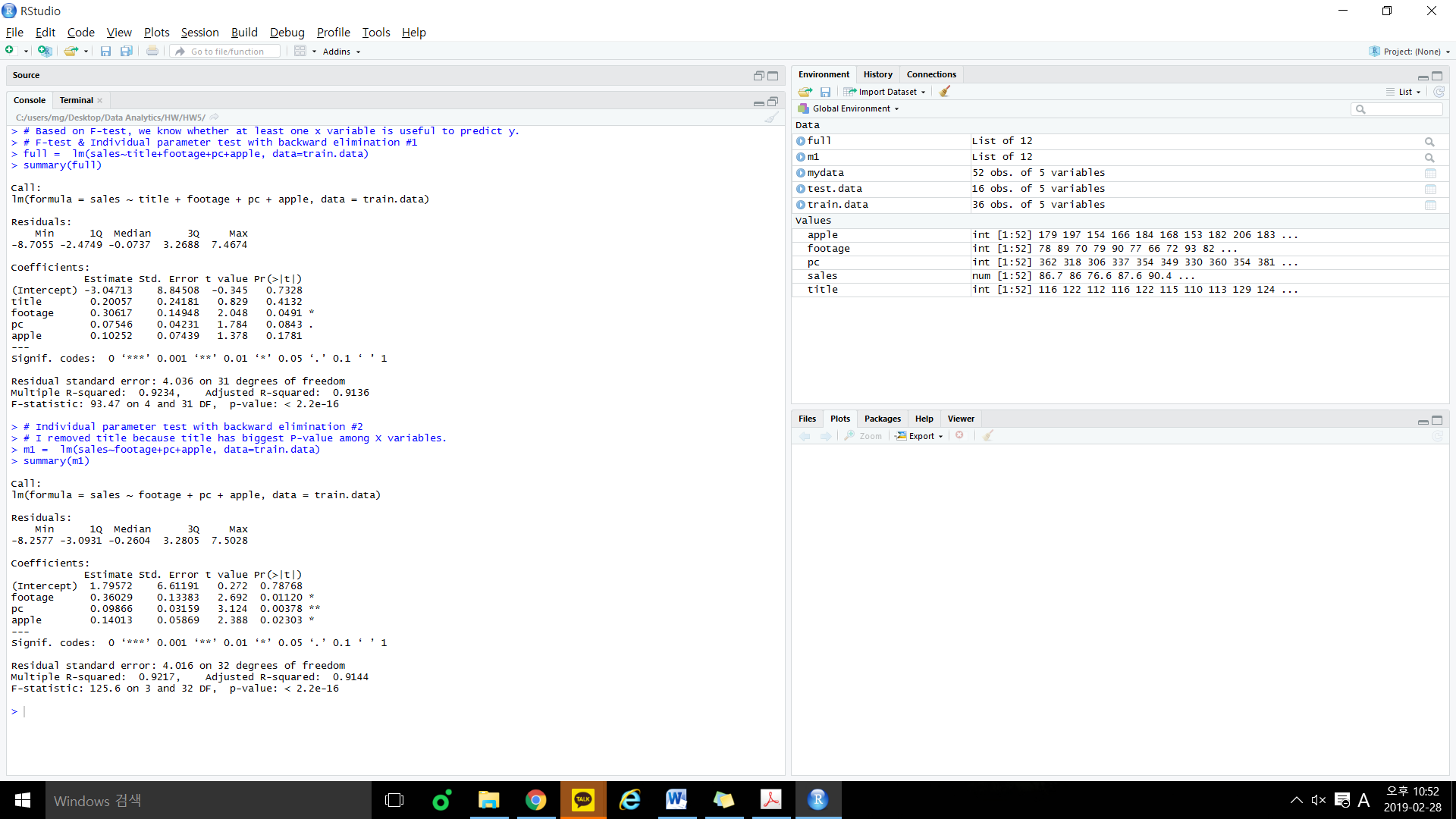
Second, footage, pc, apple have a significant effect on Y.

Third, title is not useful so it is remove but footage is useful and it’s sign is positive that I mentioned in 2)

.

6.[10] Compare the values for the coefficients of determination (adj-R2) for the full model fitted in 4) and the reduced model fitted in 5). Interpret the adj-R2 in part 5), and Discuss what they indicate in terms of model fit, and give me your conclusions

 = Model in (4)

 = Model in (5)

Adj- is useful when comparing two models with a different set of x-variables.

A higher Adj- typically indicates a better model, in terms of the training data set.

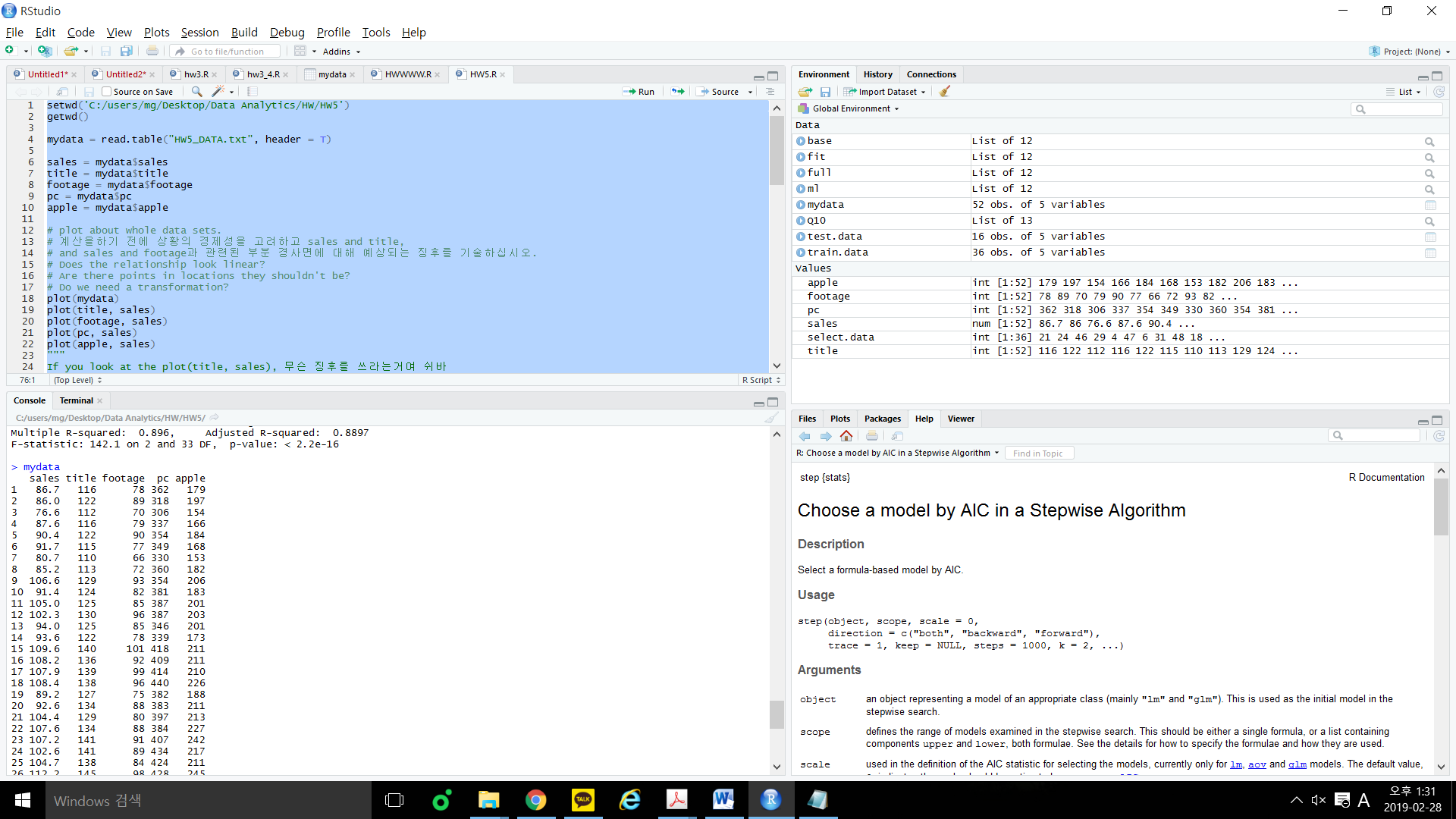
In this case, reduced model is better model comparing to full model.

7. [10] Use the reduced model in 5) to predict the sales value for 130 titles displayed in 80 foot display area with 400 customers using PC’s and 215 customers using Apple computers. Find an  
observation in the dataset that is similar to this case and compute the prediction error. Note, if there are multiple observations which meet the requirements, choose the one with larger row index.

=

Title = 130, footage = 80, pc = 400, apple = 215. So if I apply these values on the model above,

= = 100.7 +



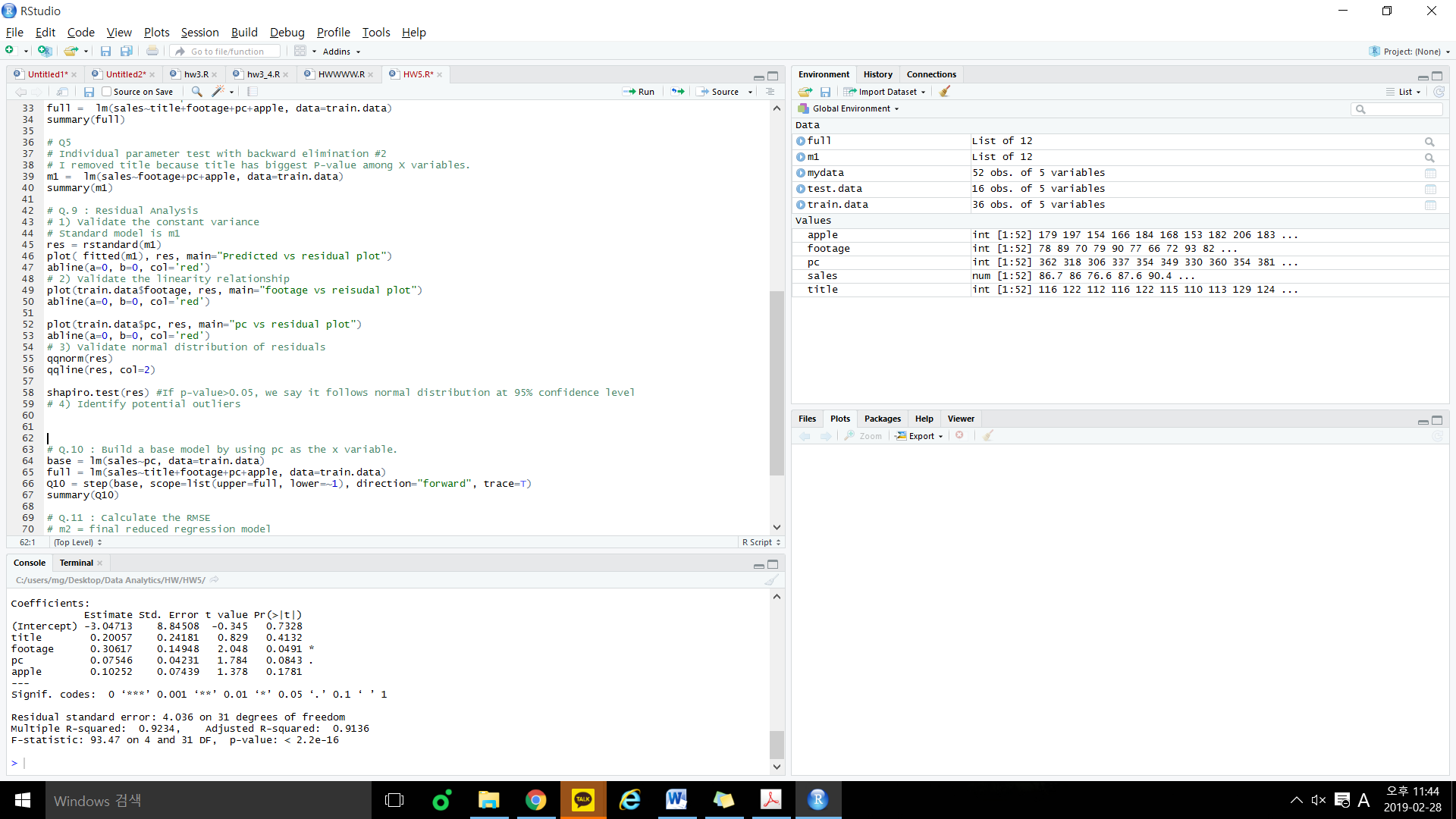
Index[21]: title = 129, footage = 80, pc = 397, apple=213 and sales = 104.4

So prediction error = observed value – predicted value = 104.4 – 100.7 = 3.7

8.[10] Explain the slopes in your final model in 7)

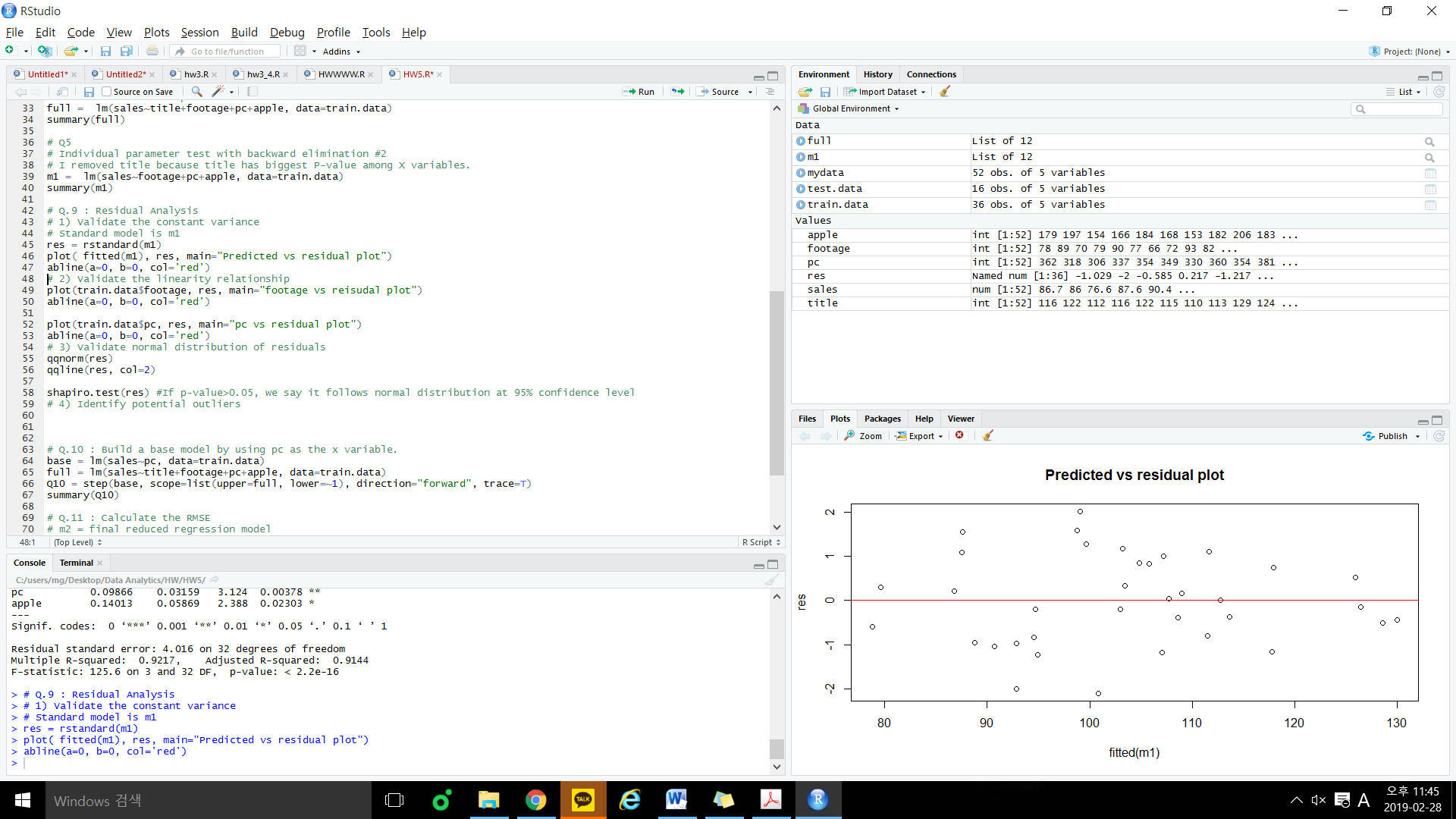
Final model have partial slopes. Slope of footage is 0.36, slope of pc is 0.10 and slope of apple is 0.14

9. [10] Perform residual analysis, and provide your conclusions after the residual analysis. Note: if you are going to use the normality test, let’s use the Shapiro-Wilk Normality Test



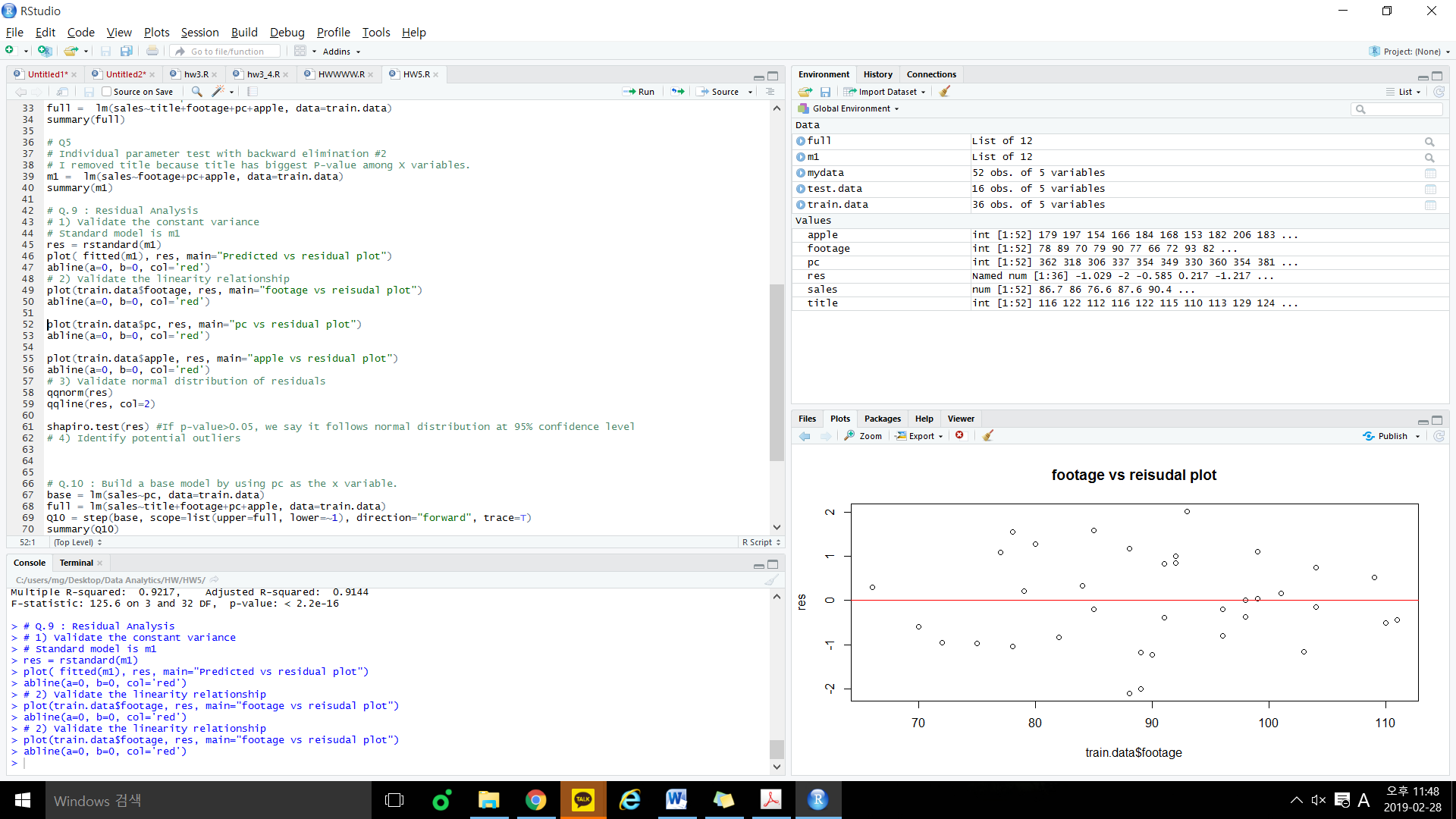
Goals in Residual Analysis

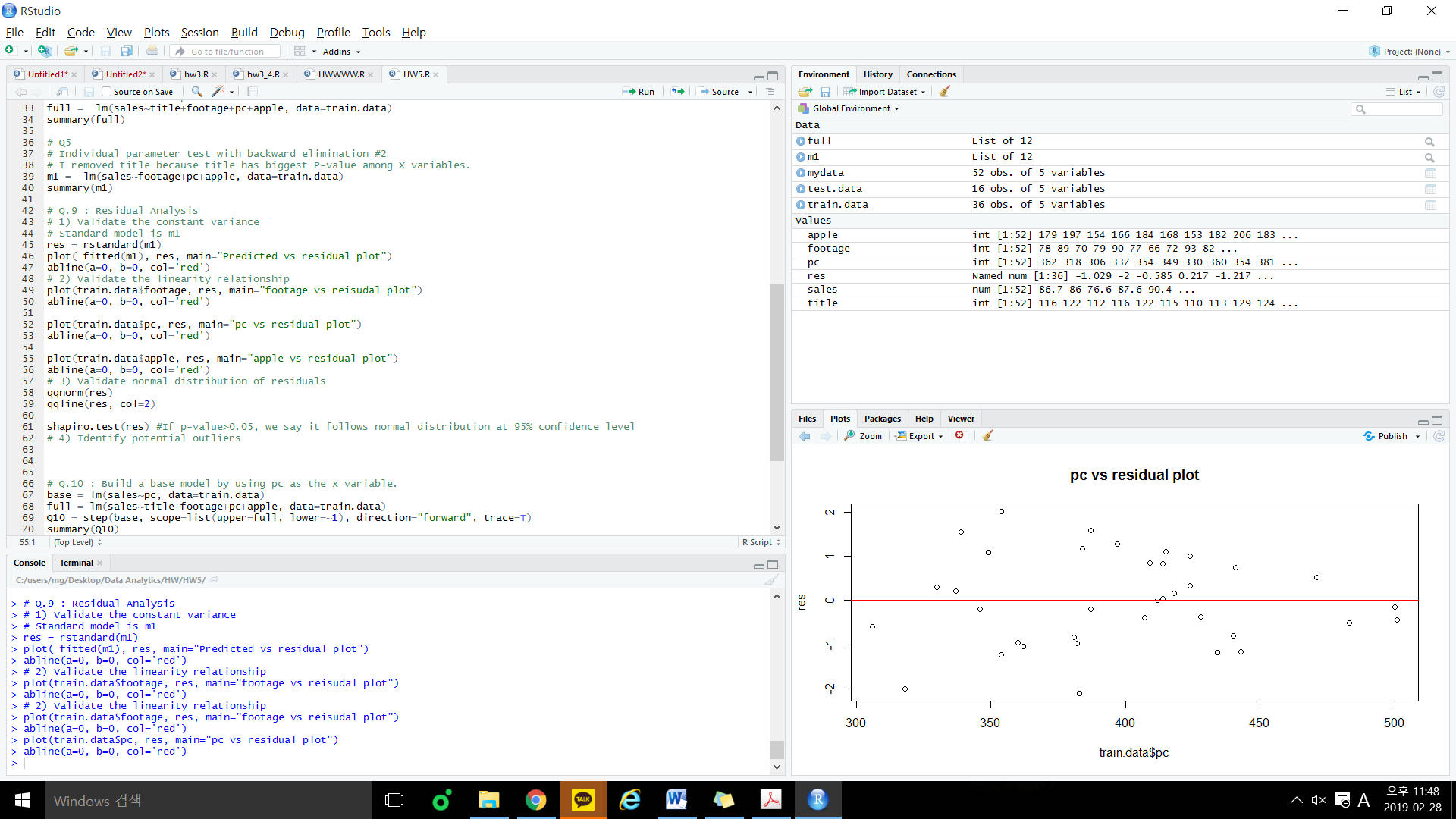
1. Validate the constant variance
   1. Plot residuals vs predicted values

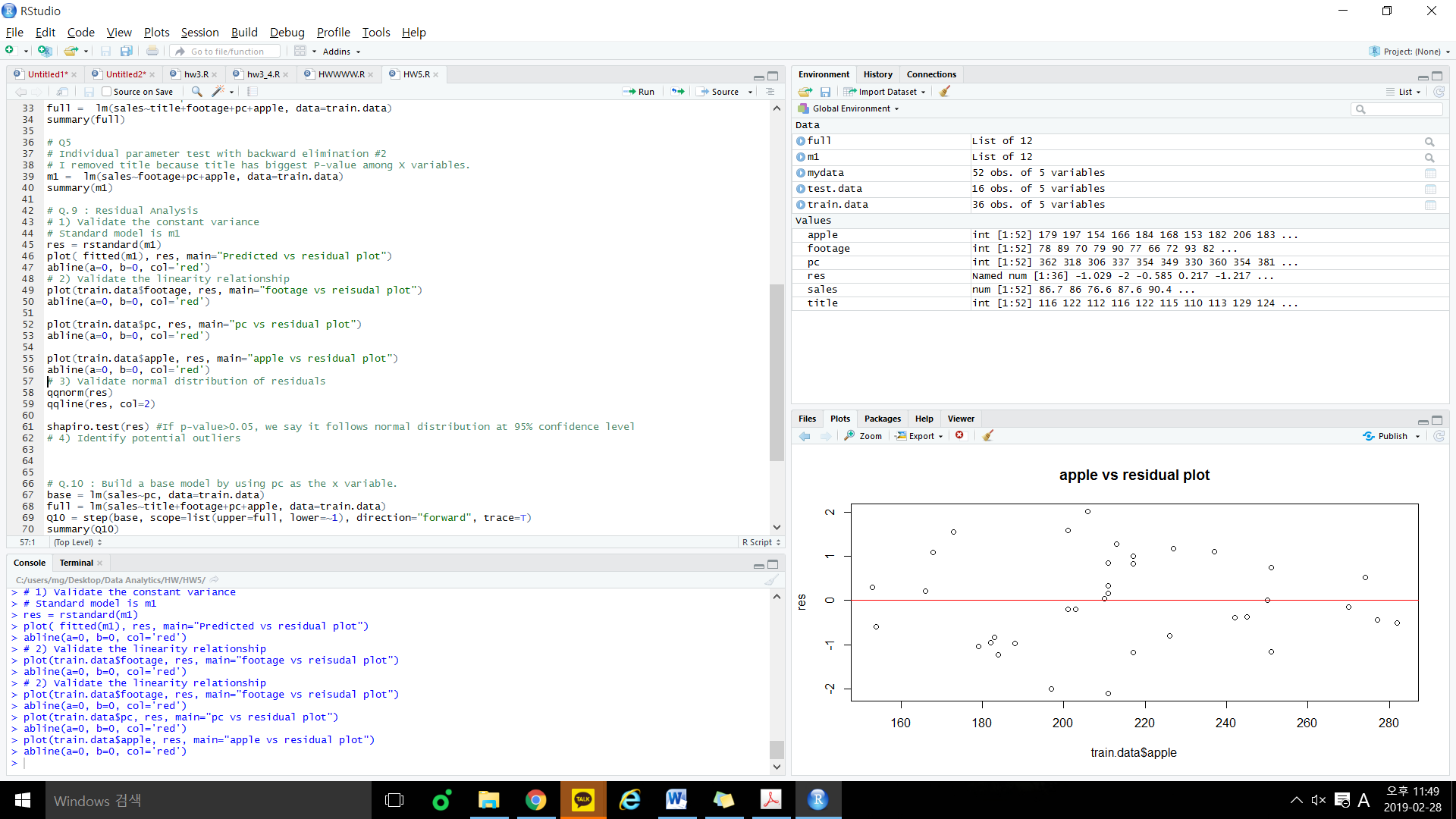


We know that constant variance from “Predicted vs residual plot” because we can see a constant spread.

1. Validate the linearity relationship
   1. Plot residuals vs each x-variable

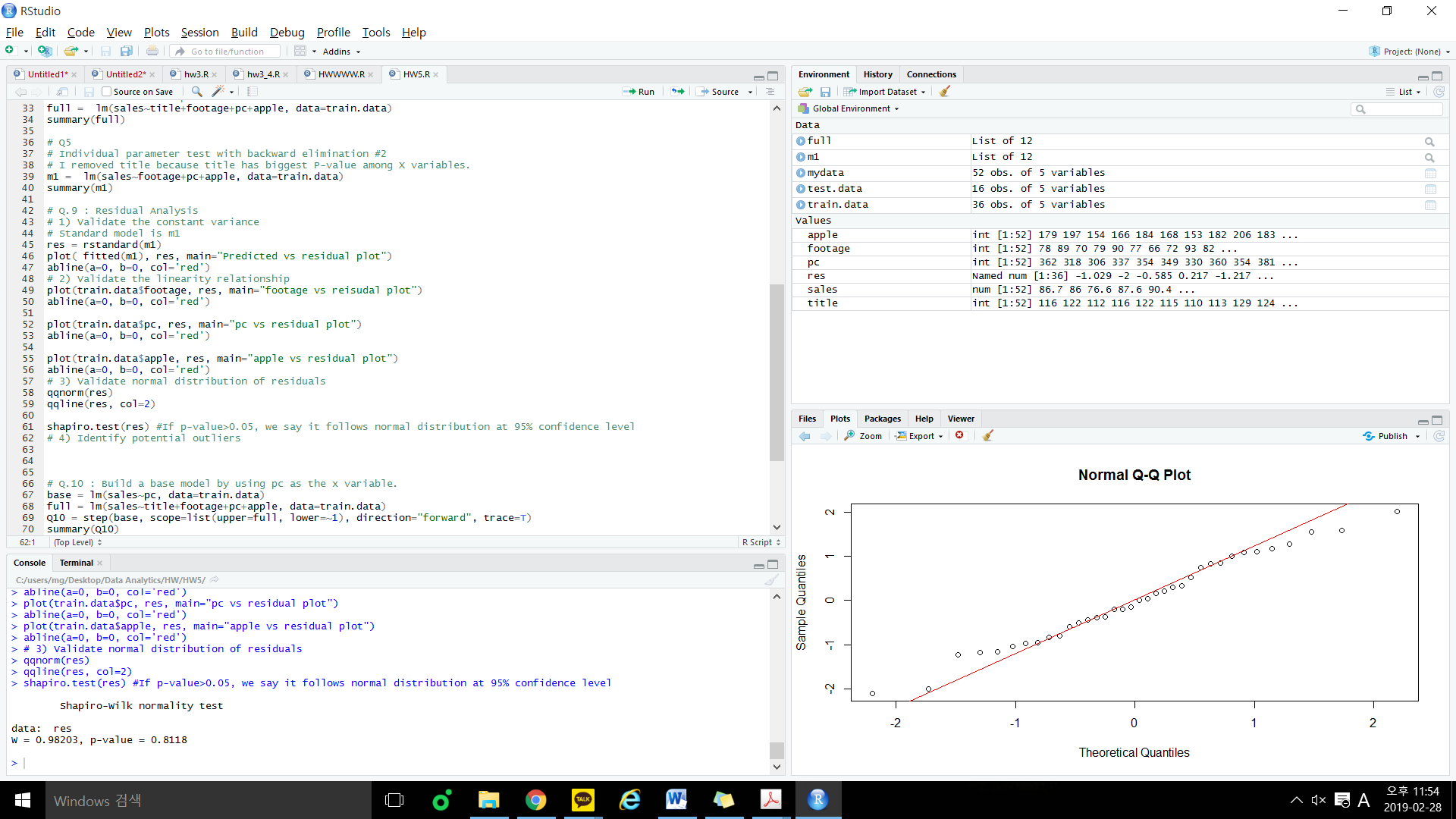






We know that linearity relationship of variables from plots above because points are randomly scattered around the zero line.

1. Validate normal distribution of residuals



Since p-value>0.5, we say that it follows normal distribution at 95% confidence level.

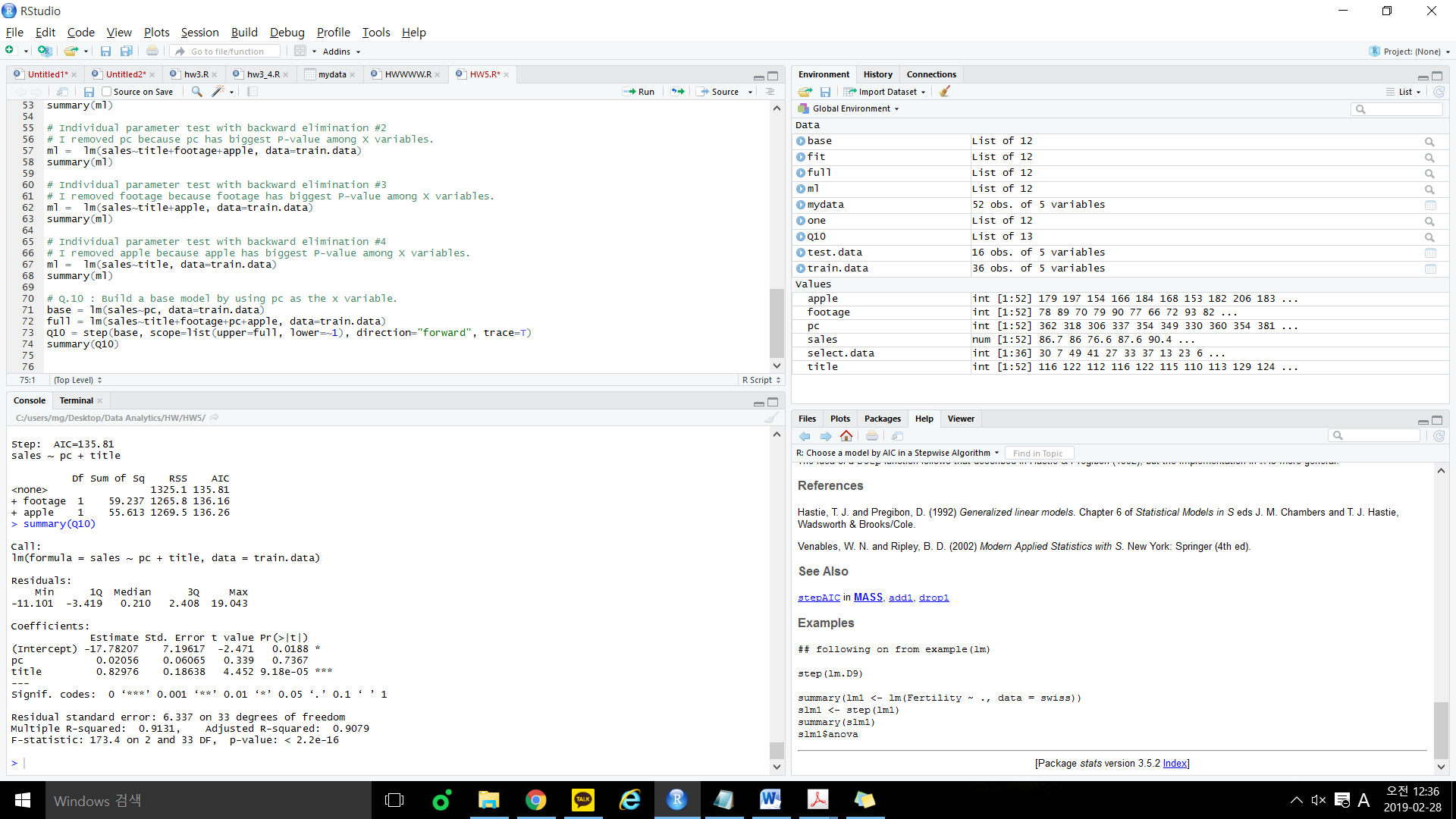
1. Identify potential outliers

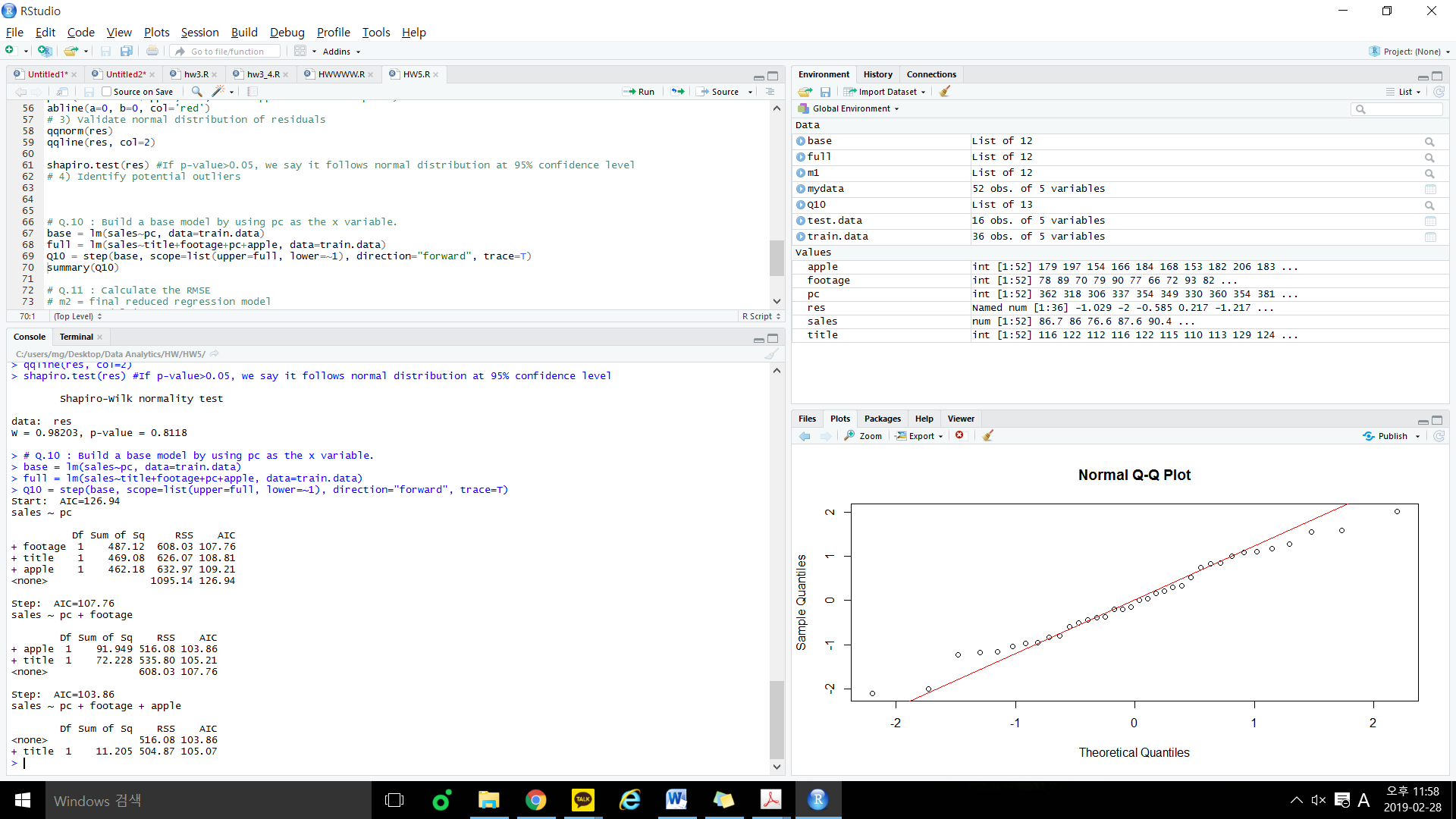
Possible outliers are observations with standardized/studentized residuals ||>3.

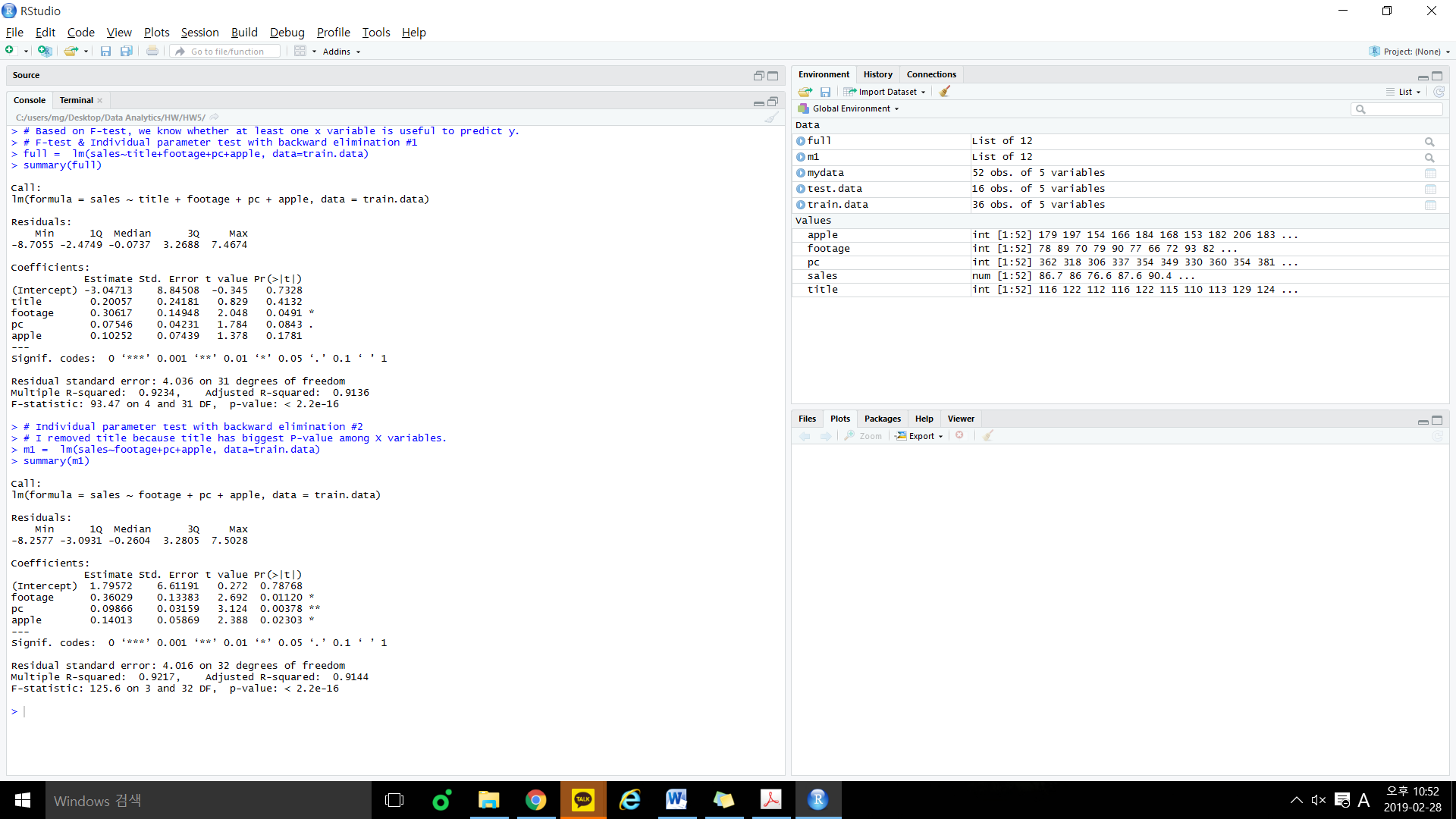
But I can’t see any outliers in plots above.

Because of these grounds, now I can say that this model is qualified.

10.[10] Build a base model by using pc as the x variable. Then perform forward selection and use AIC/BIC as the metric to build the model. Validate this model is qualified or not, if the model is different from the one you built in step 5). Compare the adj-R2 of these two models.





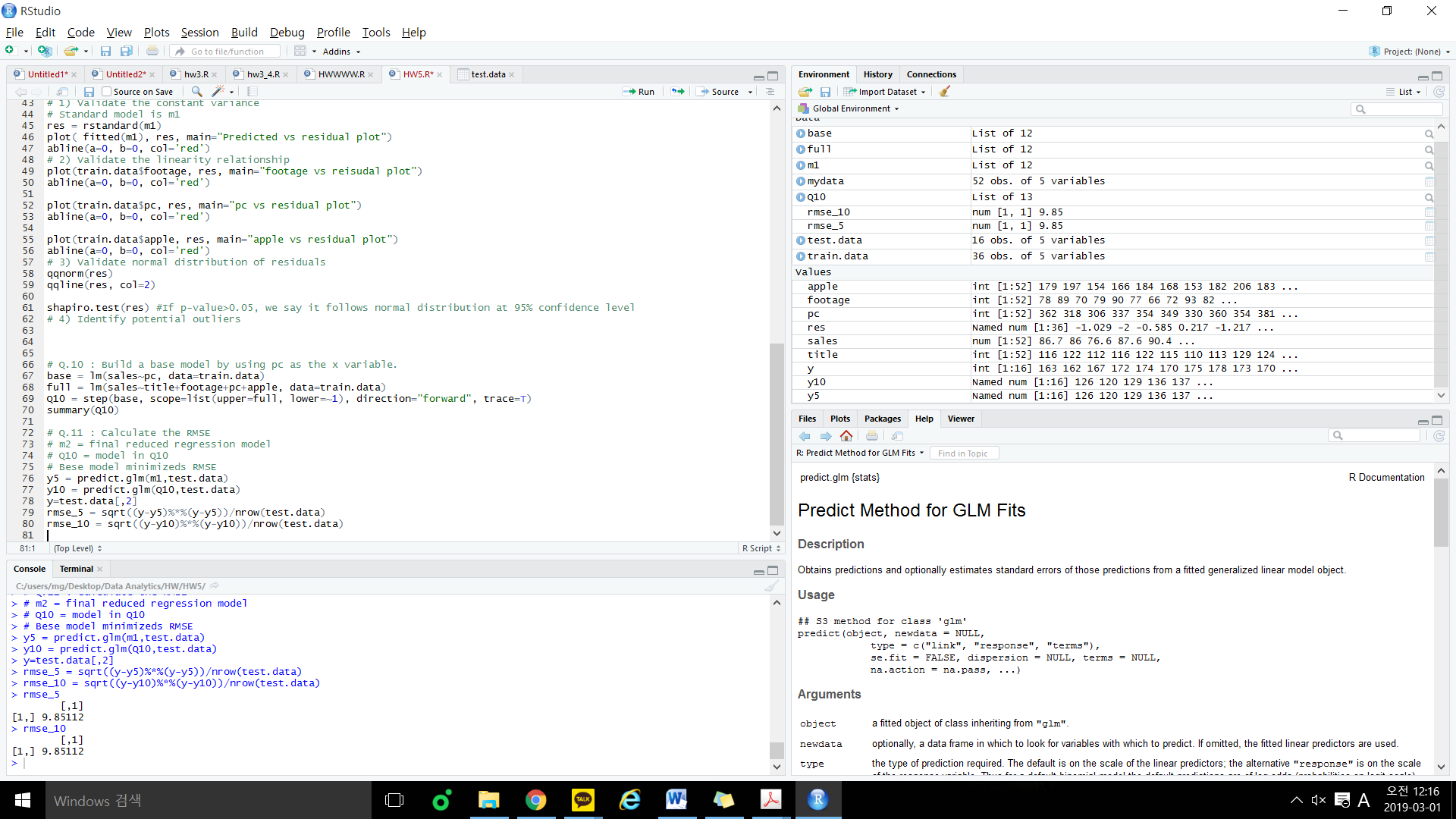


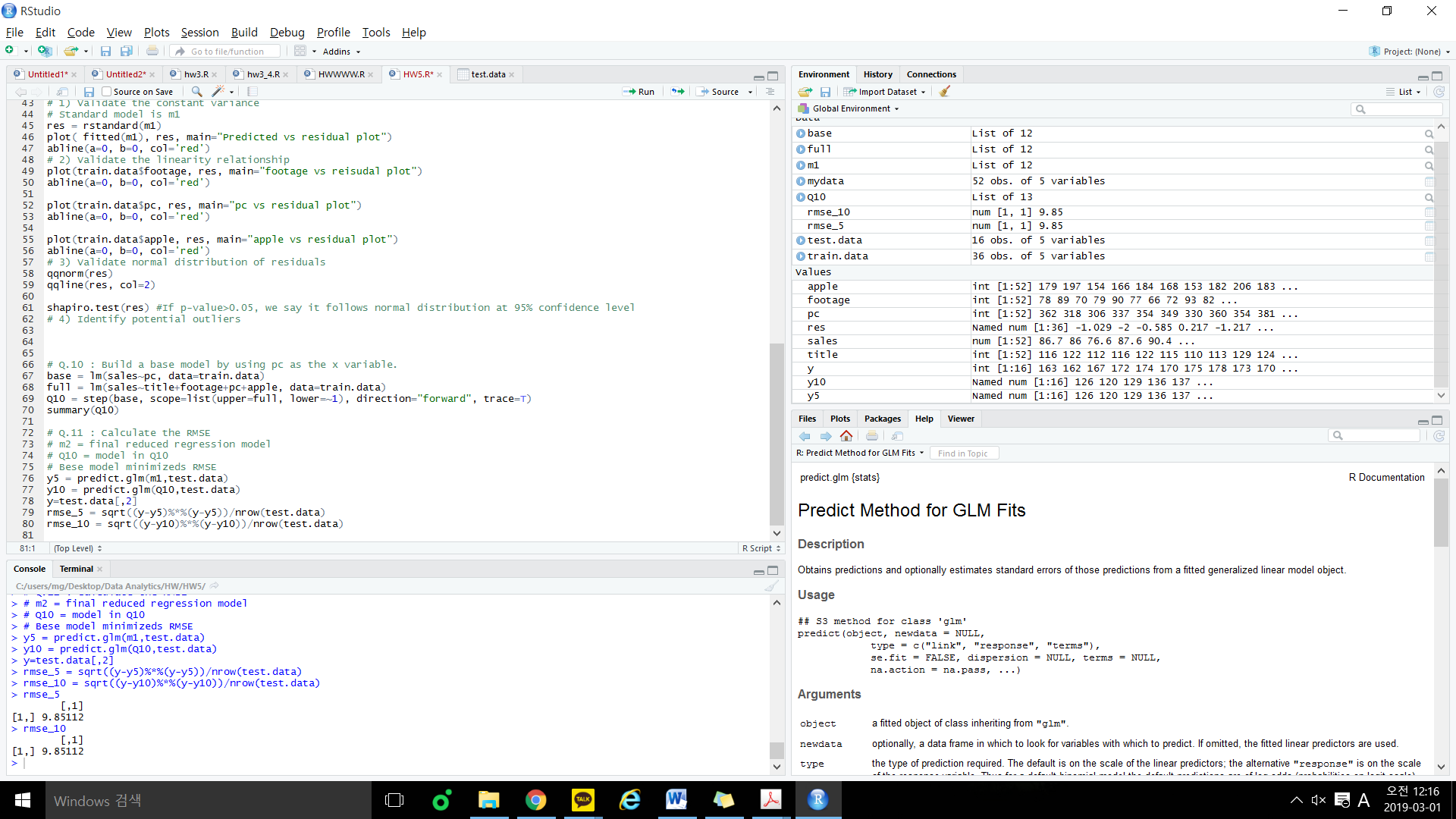
Model in Q10) Model in Q5)

Adj- of model in 10) = 0.9144

Adj- of model in 5) = 0.9144

11. [10] Calculate the RMSE for the model in 5) and 10), conclude that which model is better.





RMSE of model 5) = 9.85112

RMSE of model 10) = 9.85112, So both are same and qualified model.

#R commands

setwd('C:/users/mg/Desktop/Data Analytics/HW/HW5')

getwd()

mydata = read.table("HW5\_DATA.txt", header = T)

sales = mydata$sales

title = mydata$title

footage = mydata$footage

pc = mydata$pc

apple = mydata$apple

# plot about whole data sets.

plot(mydata)

plot(title, sales)

plot(footage, sales)

plot(pc, sales)

plot(apple, sales)

# Collinearity problem detections

# If there are two factors with strong correlations, i should remove one of them.

cor(cbind(sales,title,footage,pc,apple))

# Split data for hold-out evaluation

train.data = mydata[1:36,] # 70% of data set for training data

test.data = mydata[37:52,] # 30% of data set for testing data

# Build models

# Feature Selection // to distinguish useful x variables or not.

# Q4

# Based on F-test, we know whether at least one x variable is useful to predict y.

# F-test & Individual parameter test with backward elimination #1

full = lm(sales~title+footage+pc+apple, data=train.data)

summary(full)

# Q5

# Individual parameter test with backward elimination #2

# I removed title because title has biggest P-value among X variables.

m1 = lm(sales~footage+pc+apple, data=train.data)

summary(m1)

# Q.9 : Residual Analysis

# 1) Validate the constant variance

# Standard model is m1

res = rstandard(m1)

plot( fitted(m1), res, main="Predicted vs residual plot")

abline(a=0, b=0, col='red')

# 2) Validate the linearity relationship

plot(train.data$footage, res, main="footage vs reisudal plot")

abline(a=0, b=0, col='red')

plot(train.data$pc, res, main="pc vs residual plot")

abline(a=0, b=0, col='red')

plot(train.data$apple, res, main="apple vs residual plot")

abline(a=0, b=0, col='red')

# 3) Validate normal distribution of residuals

qqnorm(res)

qqline(res, col=2)

shapiro.test(res) #If p-value>0.05, we say it follows normal distribution at 95% confidence level

# 4) Identify potential outliers

# Q.10 : Build a base model by using pc as the x variable.

base = lm(sales~pc, data=train.data)

full = lm(sales~title+footage+pc+apple, data=train.data)

Q10 = step(base, scope=list(upper=full, lower=~1), direction="forward", trace=T)

summary(Q10)

# Q.11 : Calculate the RMSE

# m2 = final reduced regression model

# Q10 = model in Q10

# Bese model minimizeds RMSE

y5 = predict.glm(m1,test.data)

y10 = predict.glm(Q10,test.data)

y=test.data[,2]

rmse\_5 = sqrt((y-y5)%\*%(y-y5))/nrow(test.data)

rmse\_10 = sqrt((y-y10)%\*%(y-y10))/nrow(test.data)