**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:

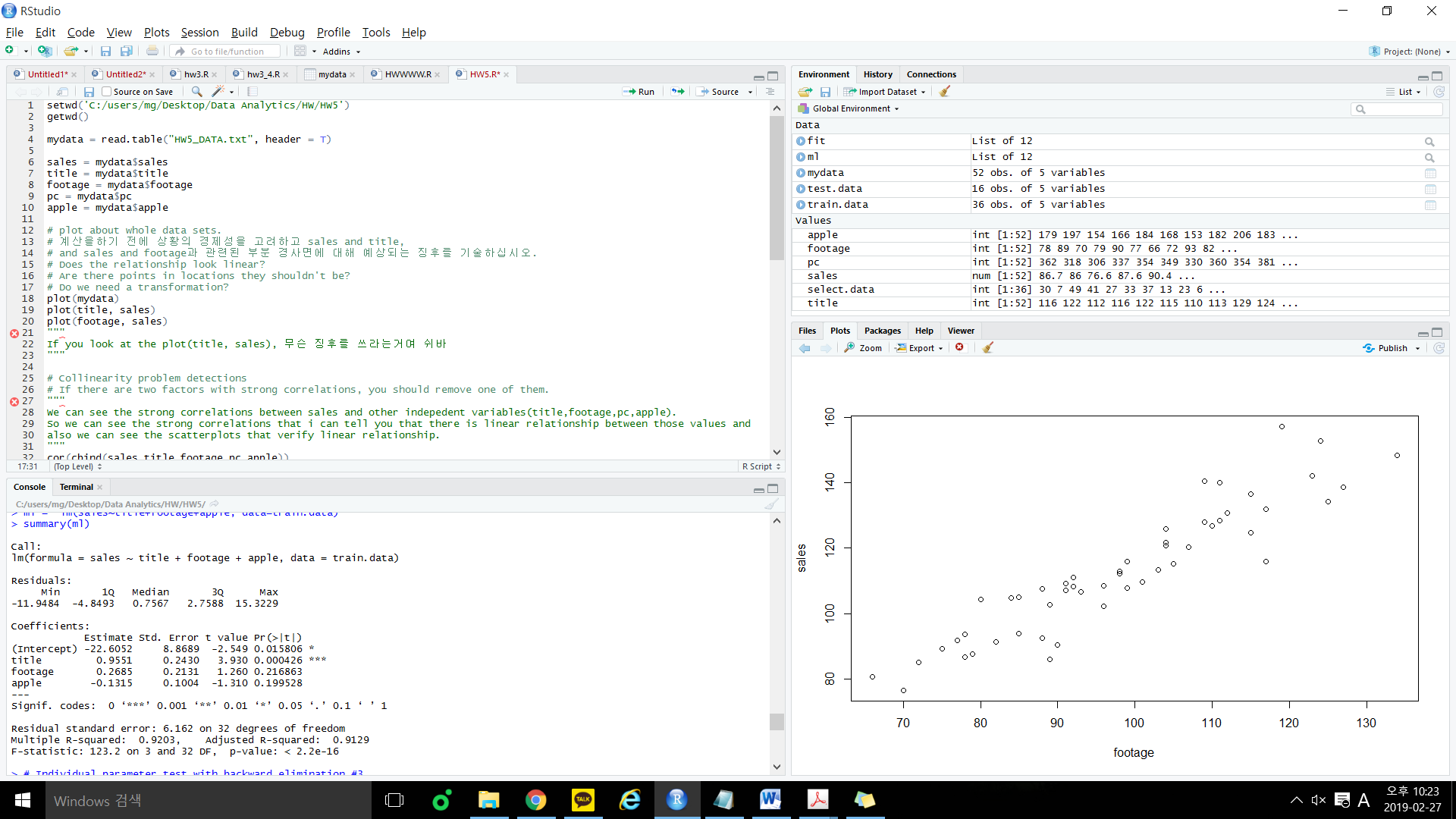
* Every step you use R, you should provide the snapshots of your R commands and R outputs, and paste the plots if it is necessary.
* 너가 R을 쓰는 단계마다 너는 너의 R커맨드와 아웃풋의 스냅샷을 제공해야하고, 필요하다면 plot를 제공해야한다.
* Use 95% as confidence level for the following questions
* 따라오는 문제에 대해서 95% 신뢰수준을 적용해라
* Do NOT shuffle 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
* Data를 섞지말아라. 앞의 0.7row를 training으로 쓰고, 남은거를 testing으로 써라.

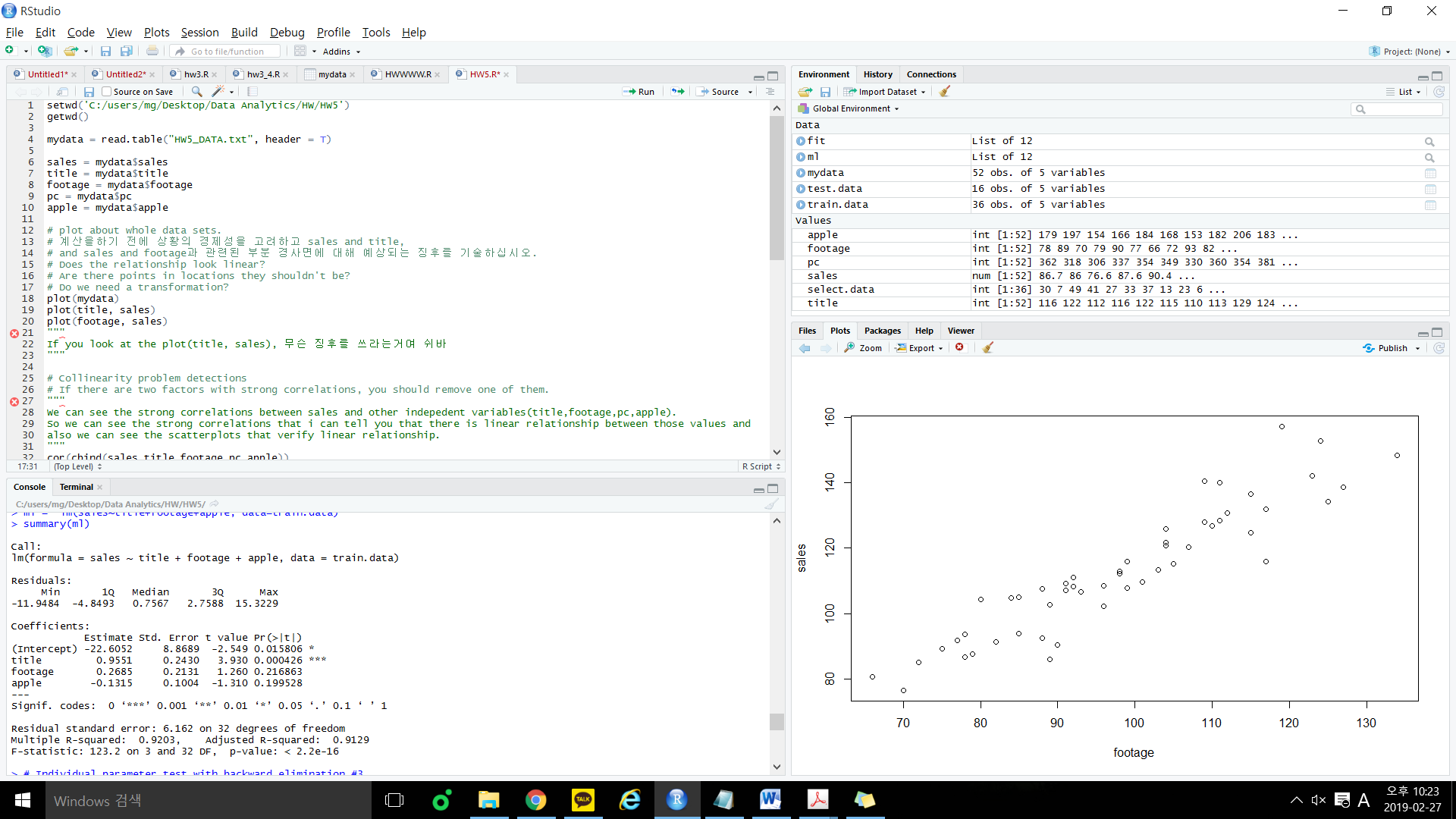
Compute the appropriate regression analysis using R and answer the following questions:

R을 이용하여 적절한 회귀분석을 계산하고 답하여라.

#코딩전문올리기

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



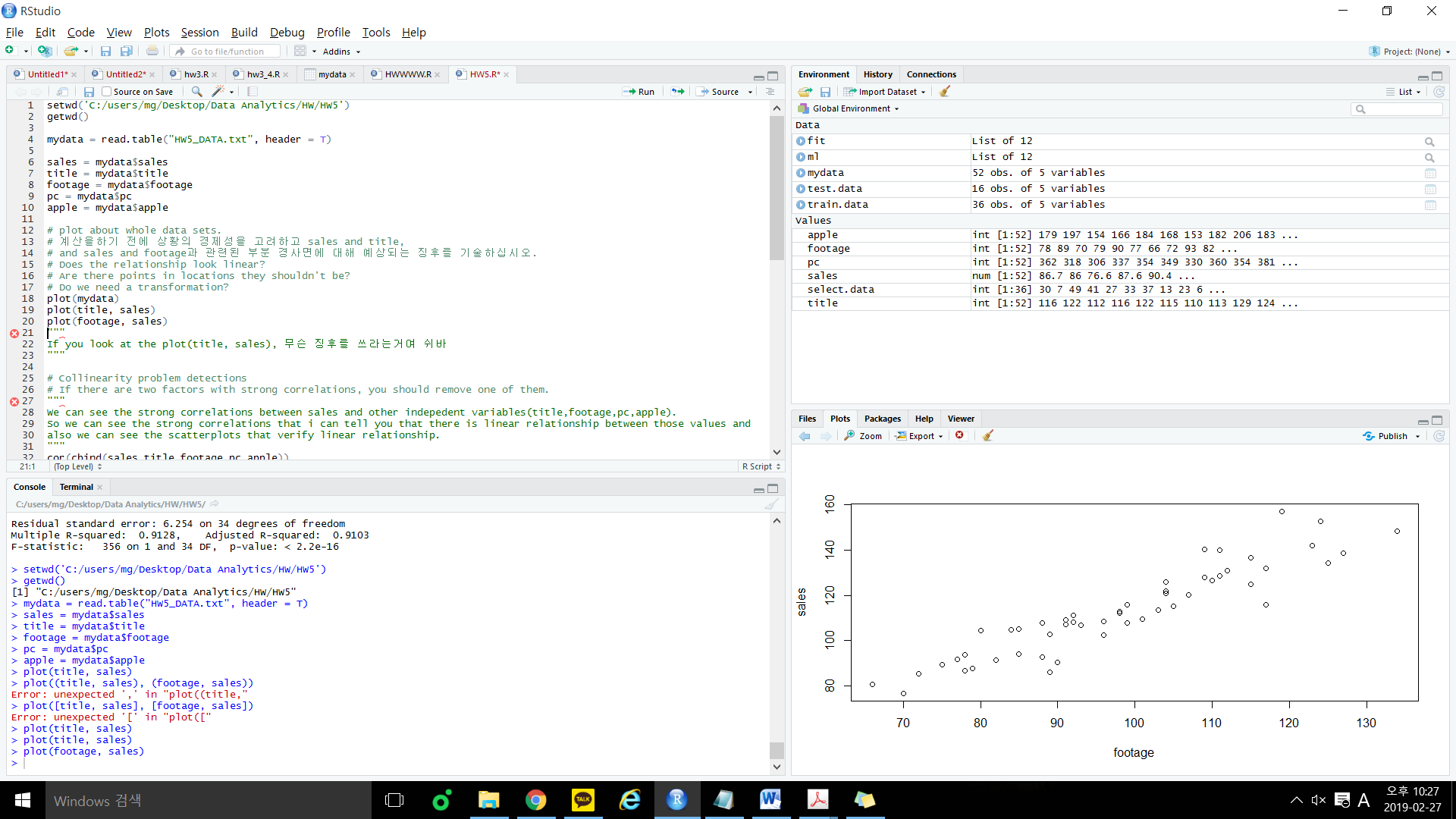
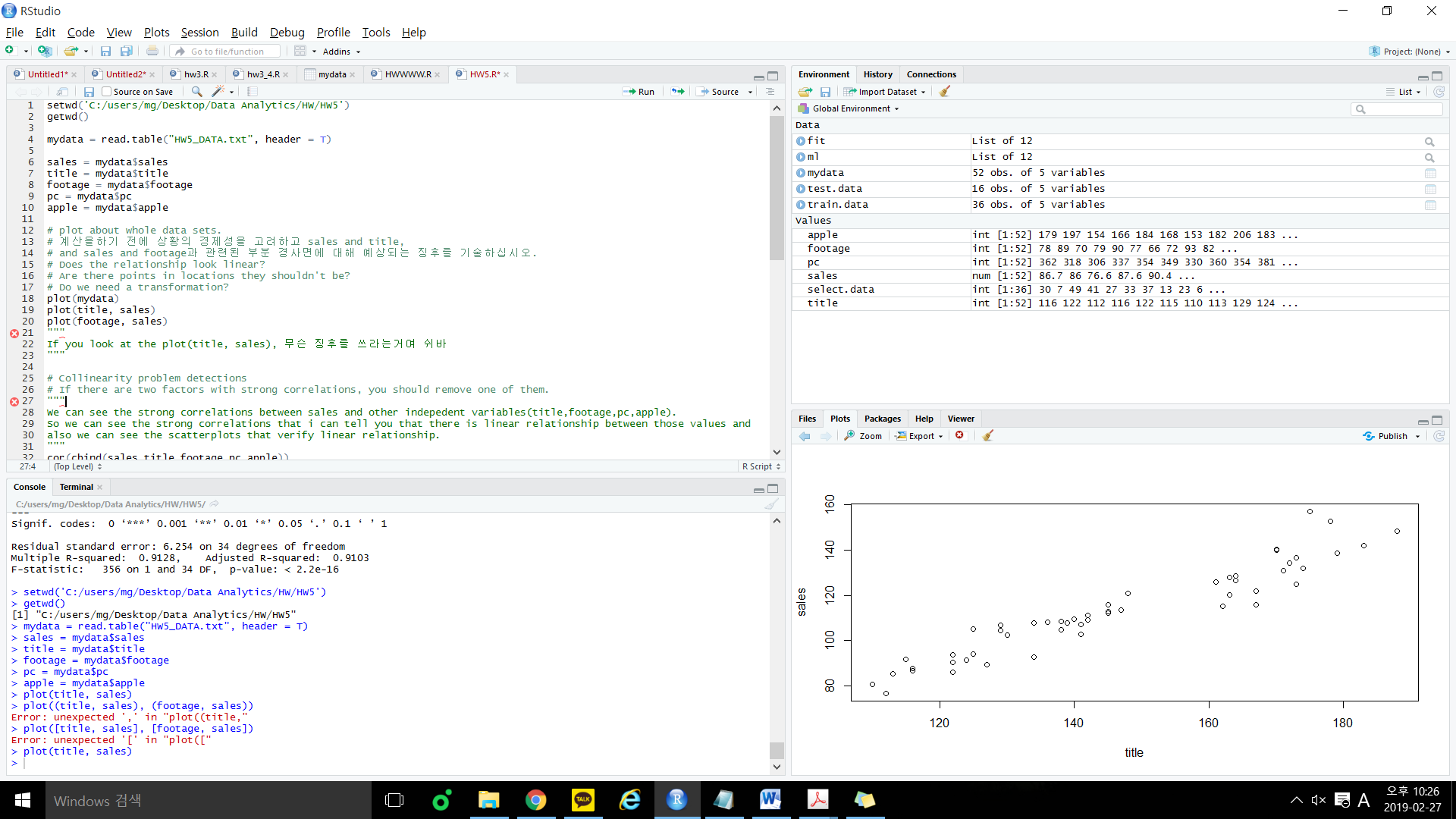


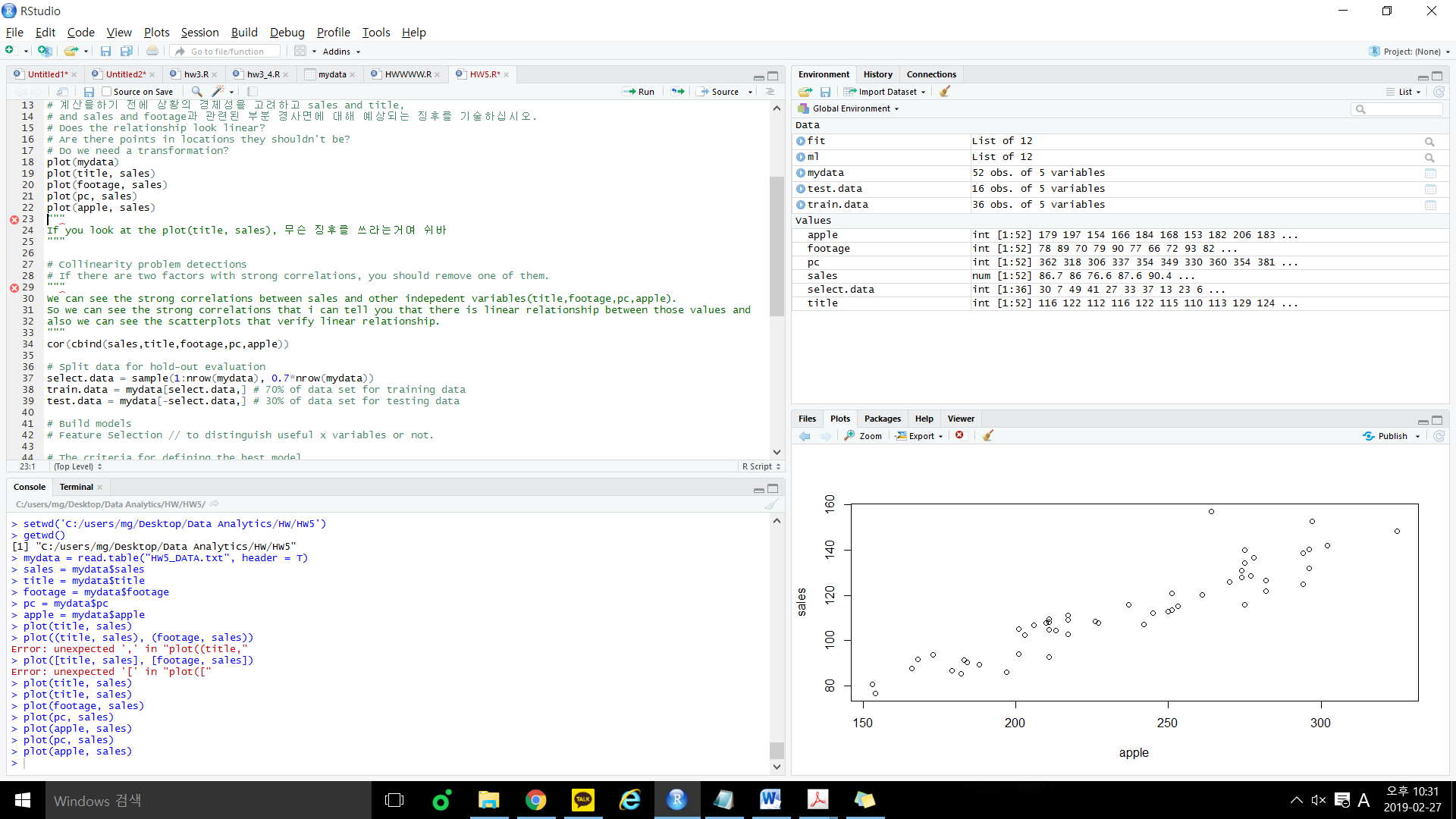
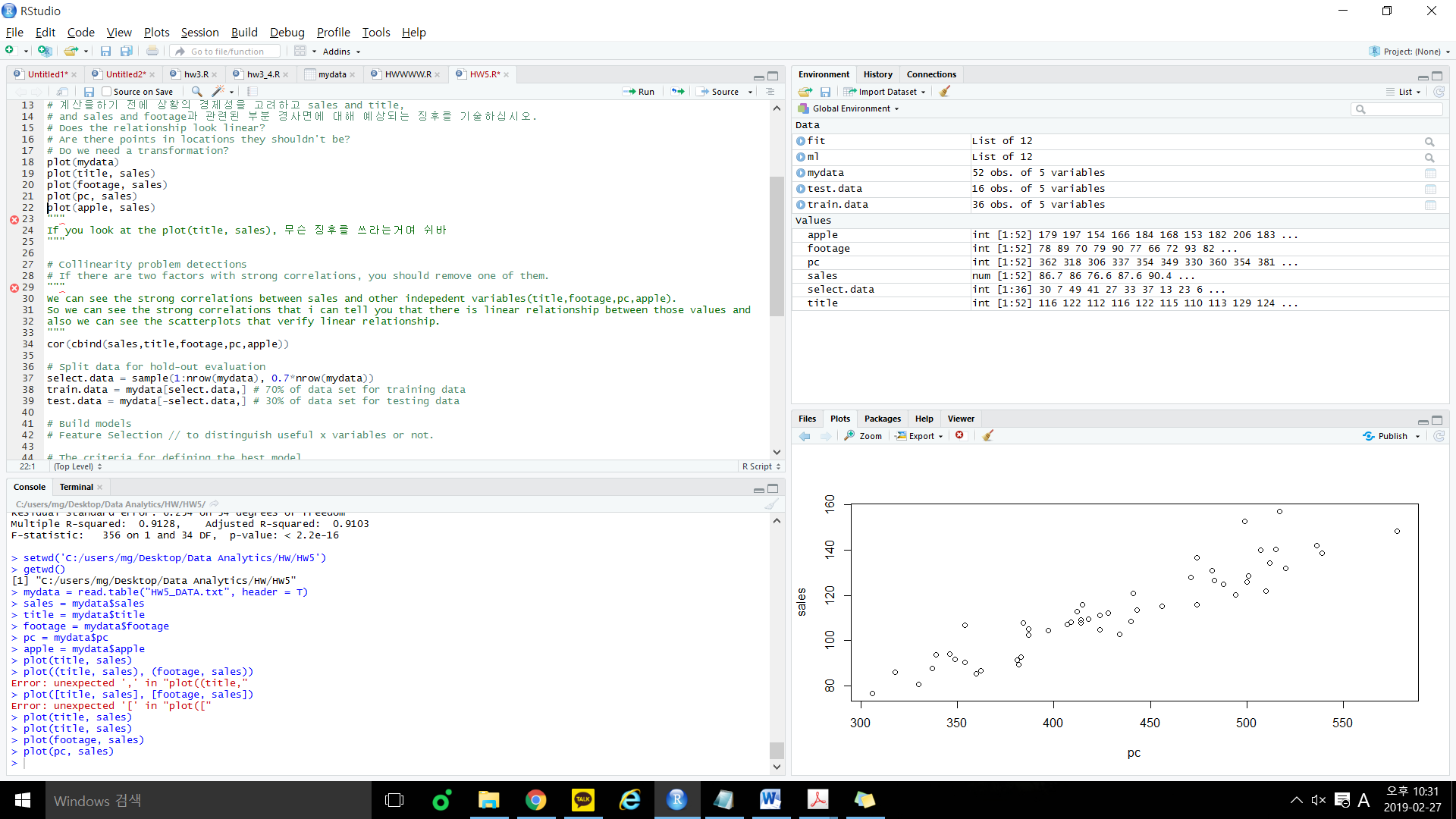
2. [10] Before doing any calculation, consider the economics of the situation and state what sign you would expect for the partial slopes relating sales and title, and sales and footage.

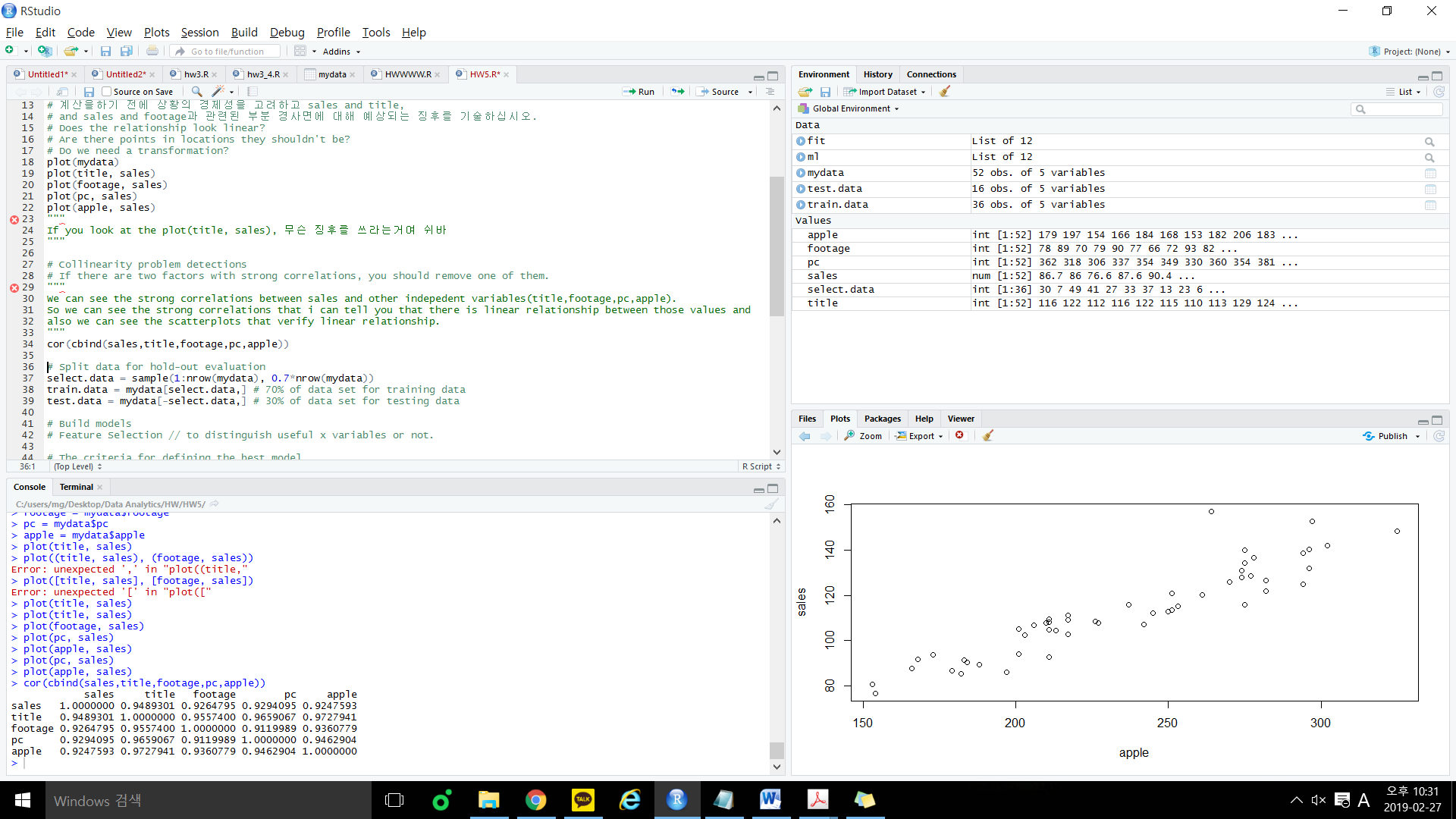
**The increase in biweekly sales of the computer software must be related to some of variables.**

**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, high number of values means strong correlation between and Y.

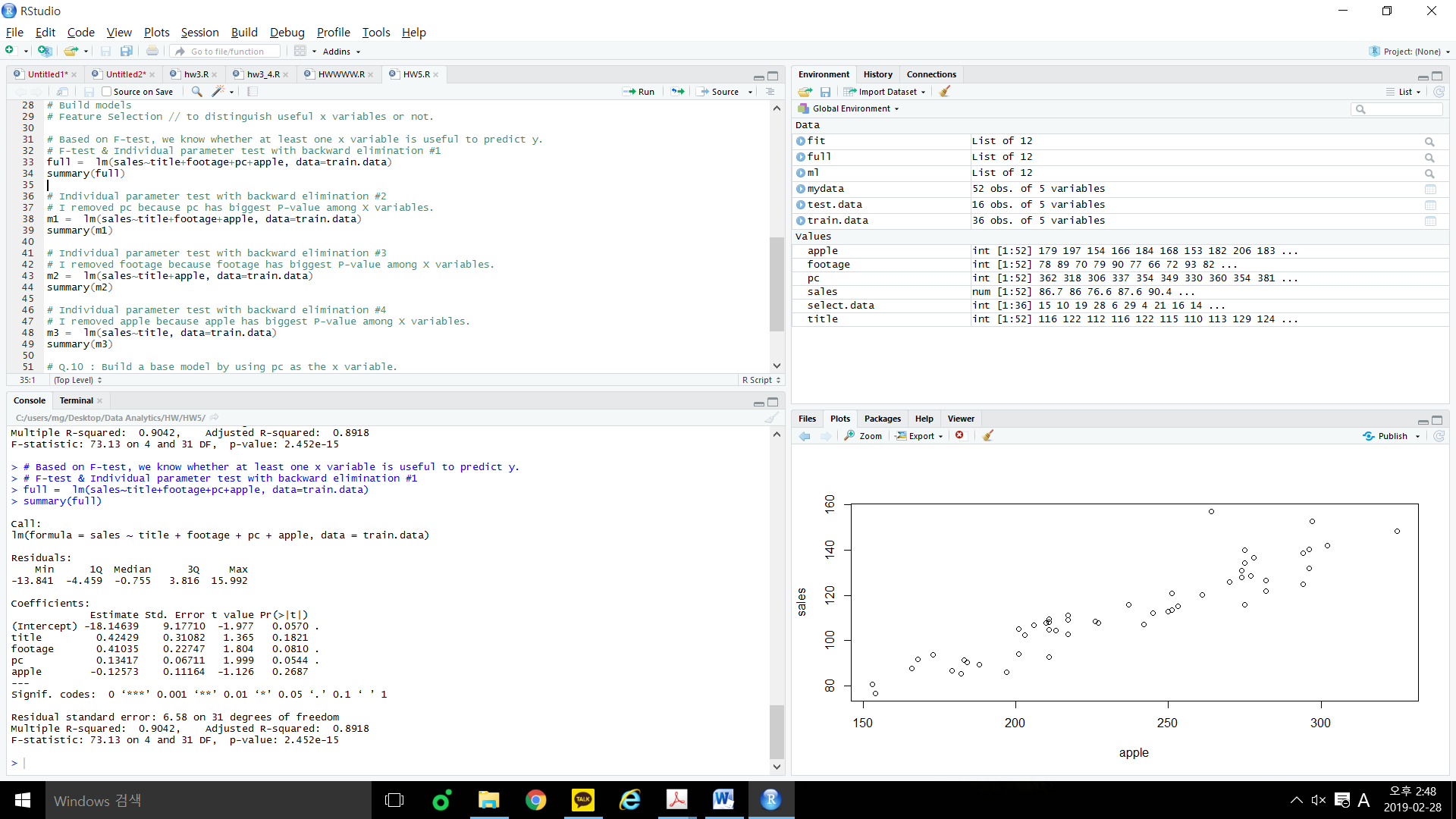
Therefore, we don’t have to use transformations.

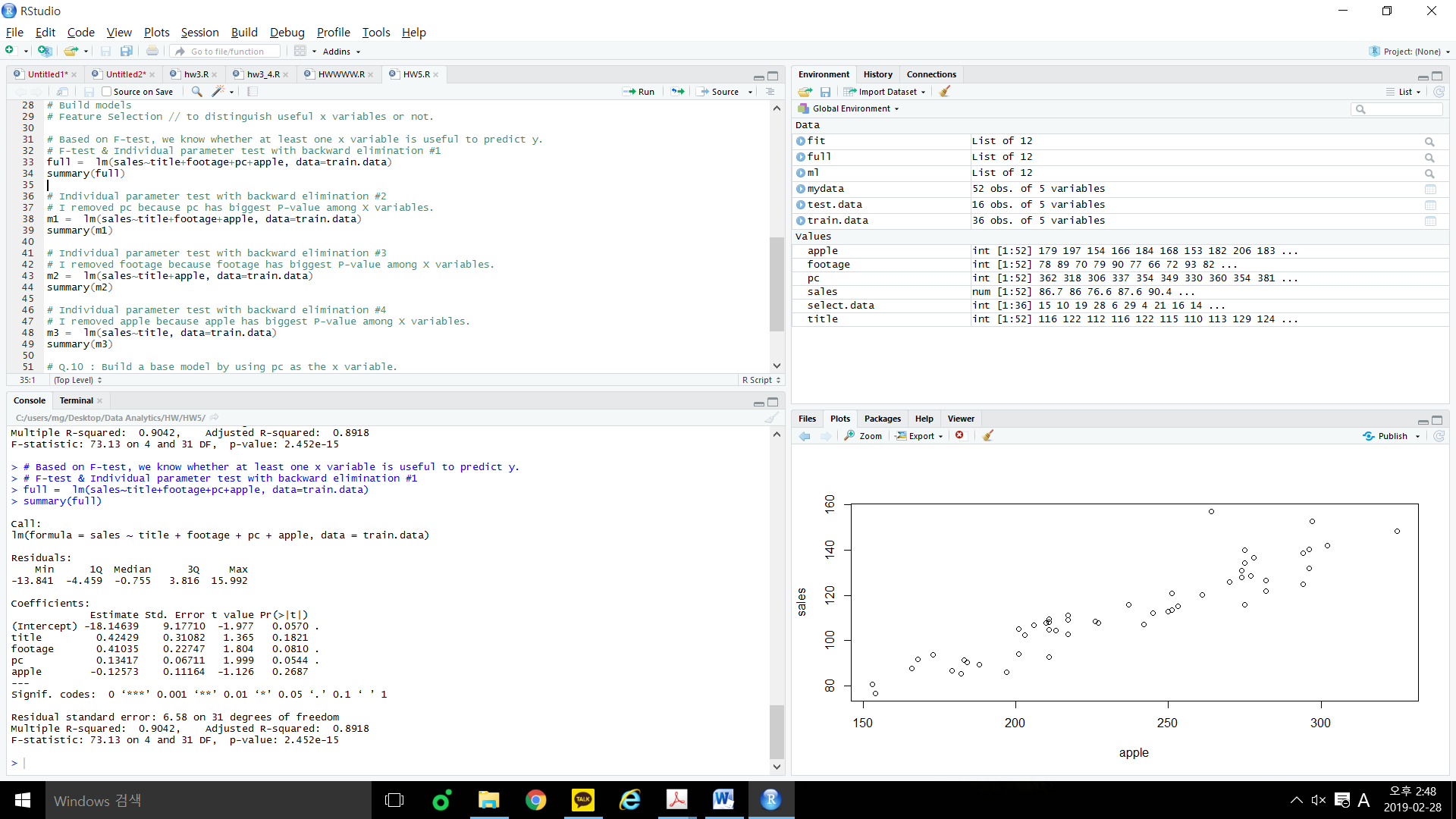
4. [10] Compute in R a multiple regression equation with sales as the dependent variable Y and  
title, footage, pc and apple as the predictors. Write down the expression for the fitted  
regression model. [Run it once, use backward elimination but you do not need to remove X variables in this step]. 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: < 2.452e-15 which means P-value is smaller than .

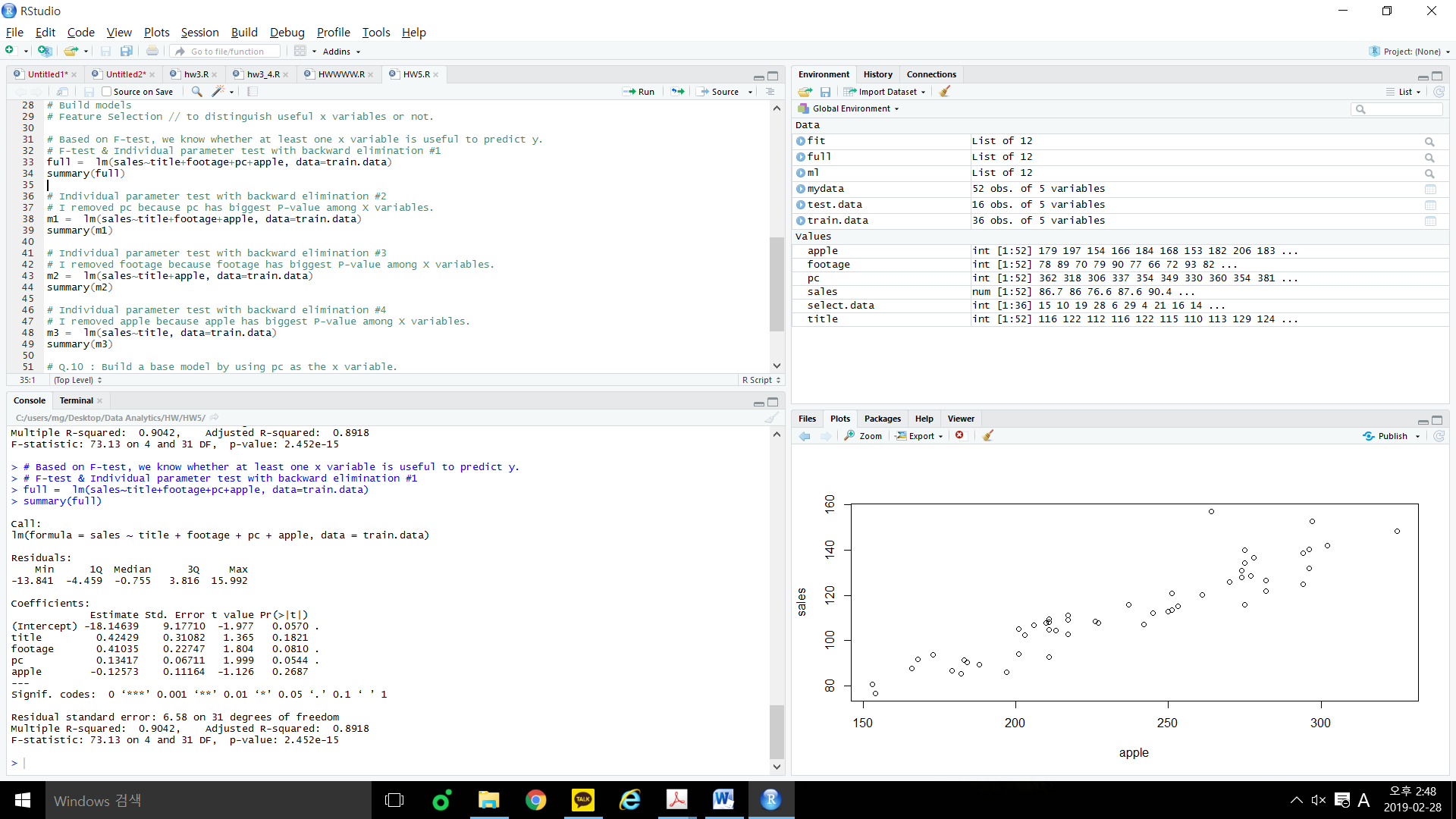
So the conclusion is at 95% confidence level, at least one X variable has significant linear relationship with Y, and it can affect the value of the Y.

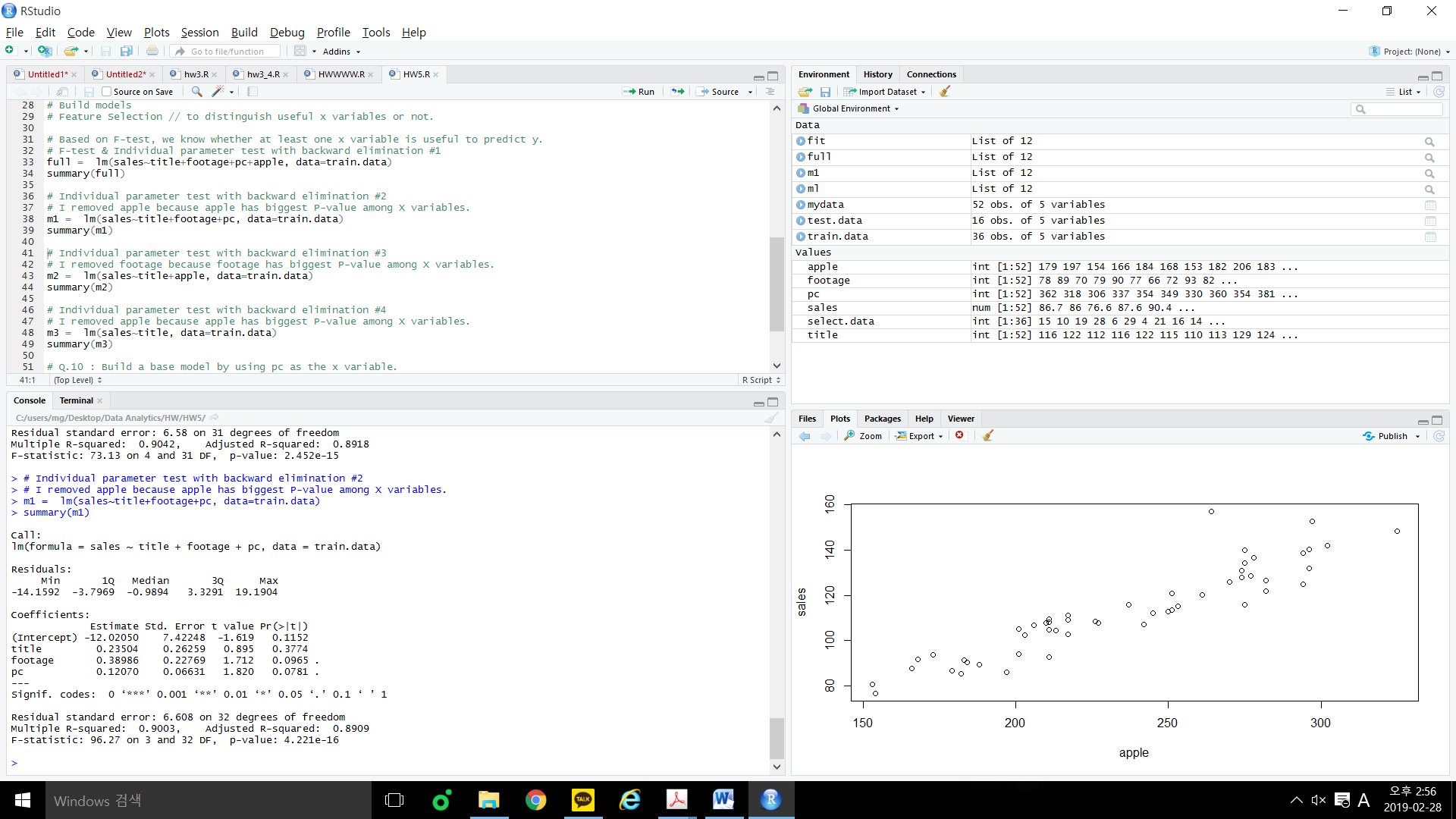
Individual parameter test)

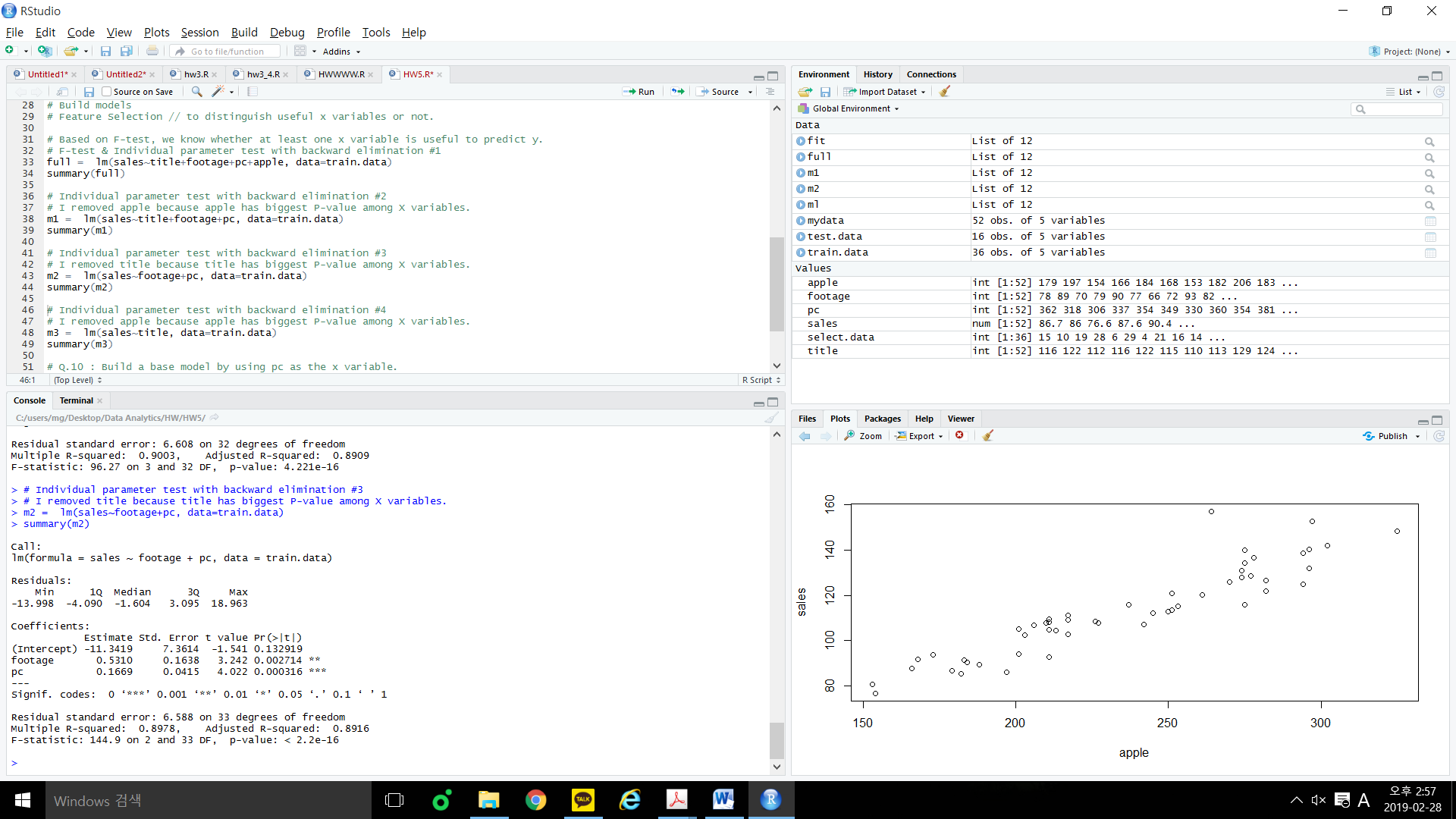
Apple has largest p-value among x variables, so we should remove apple first.

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)?

p-value를 메트릭으로 사용하고 모델을 얻기 위해 역방향 제거를 수행하여 최종 축소 회귀 모델에 대한 표현식을 작성합니다. 테스트 결과를 바탕으로 Y에 어떤 변수가 중요한 영향을 미쳤습니까? 각 추정 계수에는 2)에서 예상 한 부호가 있습니까?





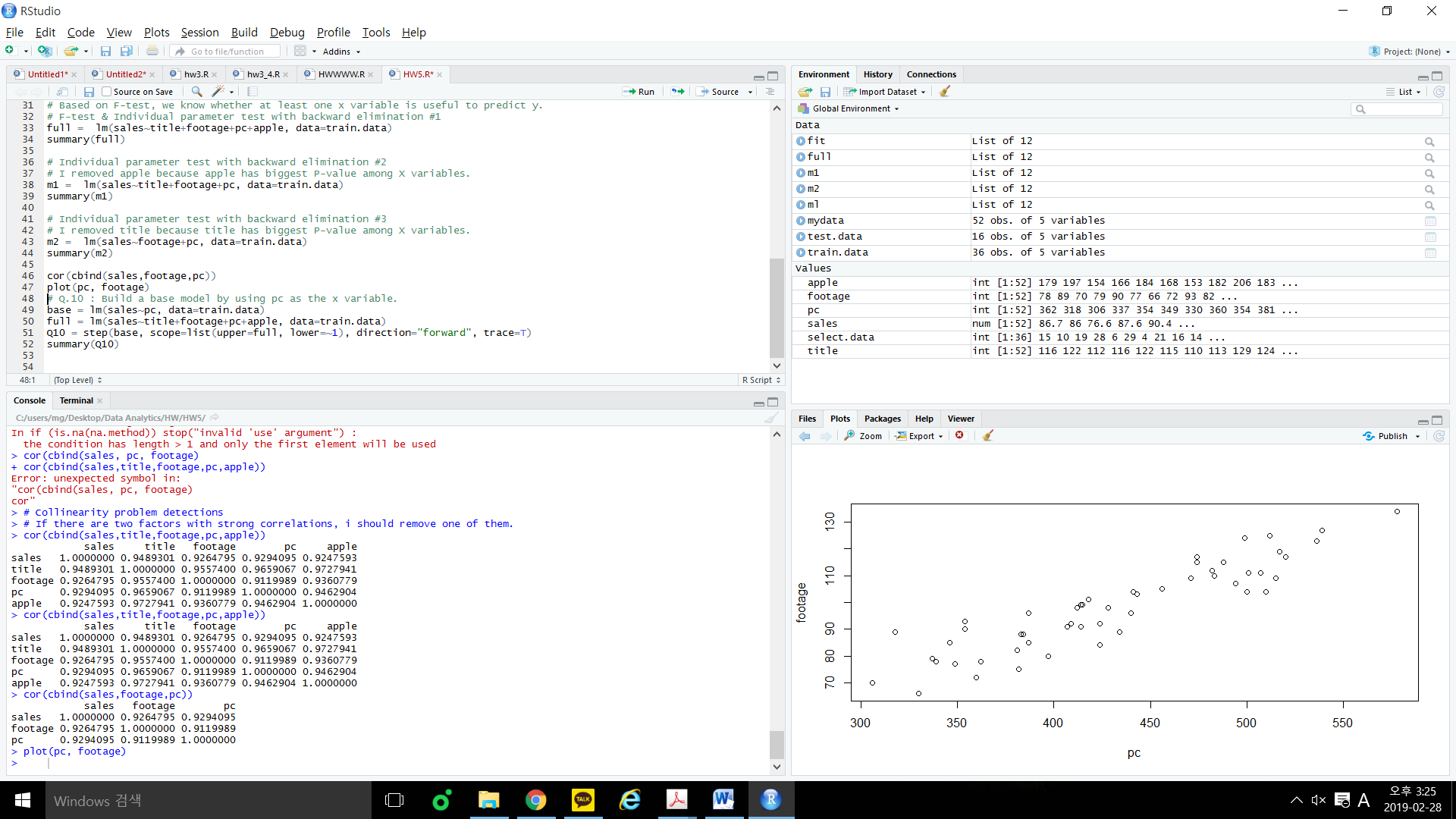
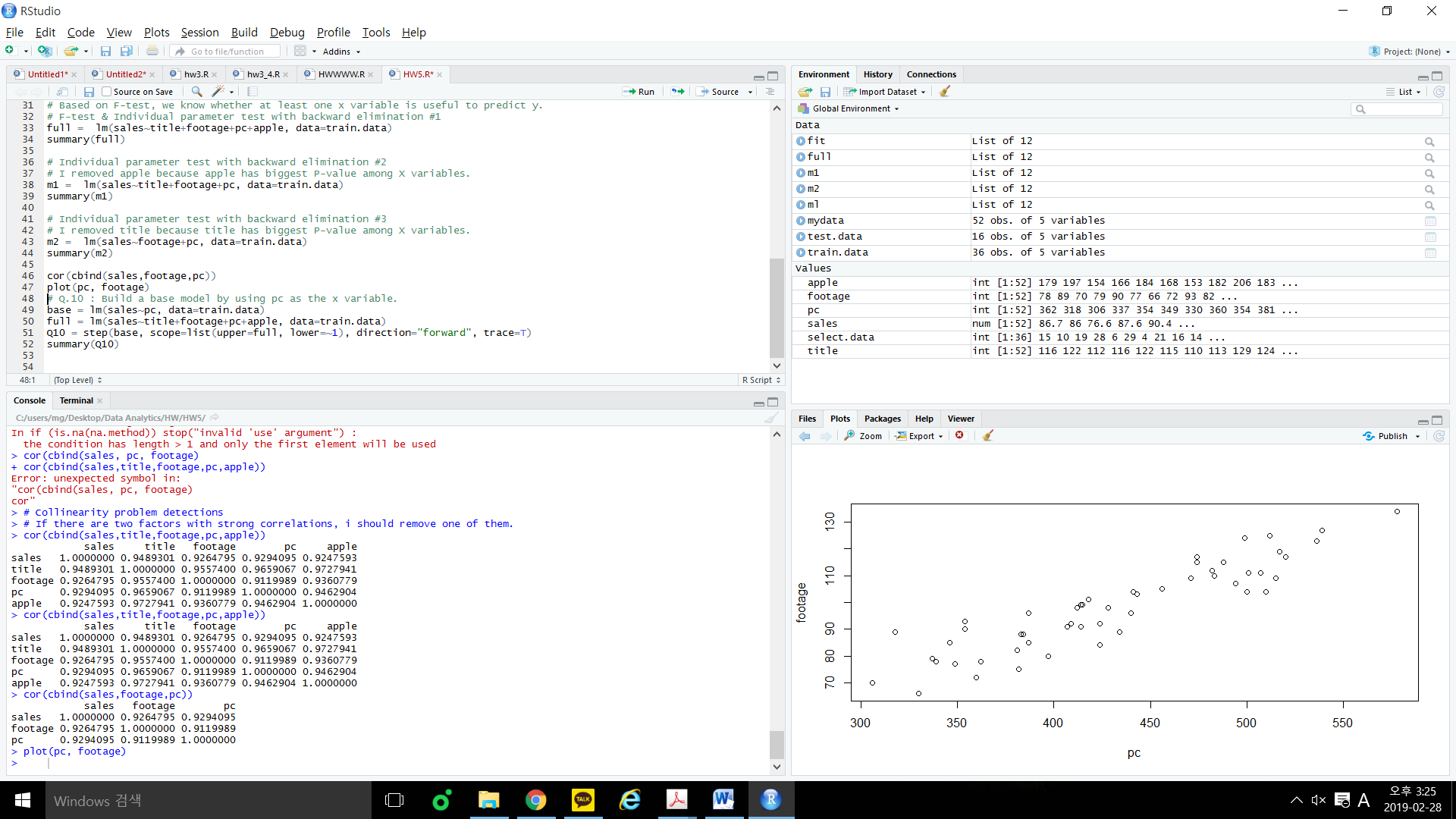


I removed x variables which have biggest P-value in each step and which is greater than 0.05 one by one in R.

Reduced regression model)

=

=



But we can see that there is strong correlation between footage and pc through these.

So here is two factors with strong correlation, I decide to remove one of them(which is pc)

Final reduced regression model)

=

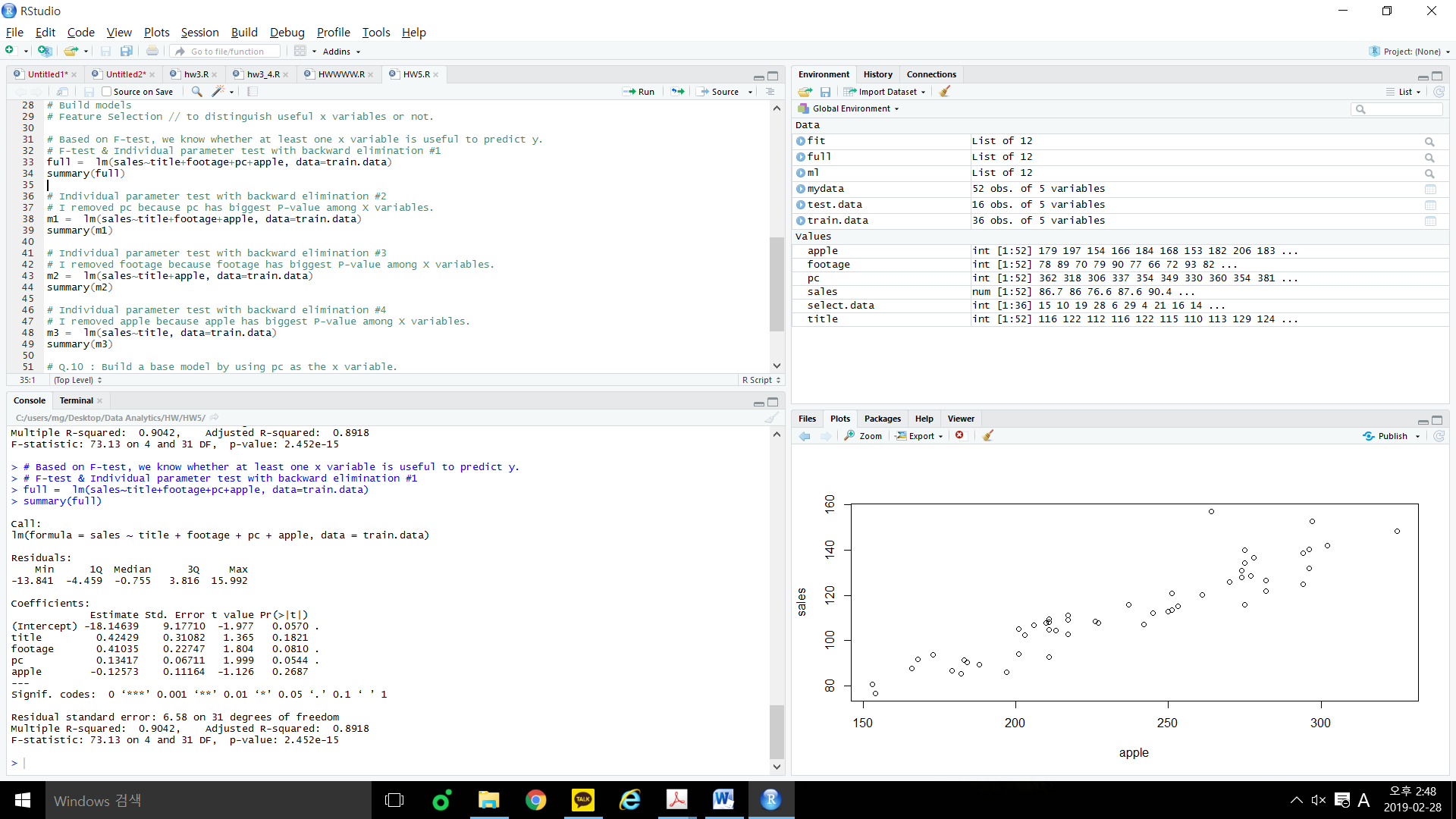
=

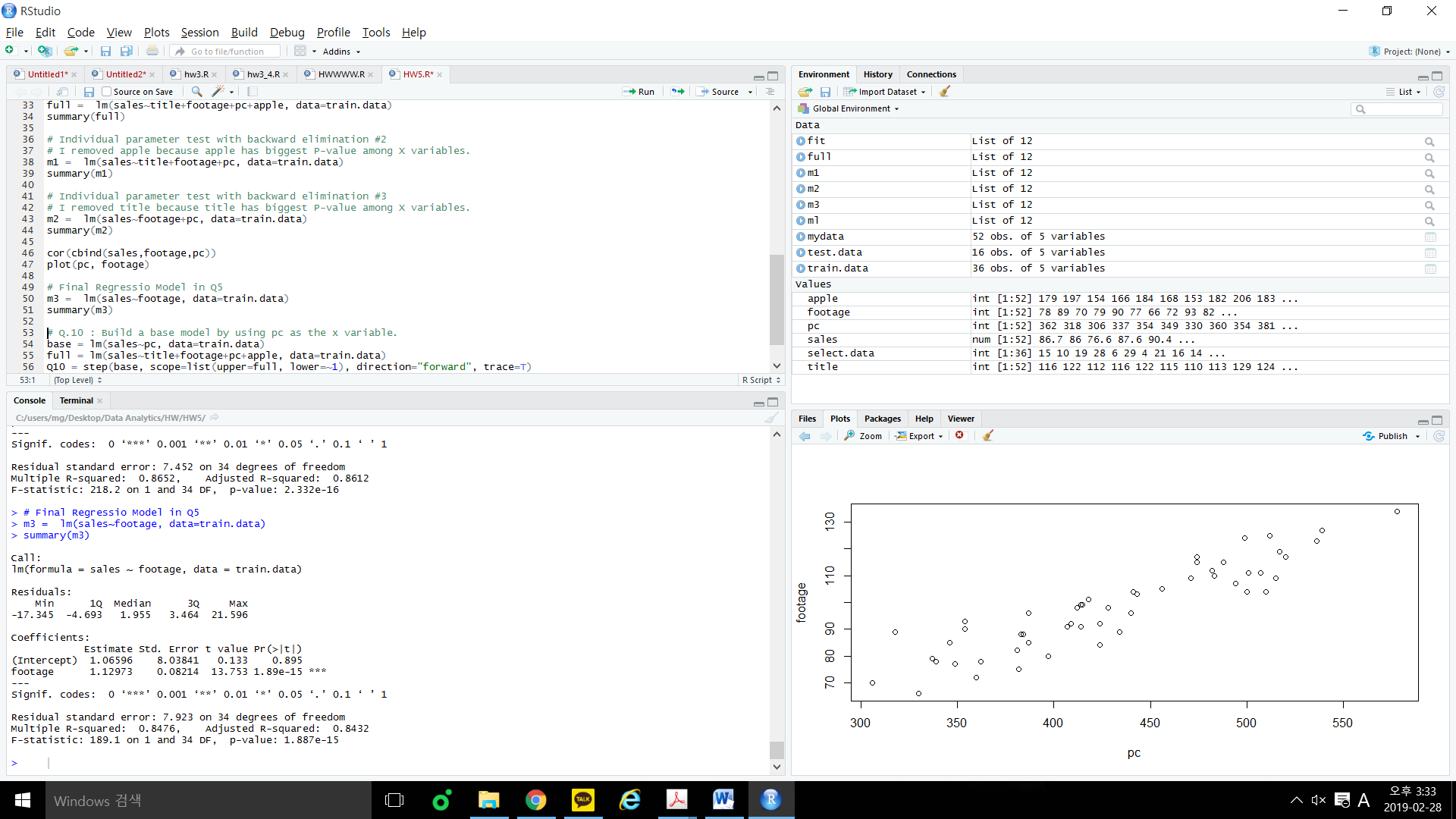
So, footage variable had a significant impact on sales.

I expected if footage variable is related to sales, the relationship between sales and footage will be linearly proportional and estimated coefficient confirms my prediction.

.

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, I thought reduced model is better model comparing to full model but Adj- values tell us full model fitted in 4) is better. RMSE도 또한 구해보아서 비교해야할 것 같다.

여기서 값이 (5)가 더 높아야 정상인데, 일단 계속 진행해봄.

아마 train data만으로 했기에 정확도가 떨어질수잇을듯

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.

<여기 다시해야해>

5)의 축소모델을 사용하여 80foot display area안에 있는 PC를 사용하는 400명의 손님들과 apple을 사용하는 215명의 손님들이 있는 130titles에 대한 sales value를 예측하십시오.

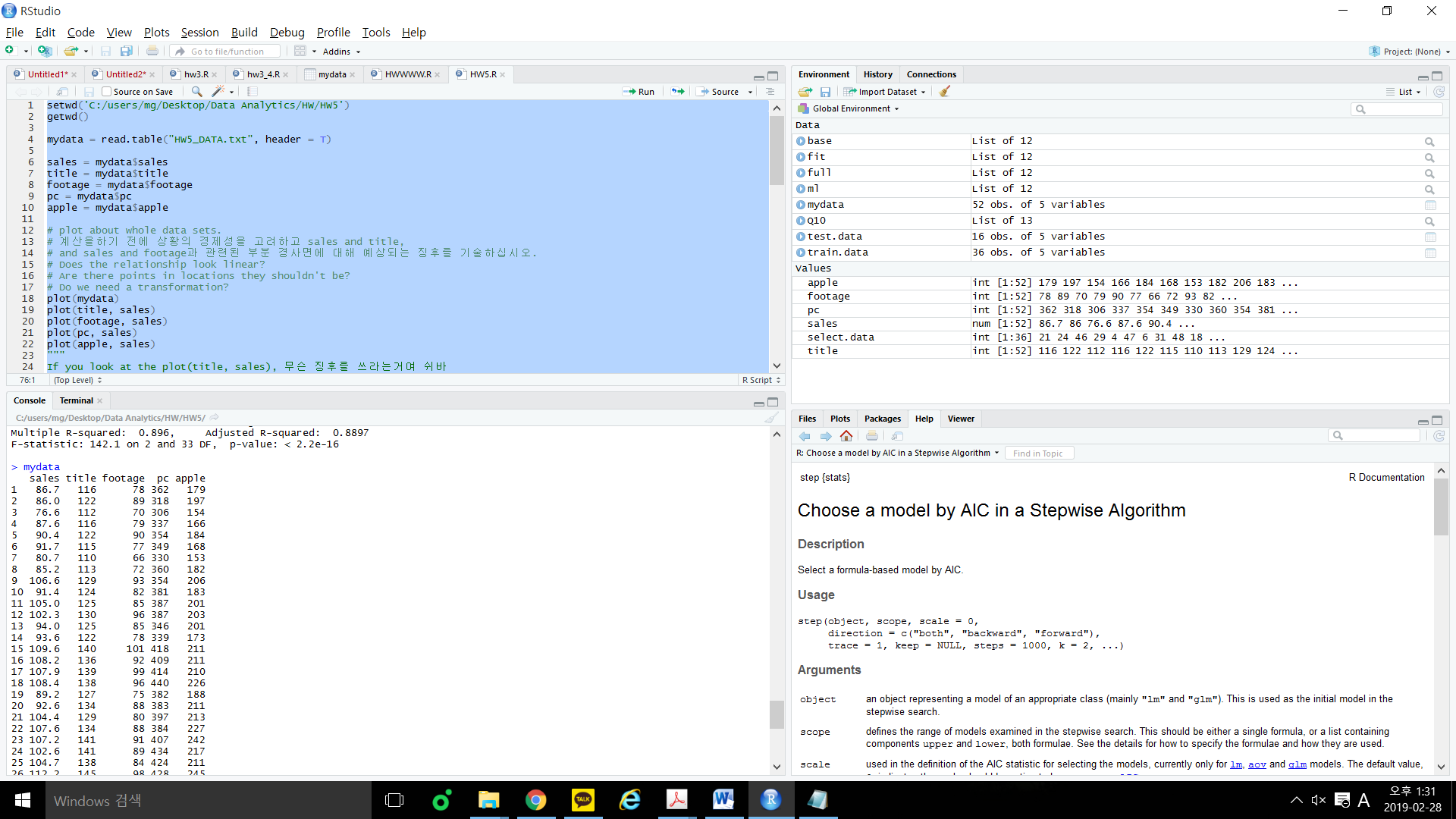
Txt의 observation중에 이와 비슷한게 잇는지 찰고 prediction error를 계산하라는 것인듯.

=

=

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

=



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

So prediction error = observed value – predicted value = 104.4 – 82.88 = 21.52

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

7)의 파이널 모델에 있는 기울기를 설명하십시오

**<여기 다시해야해>**

Slopes is equal to 0.81, so it has a gentle rising straight line.

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

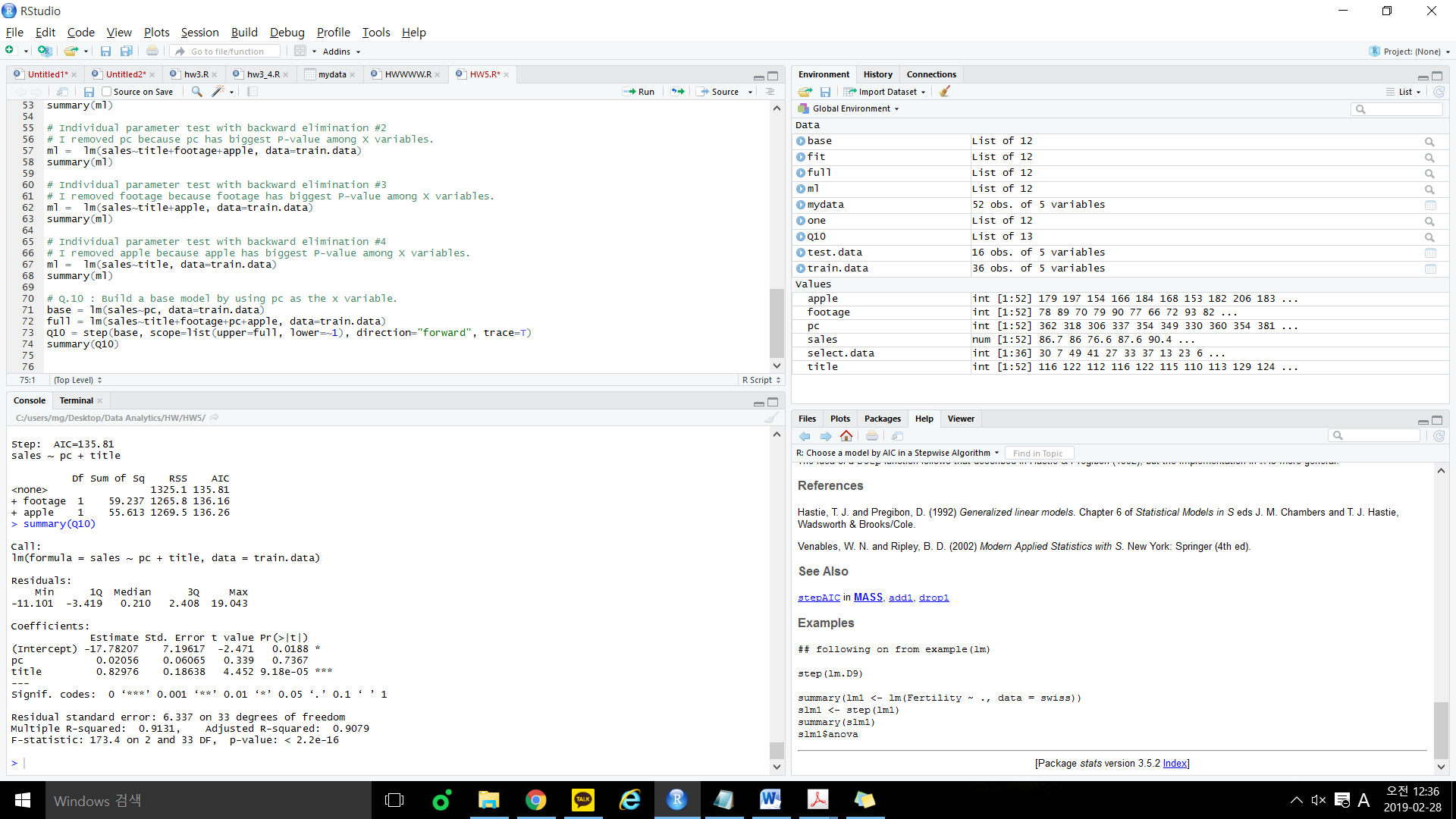
residual analysis 을 수행하고 residual analysis 후에 결론을 제공하십시오. 참고 :Normality test를 사용하려는 경우 Shapiro-Wilk 정규성 테스트를 사용하십시오

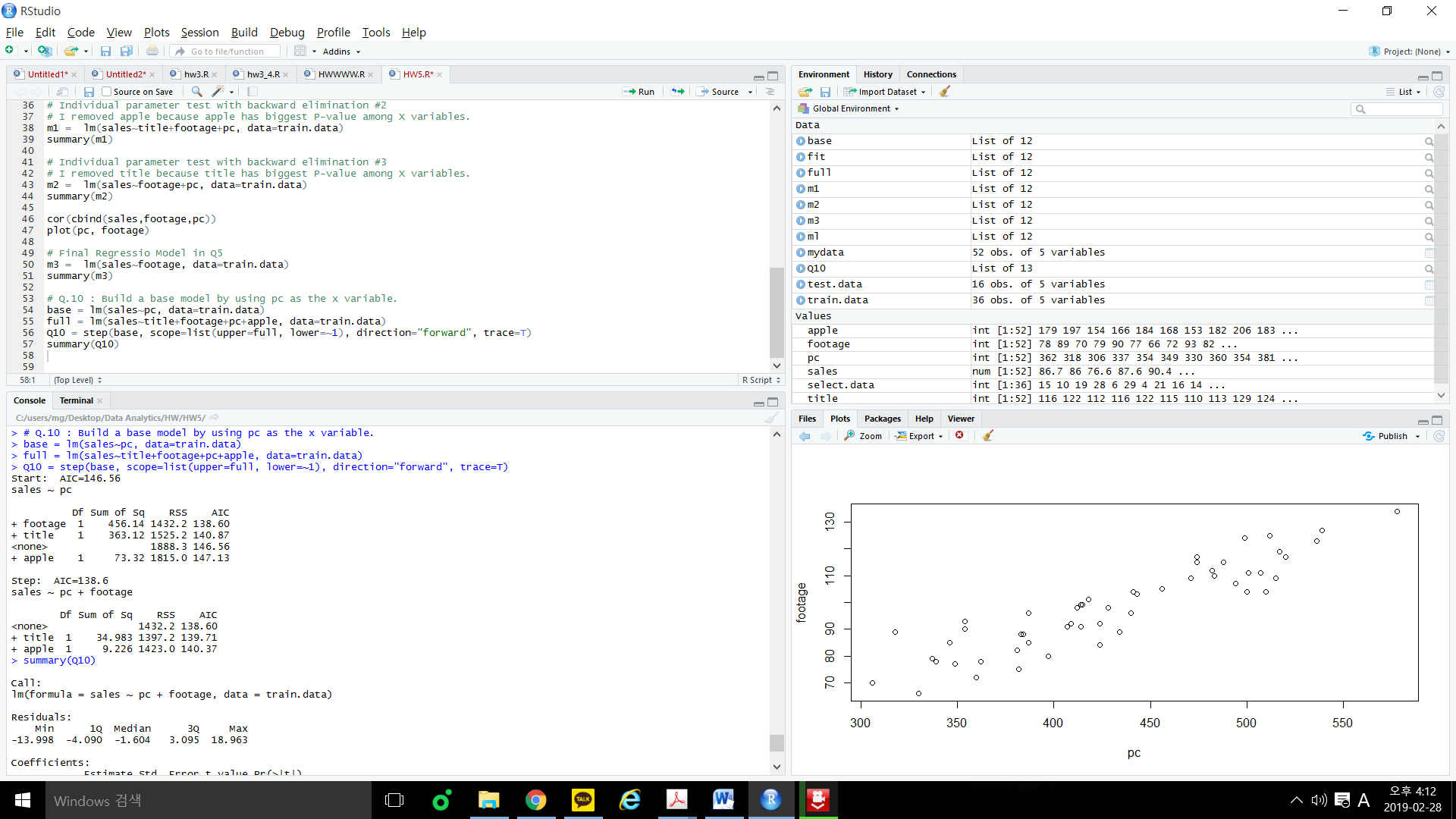
Goals in Residual Analysis

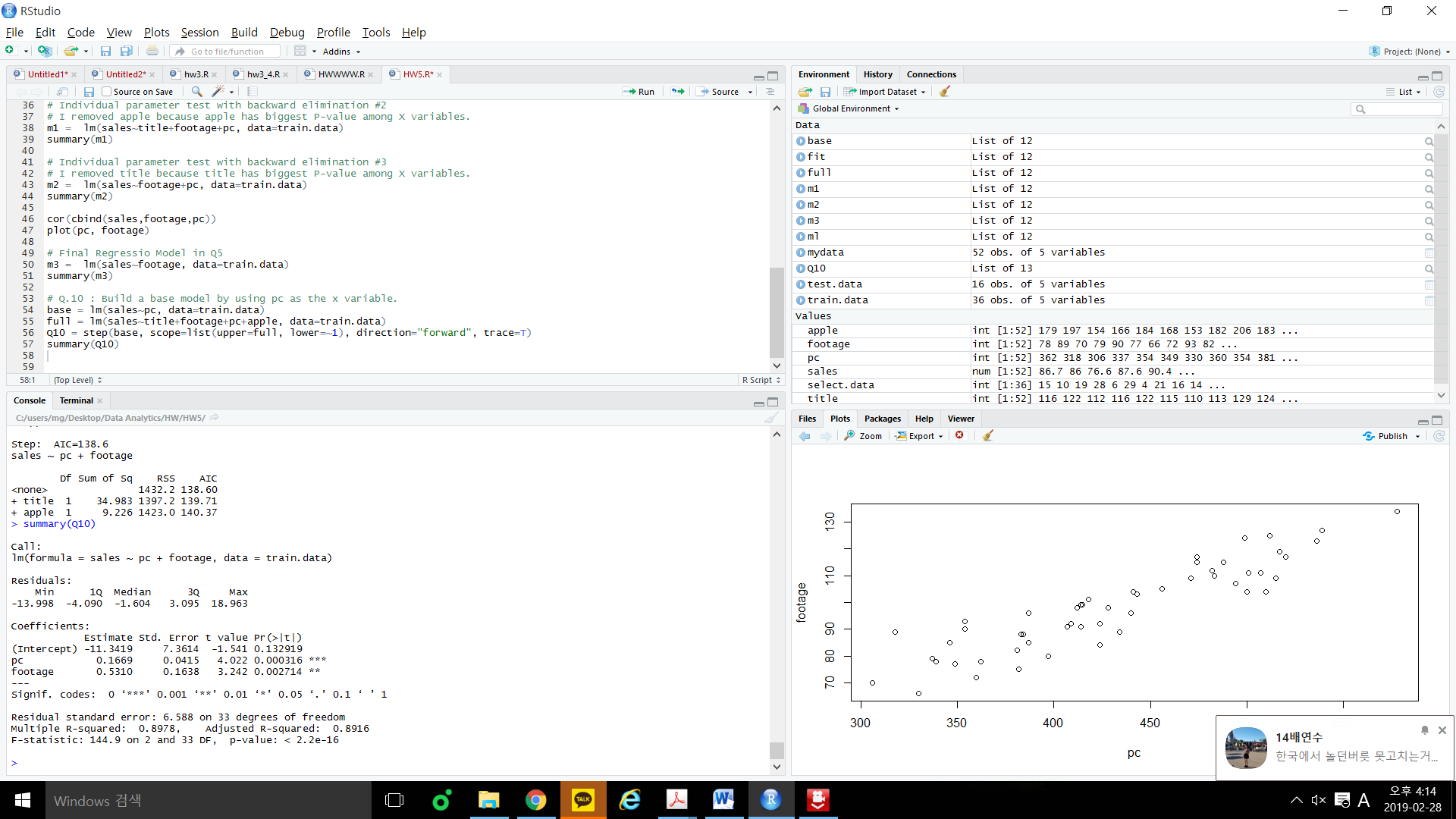
1. Validate the constant variance
   1. Plot residuals vs predicted values
2. Validate the linearity relationship
   1. Plot residuals vs each x-variable
3. Validate normal distribution of residuals
   1. Draw normal probability plot of residuals
4. Identify potential outliers

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.

pc를 x 변수로 사용하여 기본 모델을 작성하십시오. 그런 다음 forward selection을 수행하고 AIC / BIC를 메트릭으로 사용하여 모델을 작성하십시오. 모델이 5 단계에서 작성한 모델과 다른 경우 이 모델의 유효성을 검사합니다. (9의단계를 다시 거쳐야함)이 두 모델의 adj-R2를 비교하십시오.







Model in Q10) Model in Q5)

Adj- of model in 10) = 0.8916

Adj- of model in 5) = ?

+Residual analysis도 해야함

Conclusion : So we can infer that model in 5) is better model than model in 10).

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

5) 와 10)의 모델의 RMSE를 계산하고, 어떤 모델이 나은지 결론을 지으십시오.