**2018301034 컴퓨터과학과 서 용 주**

**6장 연습문제**

**[8번]**

① 처리간

② 처리내

③ 54

④ 3-1=2

⑤ 9-2=7

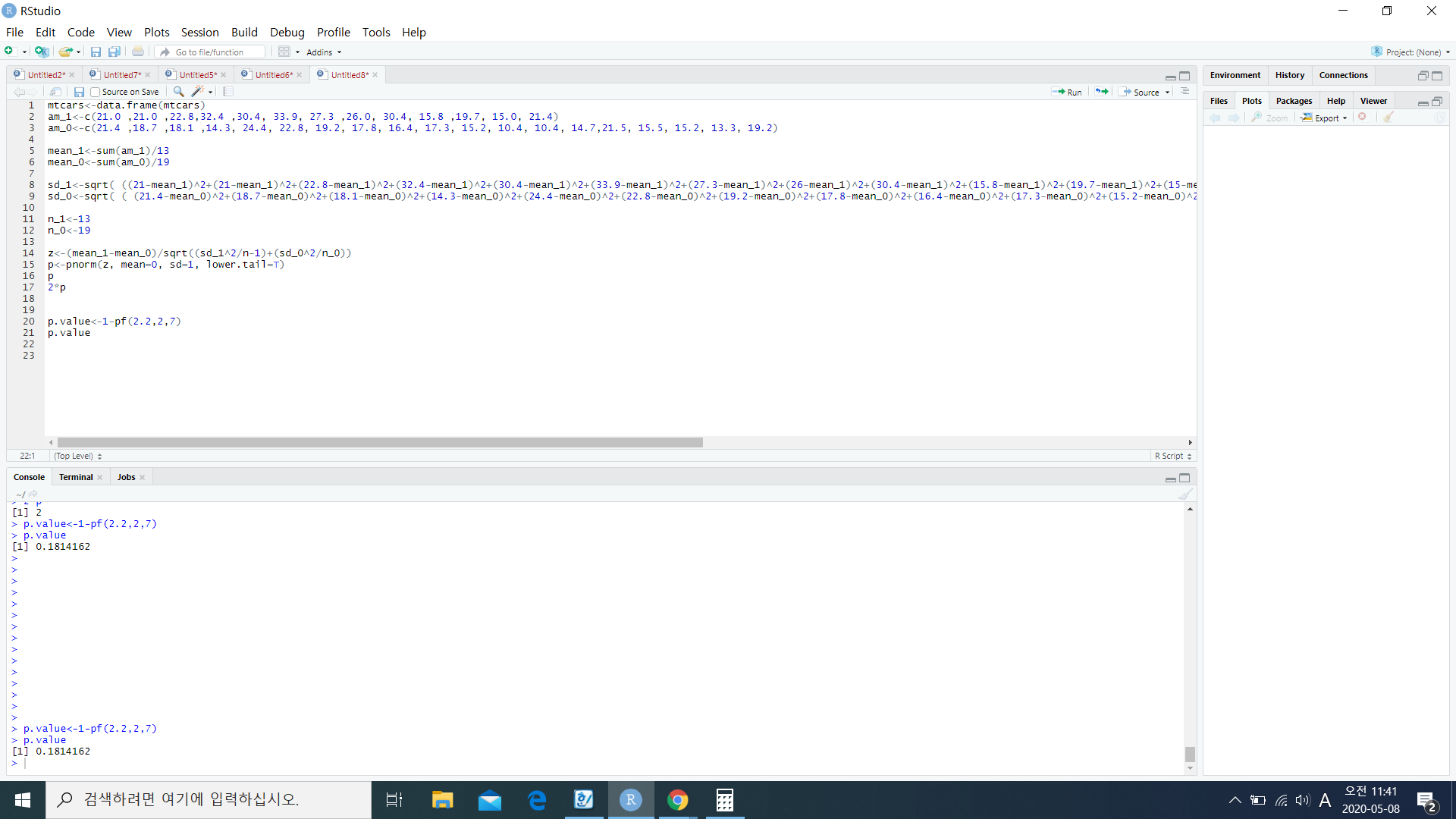
⑥ 10-1=9

⑦ 34/2=17

⑧ 54/7=7.714

⑨ 17/7.714=2.2

⑩ 0.18



자유도가 2,7인 F분포에서 서버의 응답시간에 차이가 날 확률은 0.18이므로 귀무가설을 채택한다. 즉 서버의 응답시간에 차이가 없다.

**[9번]**

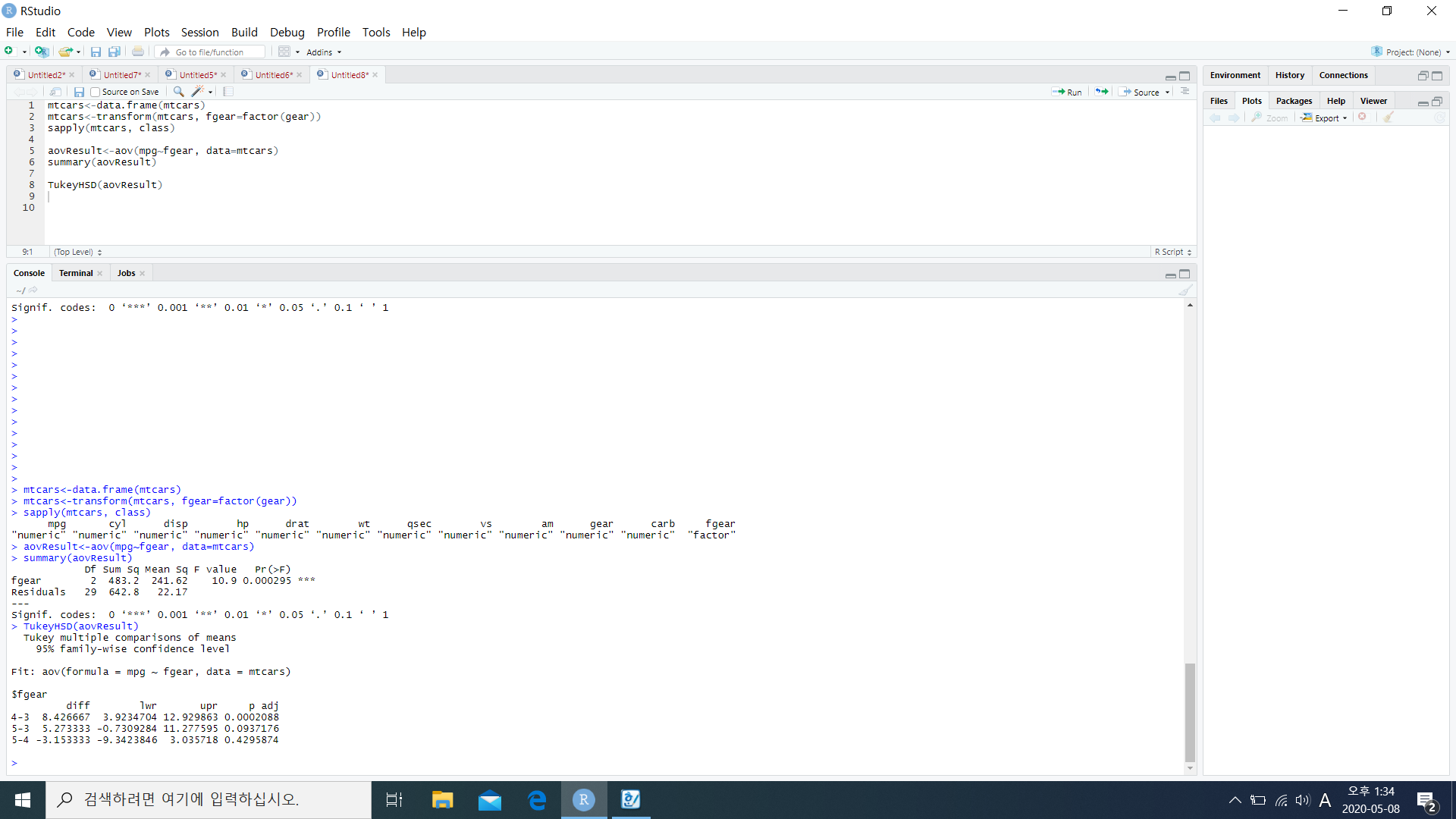
1) 귀무가설 : 기어의 개수는 연비에 영향을 미치지 못한다.

대립가설 : 기어의 개수는 연비에 영향을 미친다.

2) 유의수준은 0.05이다.

3) 검정통계량 : z

4) 기각역설정 : p<0.05

5) p값이 0.000295이므로 귀무가설을 기각한다. 즉 기어의 개수는 연비에 영향을 미친다.

**TukeyHSD()로 어떤 조건에서 차이를 보이는지 확인해보았다.**

**확인결과 4그룹과 3그룹간 차이가 있었다.**

**[10번]**

1) 귀무가설 : 실험조건은 빛의 속도에 영향을 미치지 못한다.

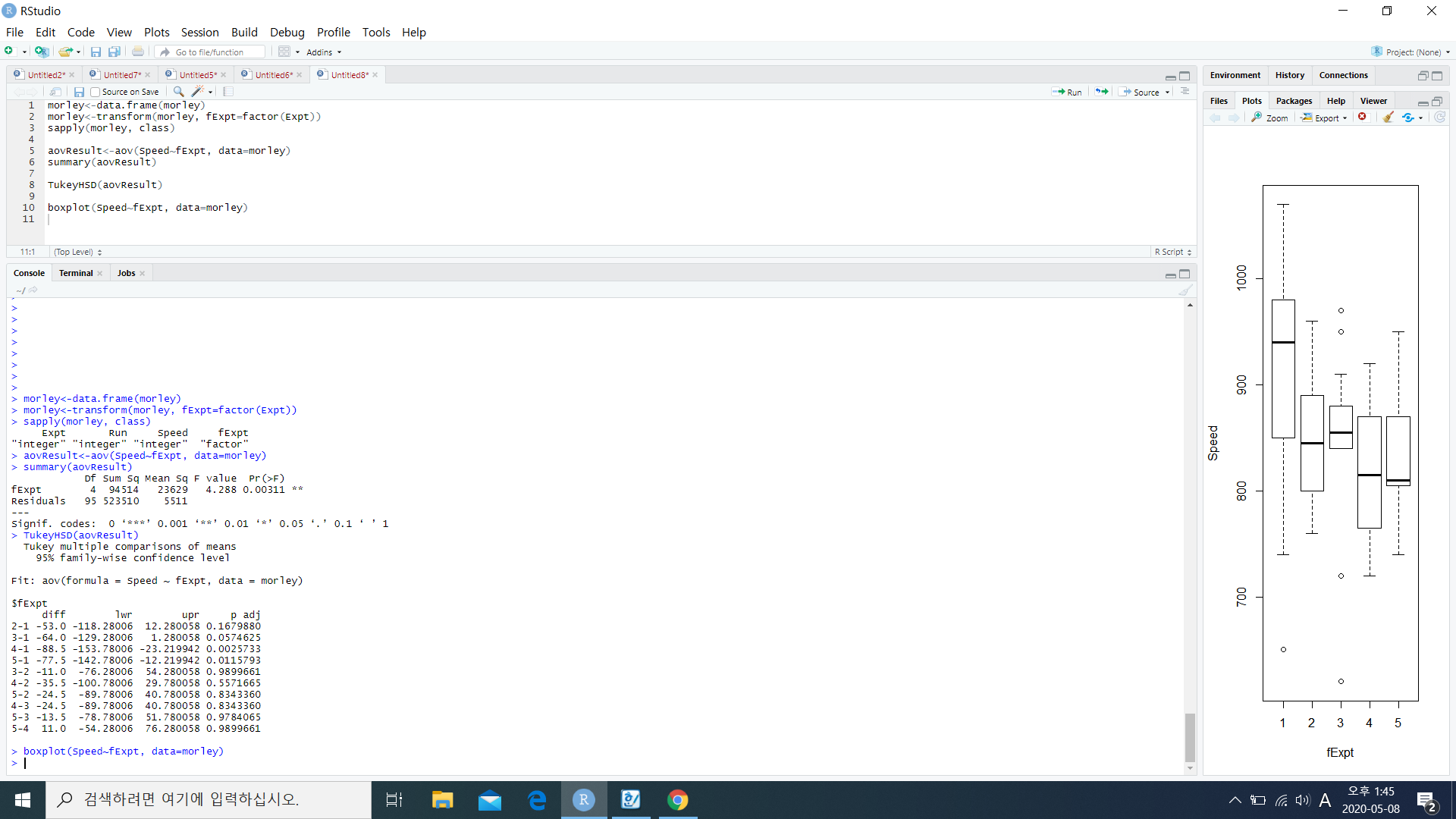
대립가설 : 실험조건은 빛의 속도에 영향을 미친다.

2) 유의수준은 0.05이다.

3) 검정통계량 : z

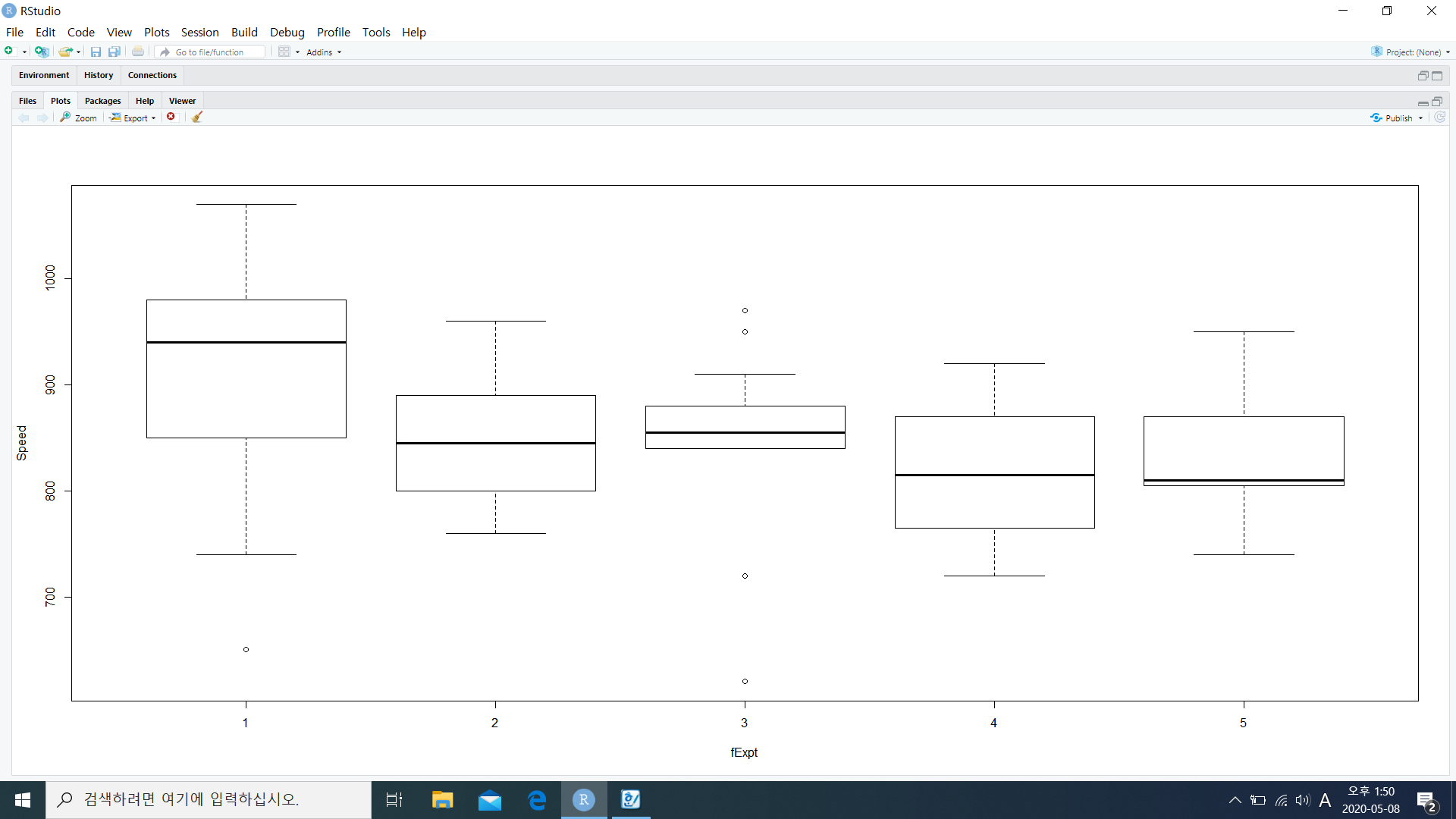
4) 기각역설정 : p<0.05

5) p값이 0.00311이므로 귀무가설을 기각한다. 즉 실험조건은 빛의 속도에 영향을 미친다.

TukeyHSD()를 사용하여 어느 조건에서 차이를 보이는지 살펴보았다.

4조건과 1조건, 5조건과 1조건 사이에 차이가 있다는 것을 p adj<0.05라는것을통해 알 수 있었다.

박스플롯을 그리면 아래와같다.



**[실습과제]**

1) 가설설정

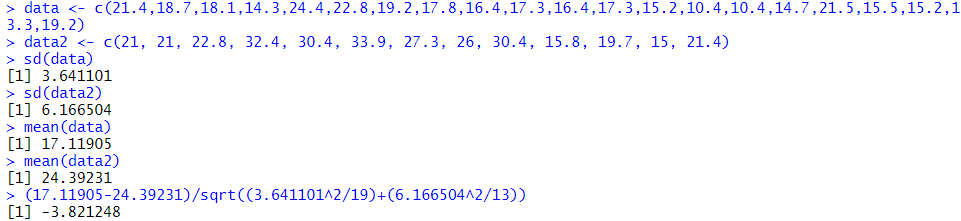
H0: u1-u2=0, Ha: u1-u2≠0

2) 유의수준결정

a=0.05

3) 검정통계량

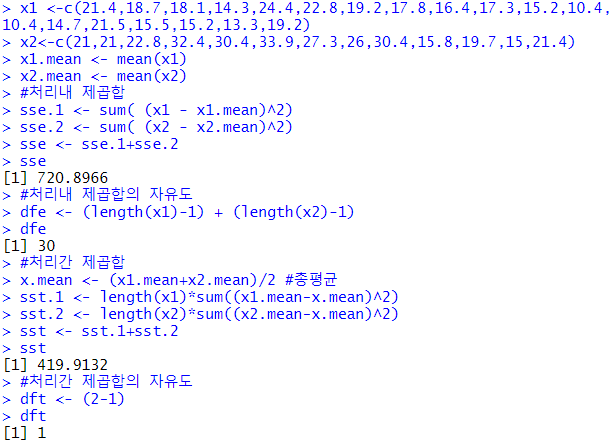
4) p-값 계산

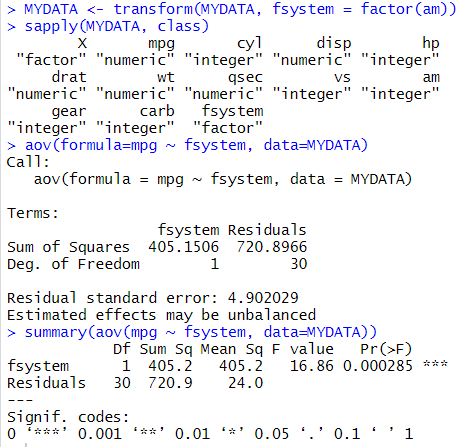


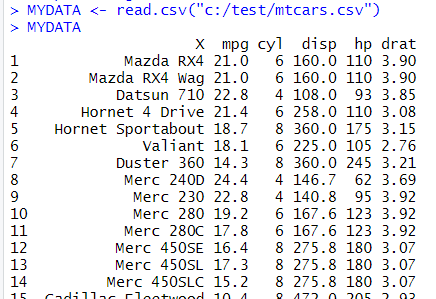


5) p-값

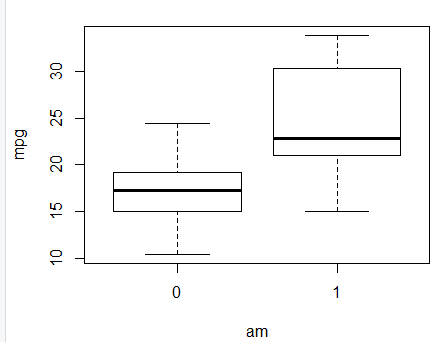
p\*2=0.001250742 > 0.05이므로 귀무가설을 기각 할 수 있다. 따라서 mtcars 데이터셋에서 변속기 종류에 따라 연비가 달라진다고 할 수 있다.

2)





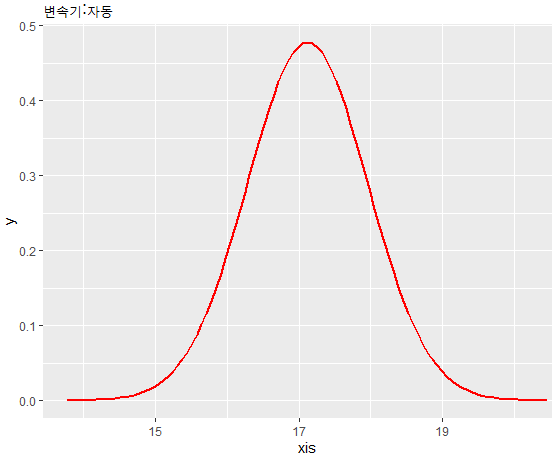
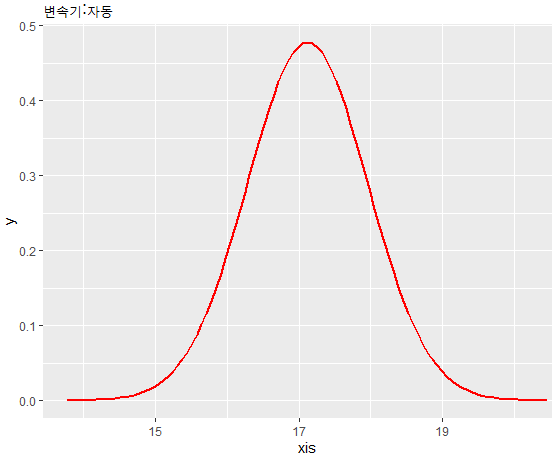
p값이 0.05보다 작기 때문에 기각할 수 있다, 기각의 유무가 1번과 같다.



3)



4)



**7장 연습문제**

**[1번]**

오차에 대한 가정 :

1) 독립변수와 오차항 간 상관관계가 없어야한다.

2) 오차항의 기댓값은 0이어야한다.

3) 오차항의 분산은 모든 관찰치에서 일정해야한다.

4) 오차항간 상과관계가 없어야한다.

5) 오차항은 정규분포를 따라야한다.

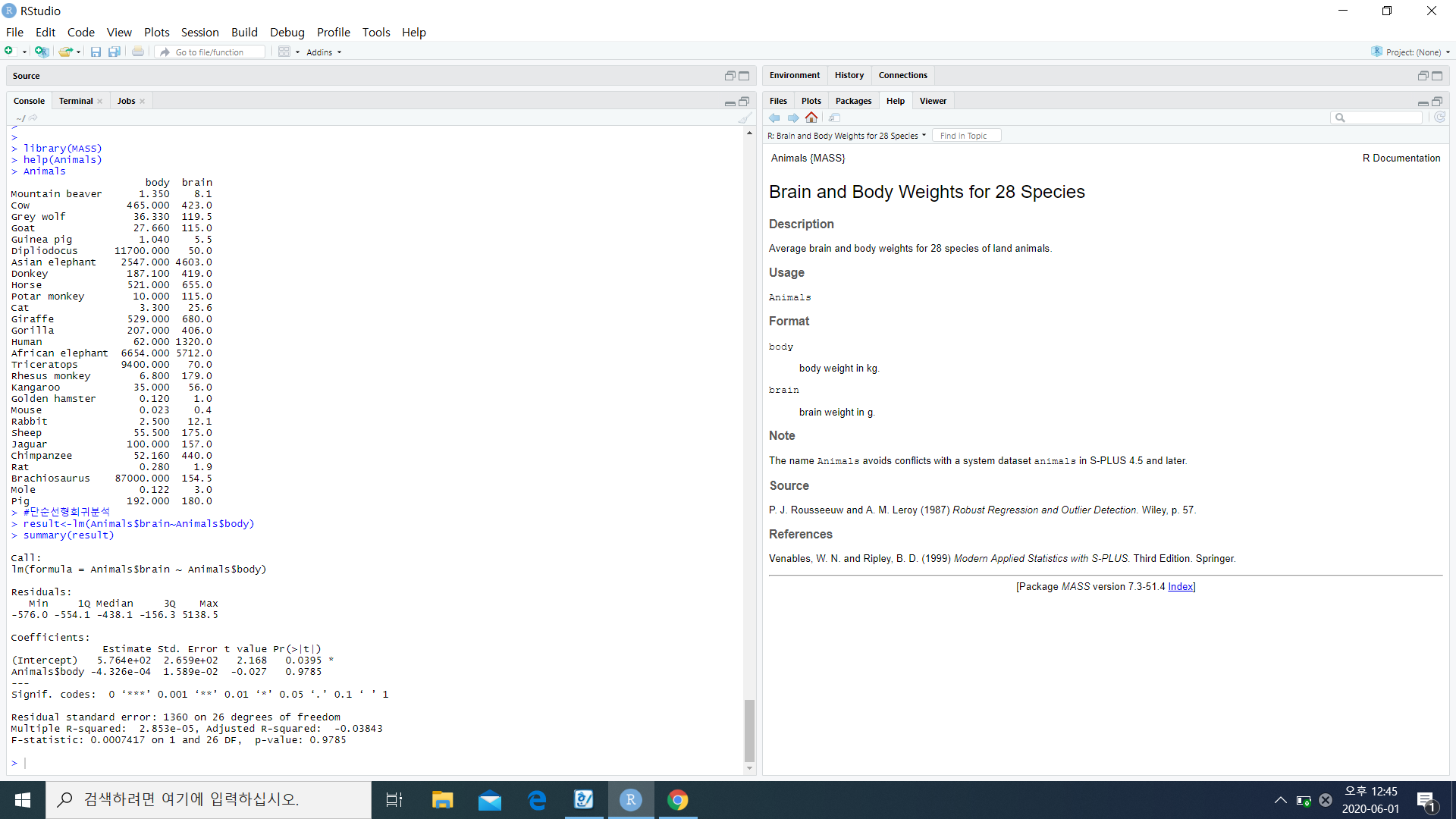
오차가 적절한지 검토하는 방법으로는

1) 회귀분석을 시행하기 전에 산포도를 그려보고 상관분석을 시행해본다.

2) 표준화 잔차들에 대한 히스토그램과 p-plot을 통해서 정규성을 검토한다.

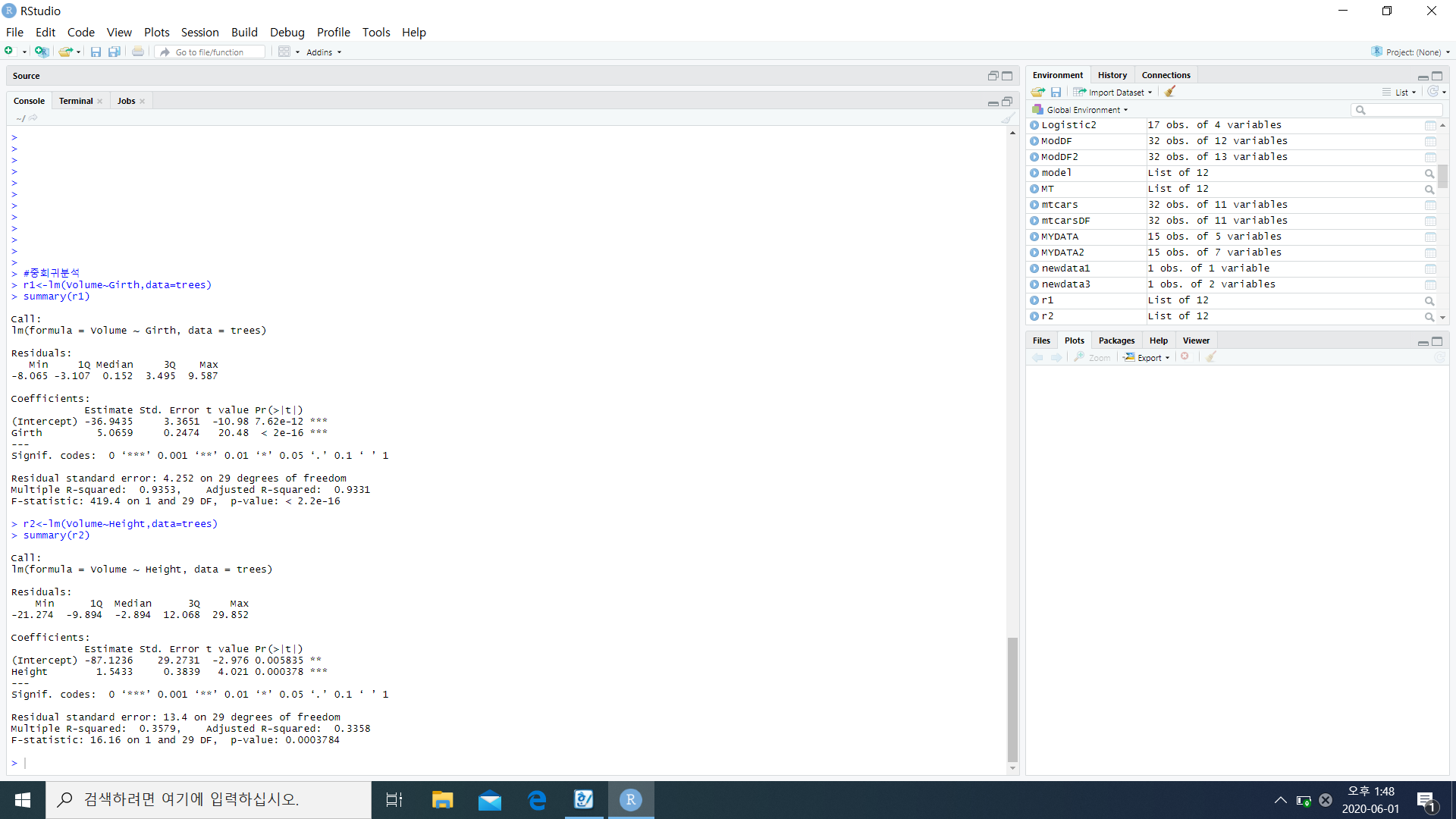
3) 선형성에대한 검토와 동일한 과정을 거쳐 plot의 점들이 흩어진 형태가 두 축 사이에 특정 패턴(깔때기 모양)을 보이지 않음을 보아 등분산성 가정을 검토할 수 있다.

**[5번]**

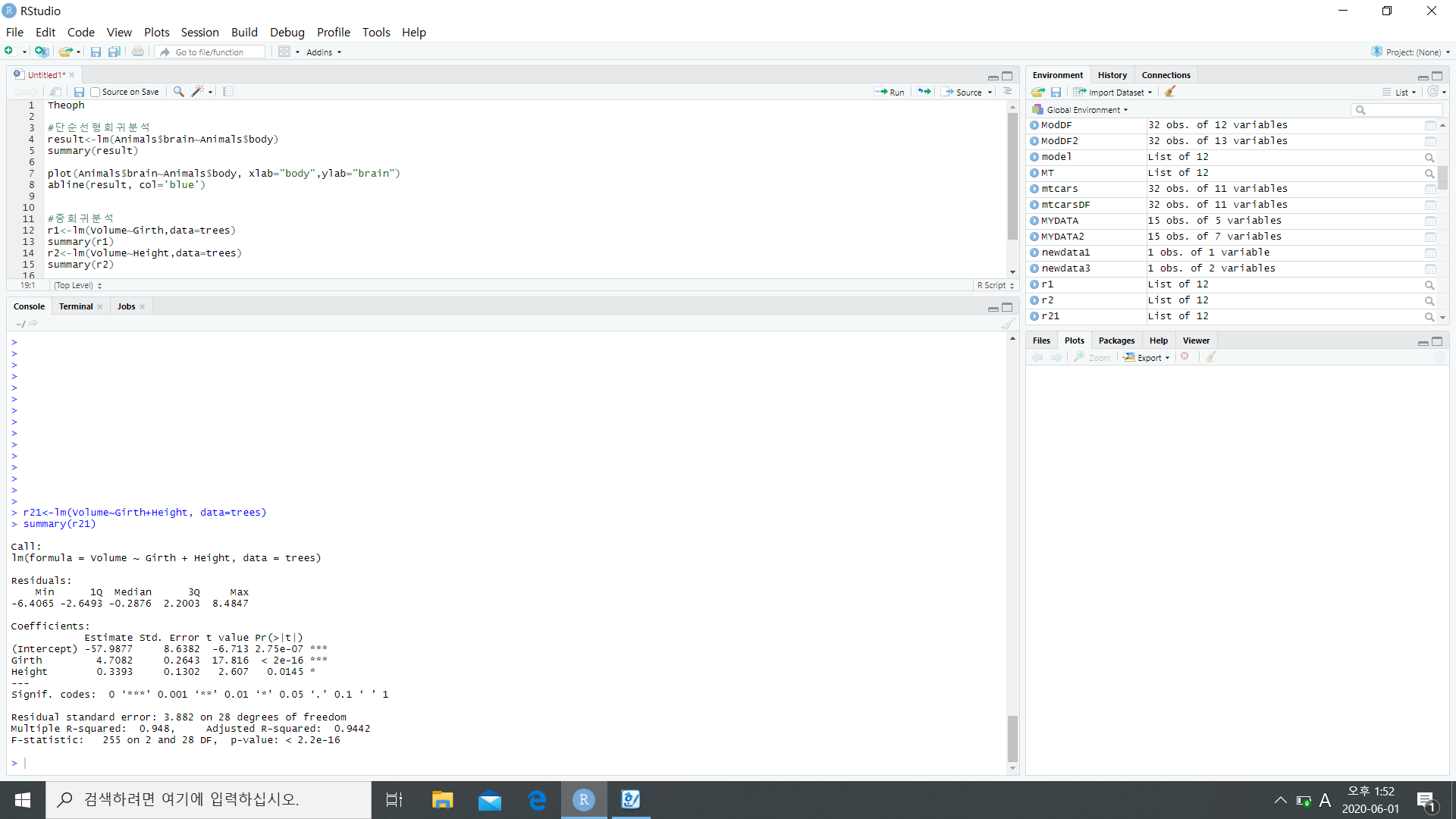
 **b1의 값이 -0.027로 0이 아니므로 귀무가설을 기각한다. 즉 선형관계가 존재한다. 맨 우측 하단의 p-value가 0.9785로 0보다 훨씬 큰 값이므로 귀무가설을 채택한다. 즉 단순선형회귀모형은 타당하지 않다.**

**[7번]**

**첫 번째 변수 선택 :** 이 단계에서 가장 유의한 변수는 P-값이 가장 작은 Girth이다.



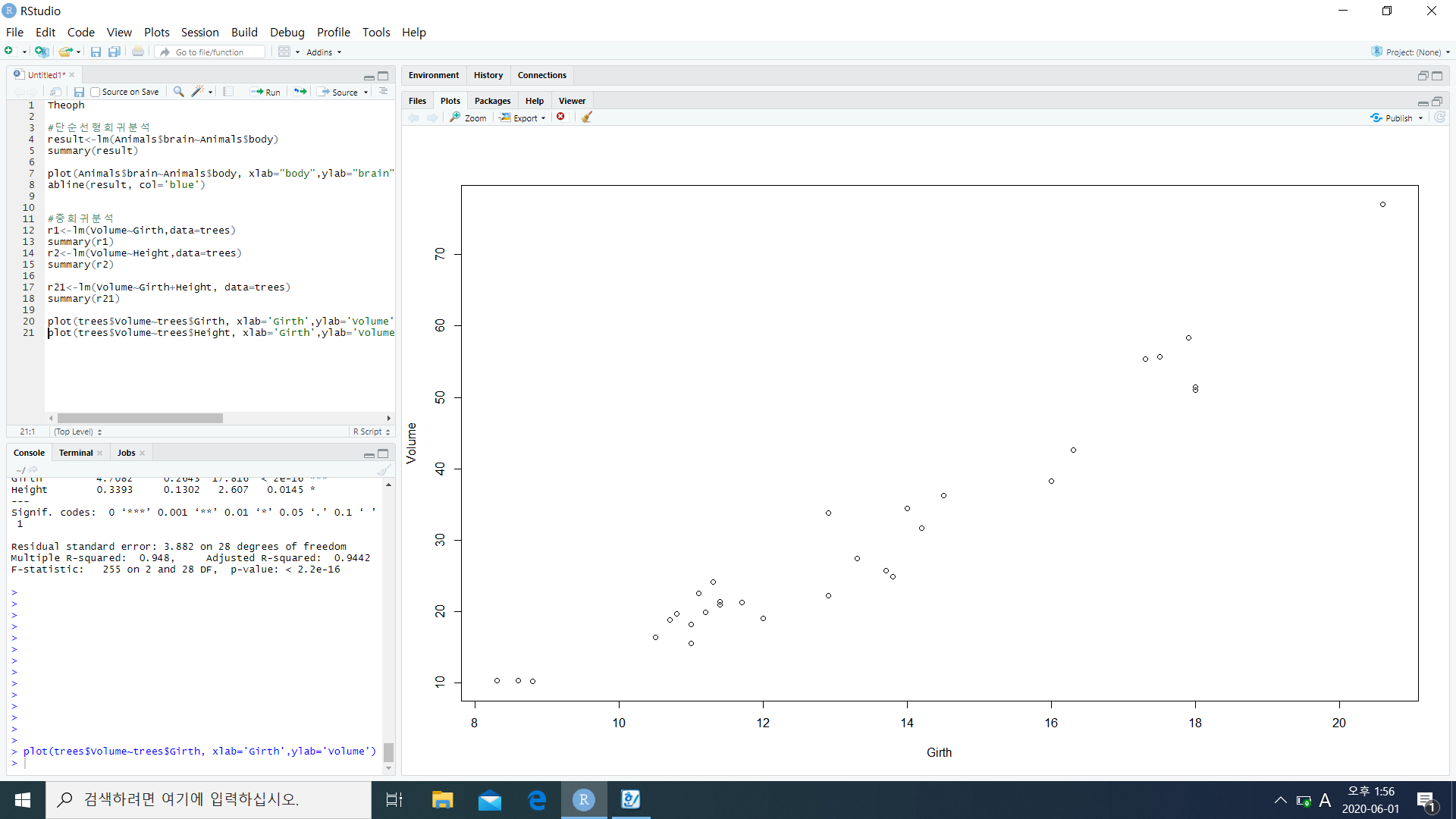
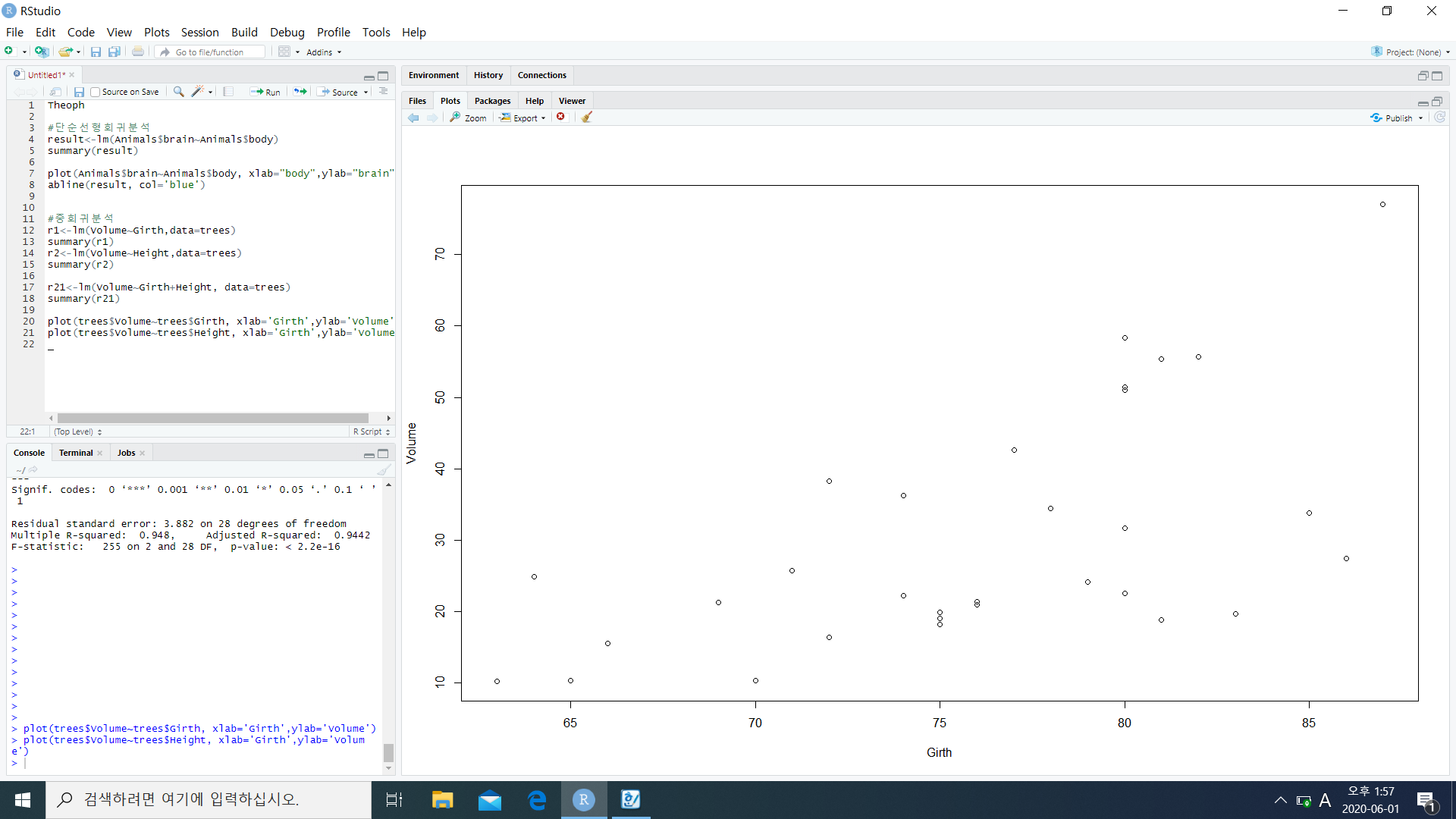
**두 번째 변수 선택**

두 번째 변수로 Height가 선택되었다.

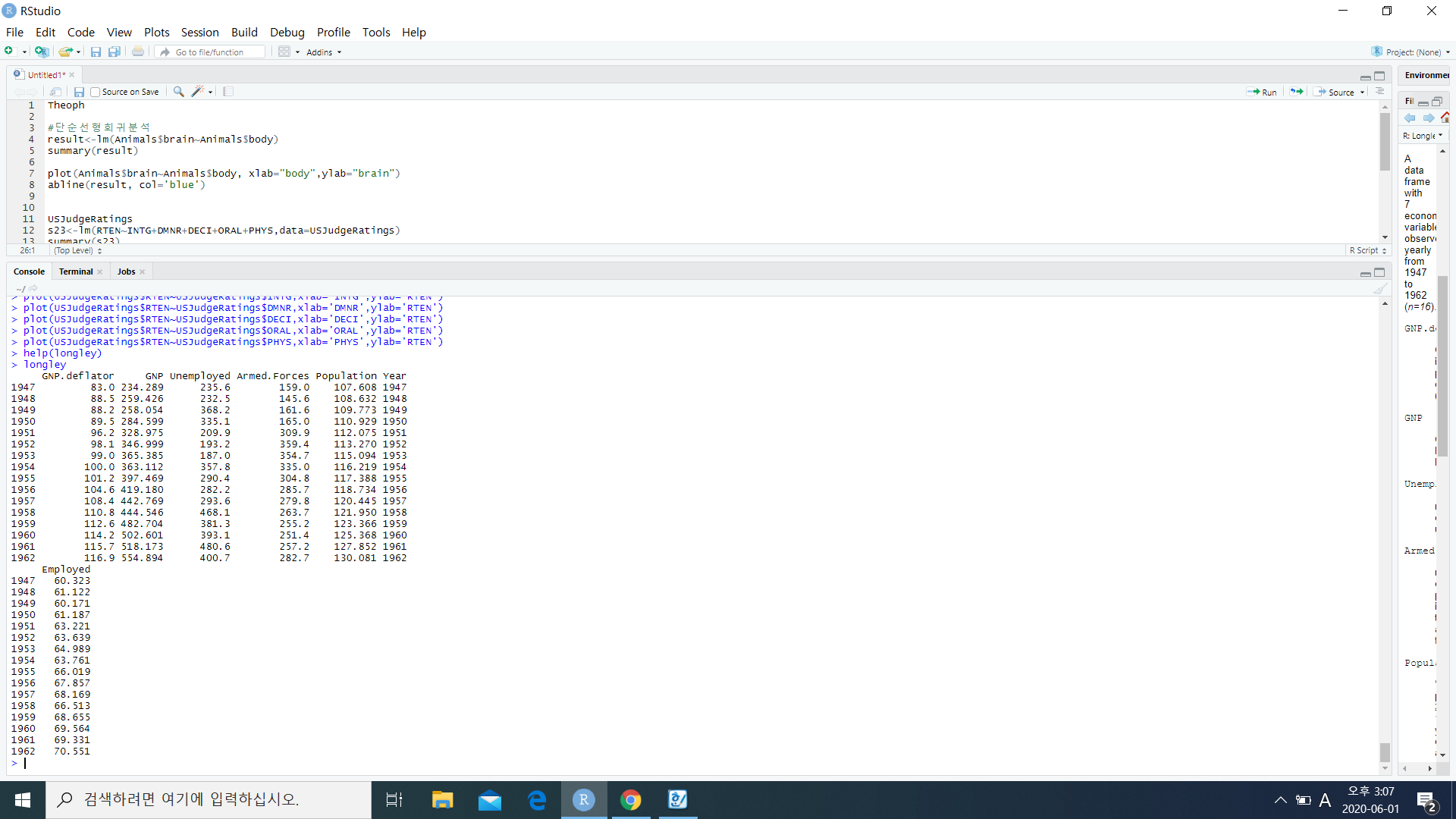
따라서 최적모형은 Volume=-57.9877+4.7082\*Girth+0.3393\*Height이다.

모형의 F값이 255이고 p-값이 2.2e-16보다 작은 값이므로 매우 유의하다.

유의한 독립변수별 종속변수의 개별 산점도를 작성하면 아래 두 가지가 나온다.



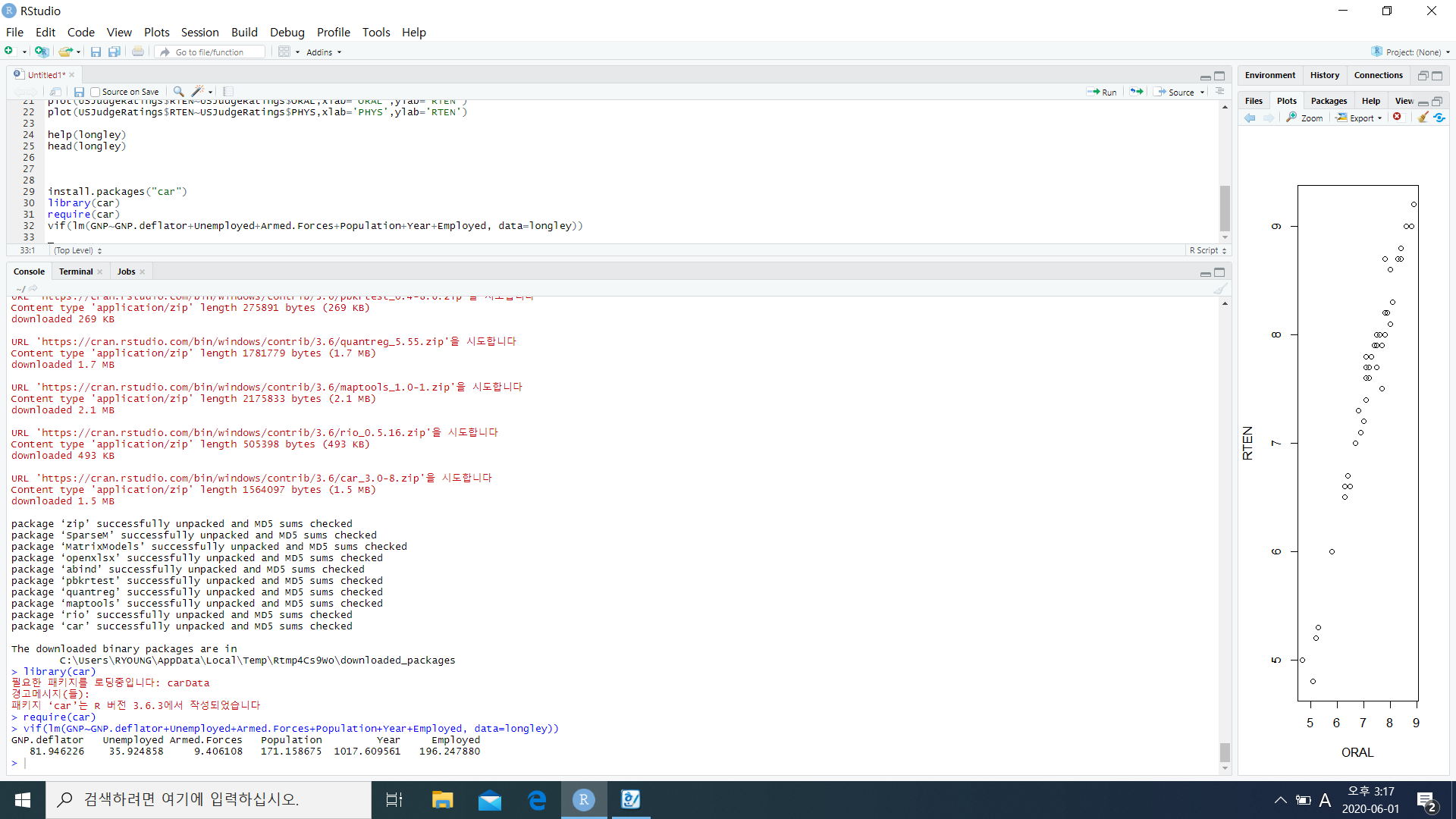
[9번]



longely는 GNP.deflator ,GNP Unemployed, Armed.Forces ,Population, Year, Employed 변수로 구성된 1947년부터 1962년까지의 데이터를 보유한 시계열데이터자료이다.

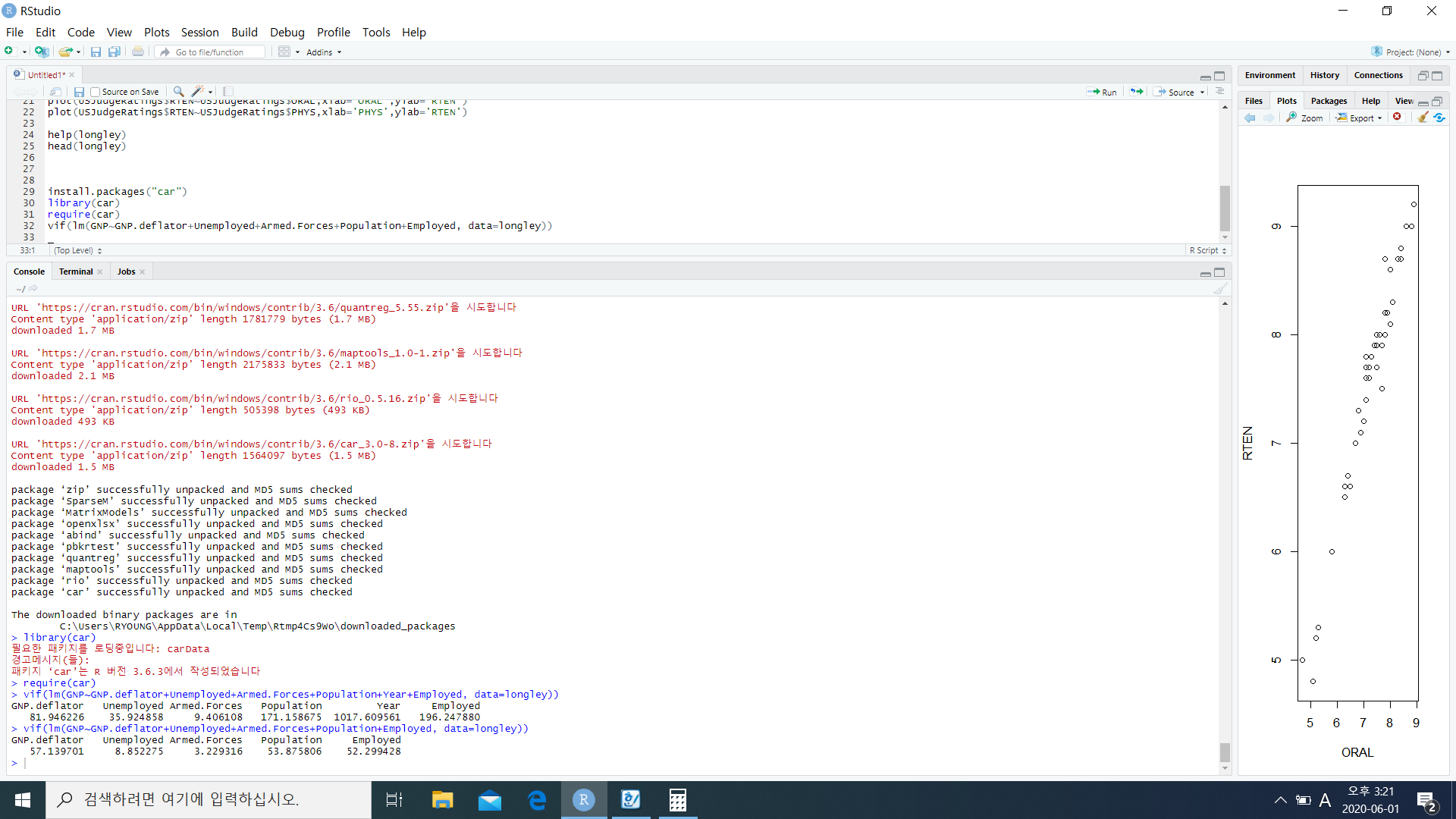
상관관계가 높은 데이터 셋이므로 다중공선성을 먼저 검토한다.

**1) 다중공선성이 있는 변수를 찾는다.**



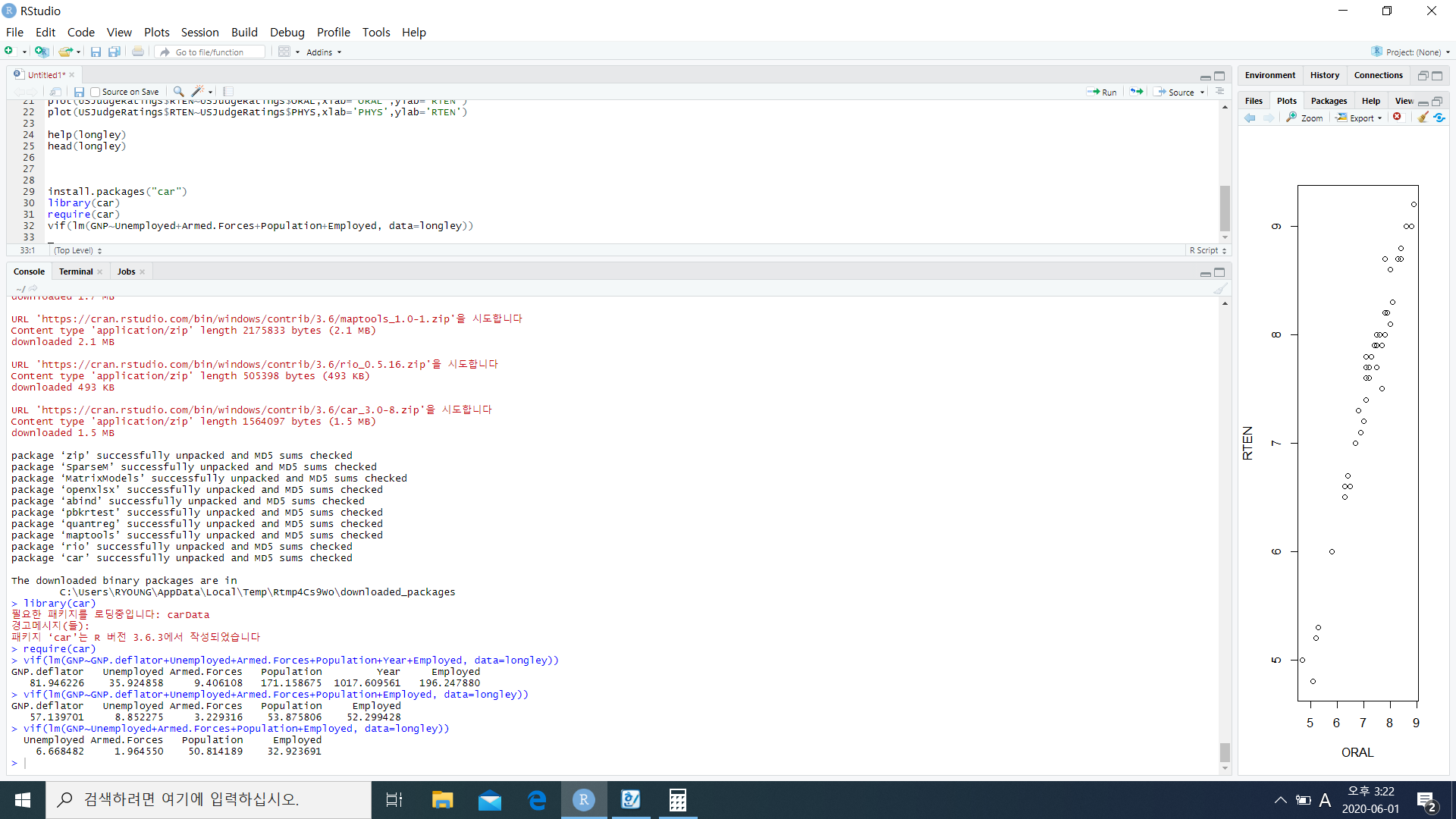
Year의 다중공선성이 가장 높고 설명비율이 1016.609561/1017.609561로 거의 100% 이므로 Year는 제거

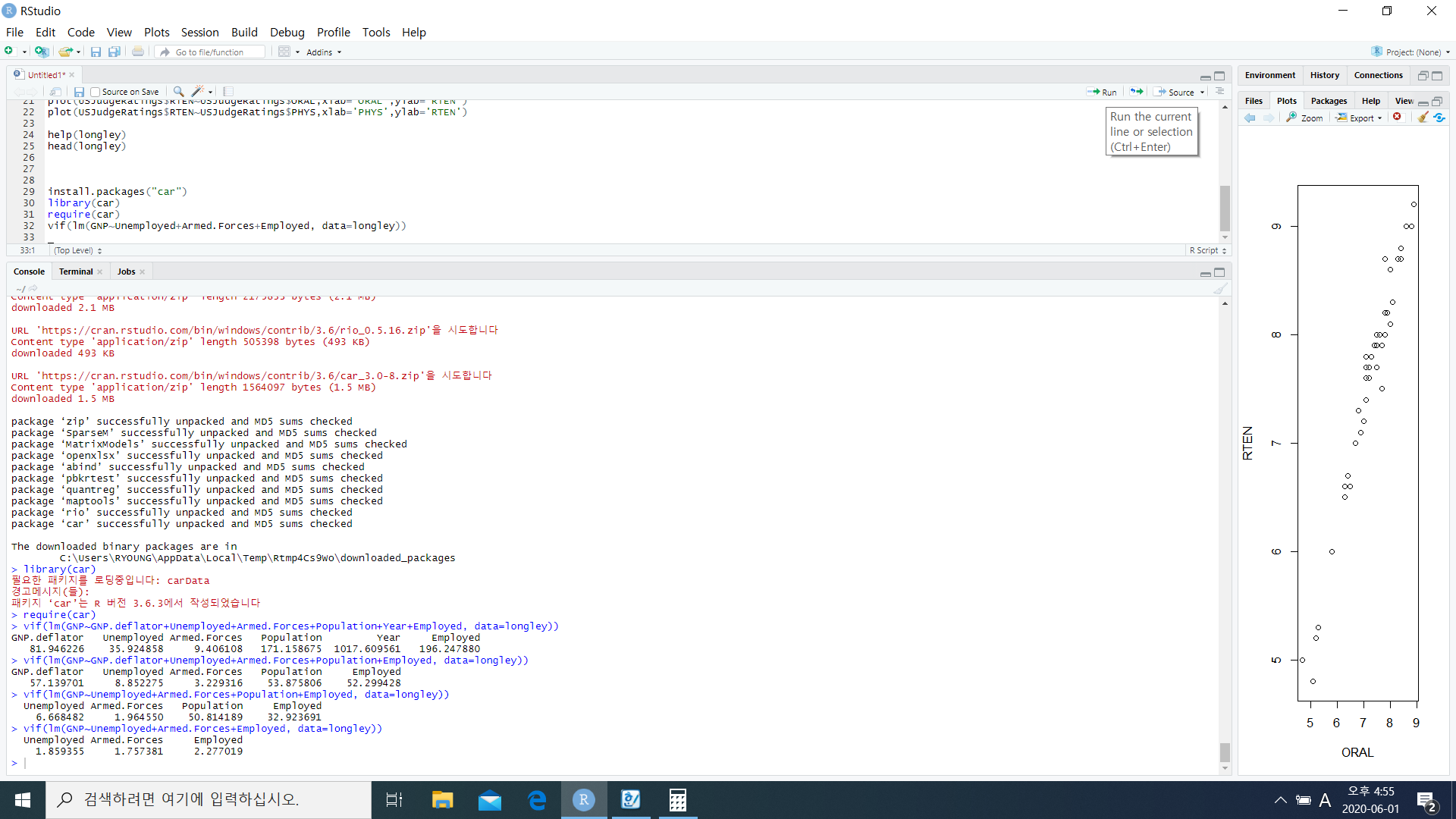
**2) Year 제거 후 vif()구동**



다음으로 GNP.deflator가 다중공선성이 높고 설명비율이 56.139701/57.139701=0.98이므로 제거.

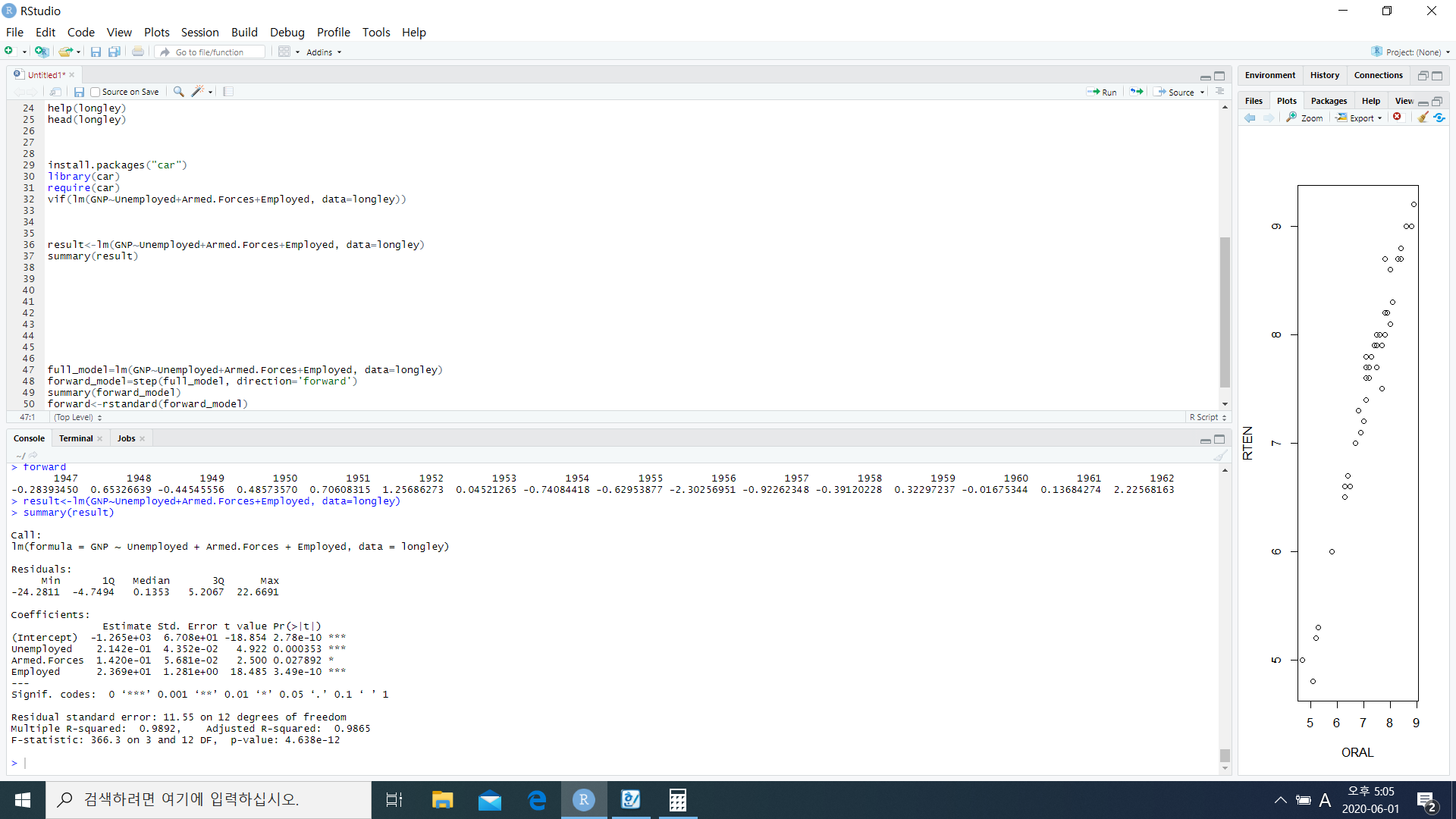
**3) GNP.deflator 제거 후 vif()구동**

다음으로 Population의 다중공선성이 높고 설명비율이 49.814189/50.814189=0.98 이므로 제거.



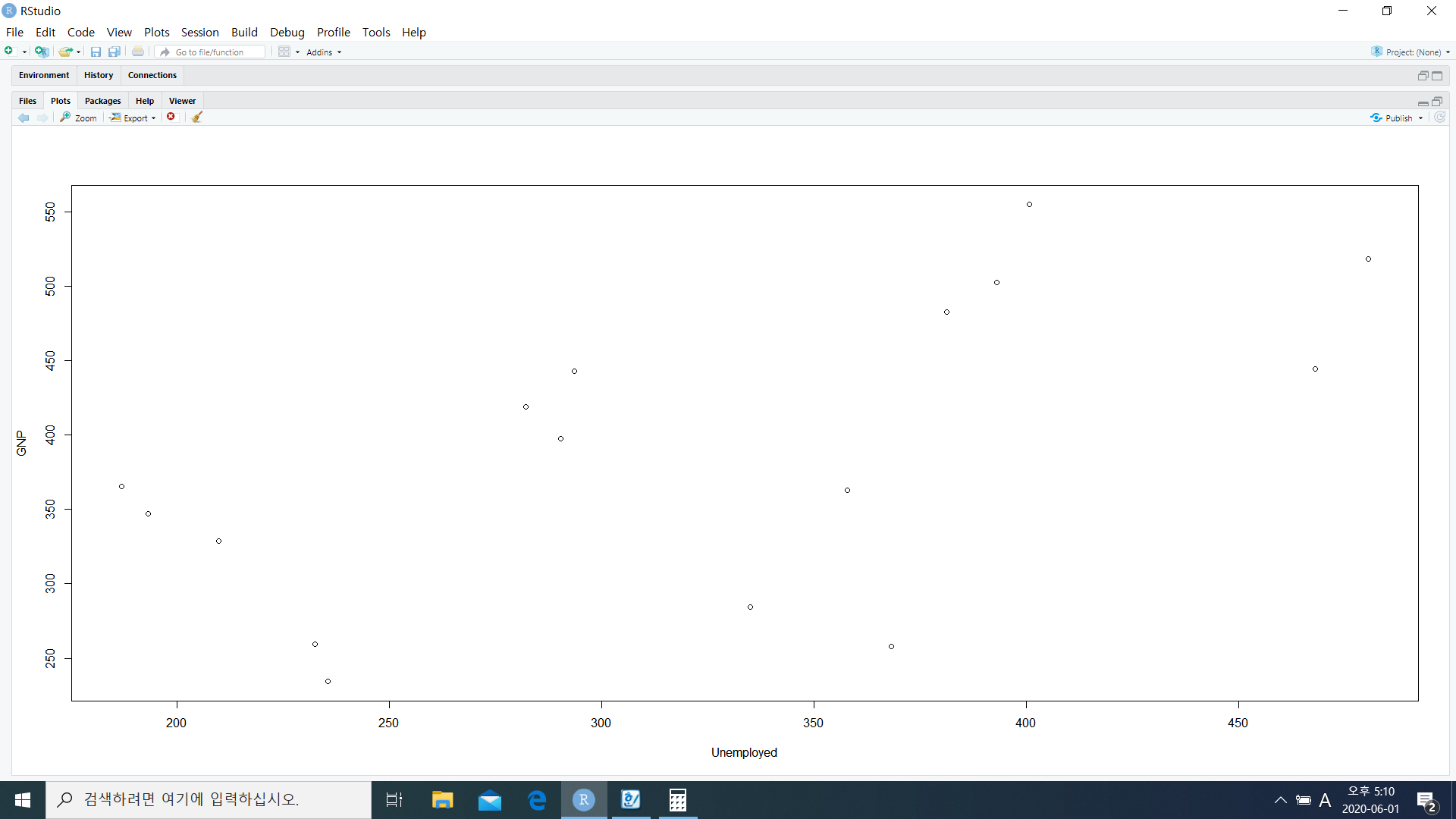
**vif값이 4이상인 변수가 없으므로 Unemployed, Armed.Forces, Employed가 최적의 독립변수들이다.**

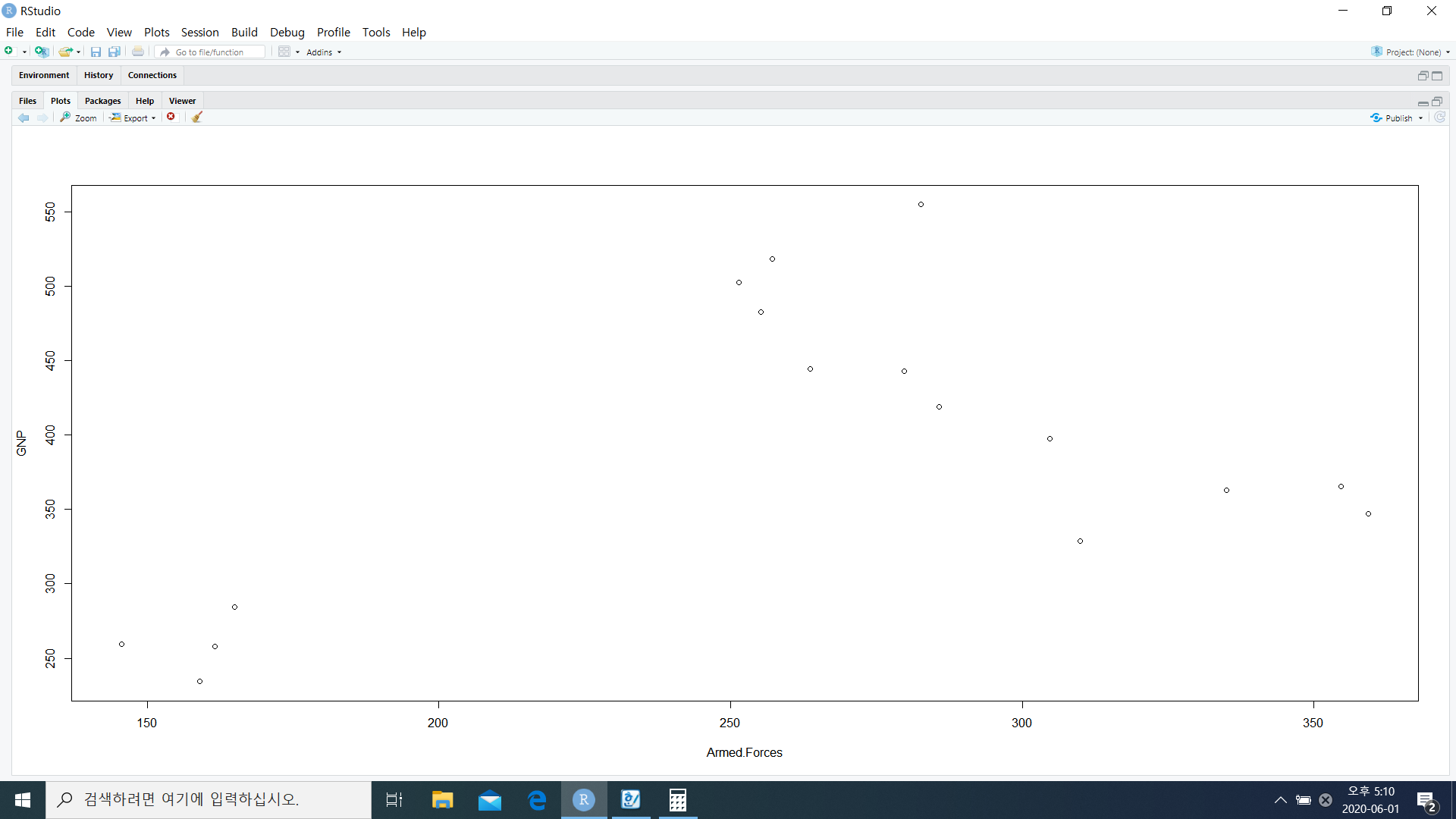
**선형관계의 타당성**

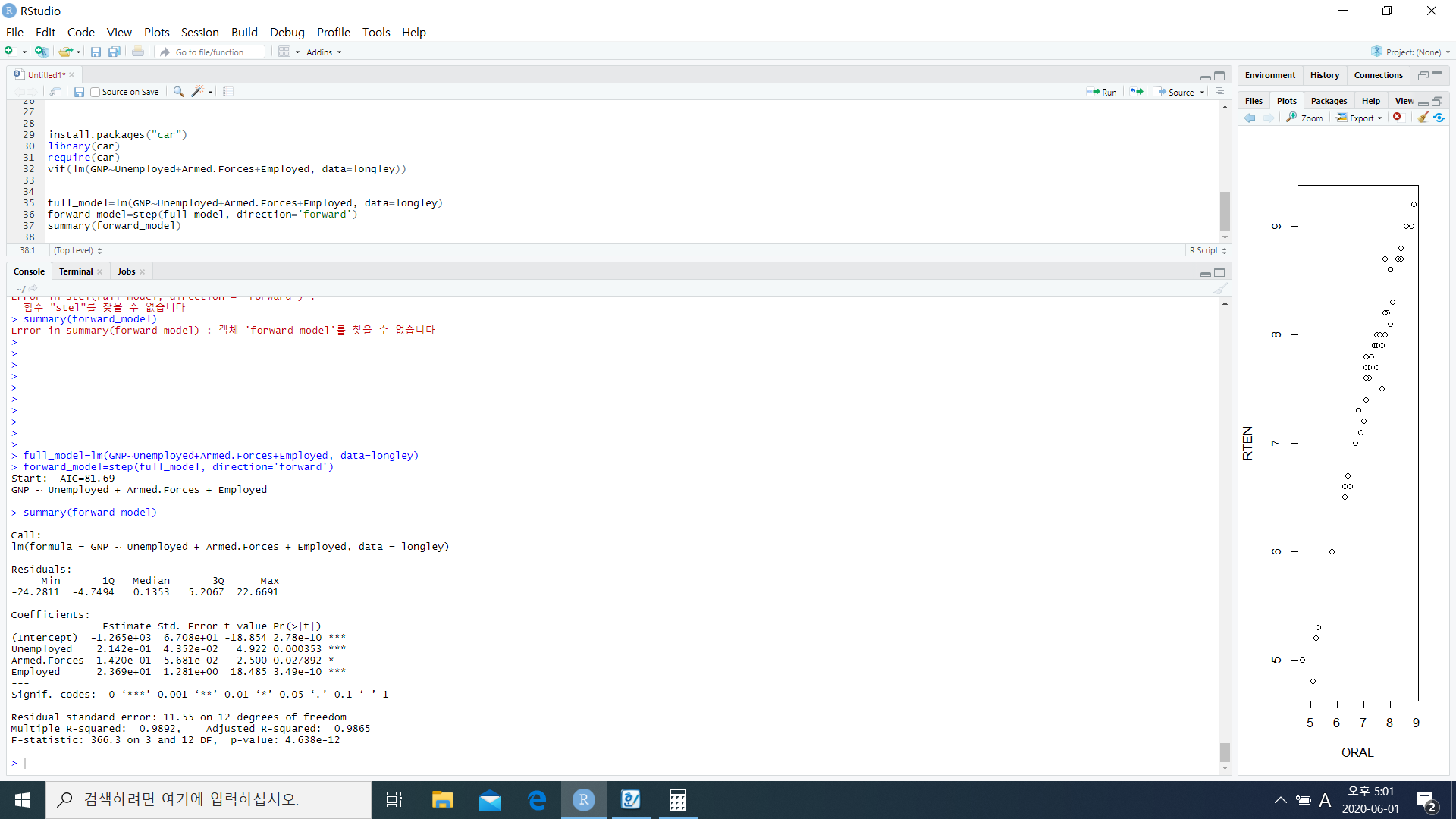


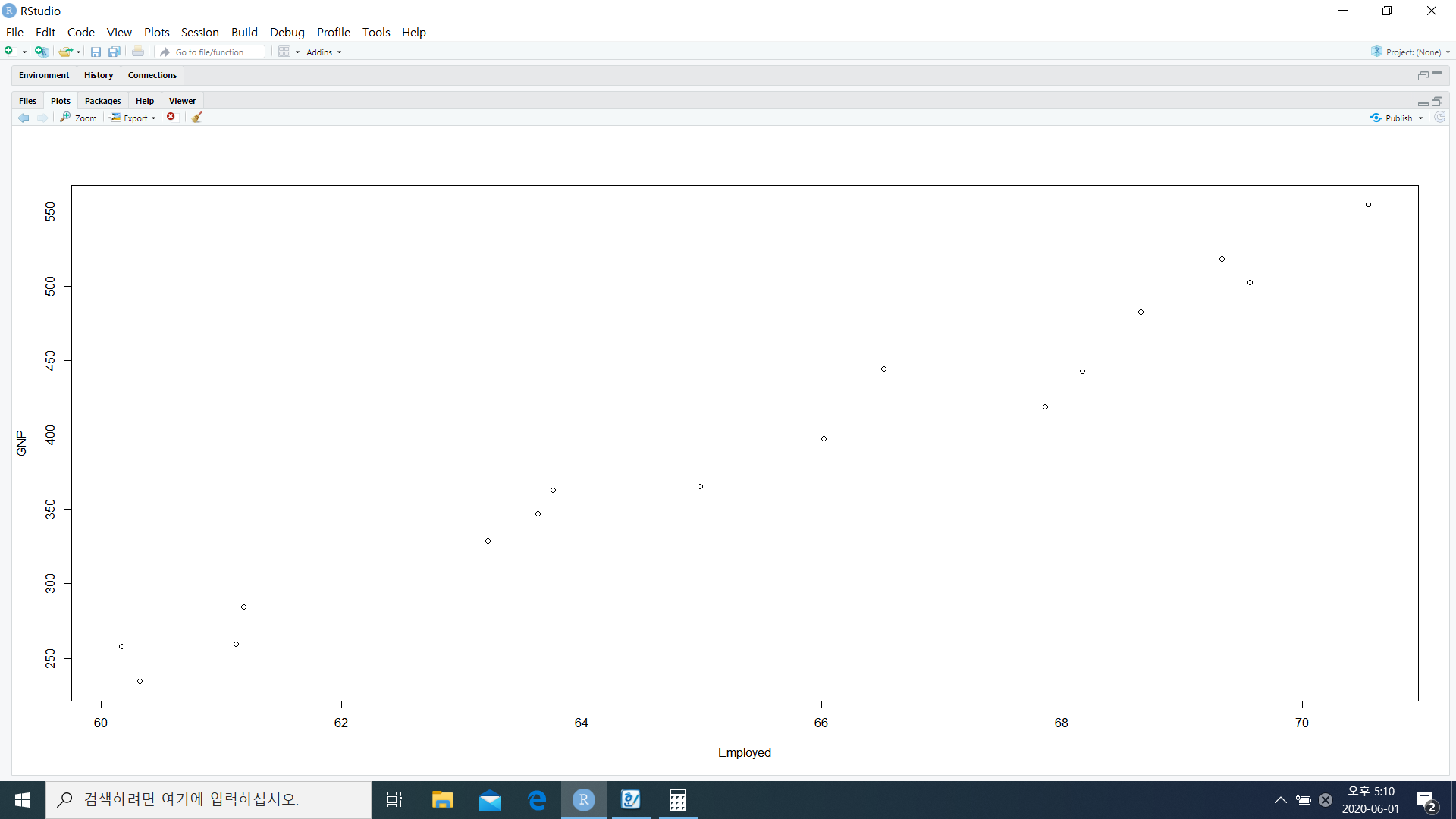
p-value가 4.638e-12로 0에 가까운 아주 작은 값을 가지므로 선형관계에 타당성이 있음을 알 수 있다.

유의한 독립변수별 종속변수의 개별 산점도를 작성하면 다음과같다.





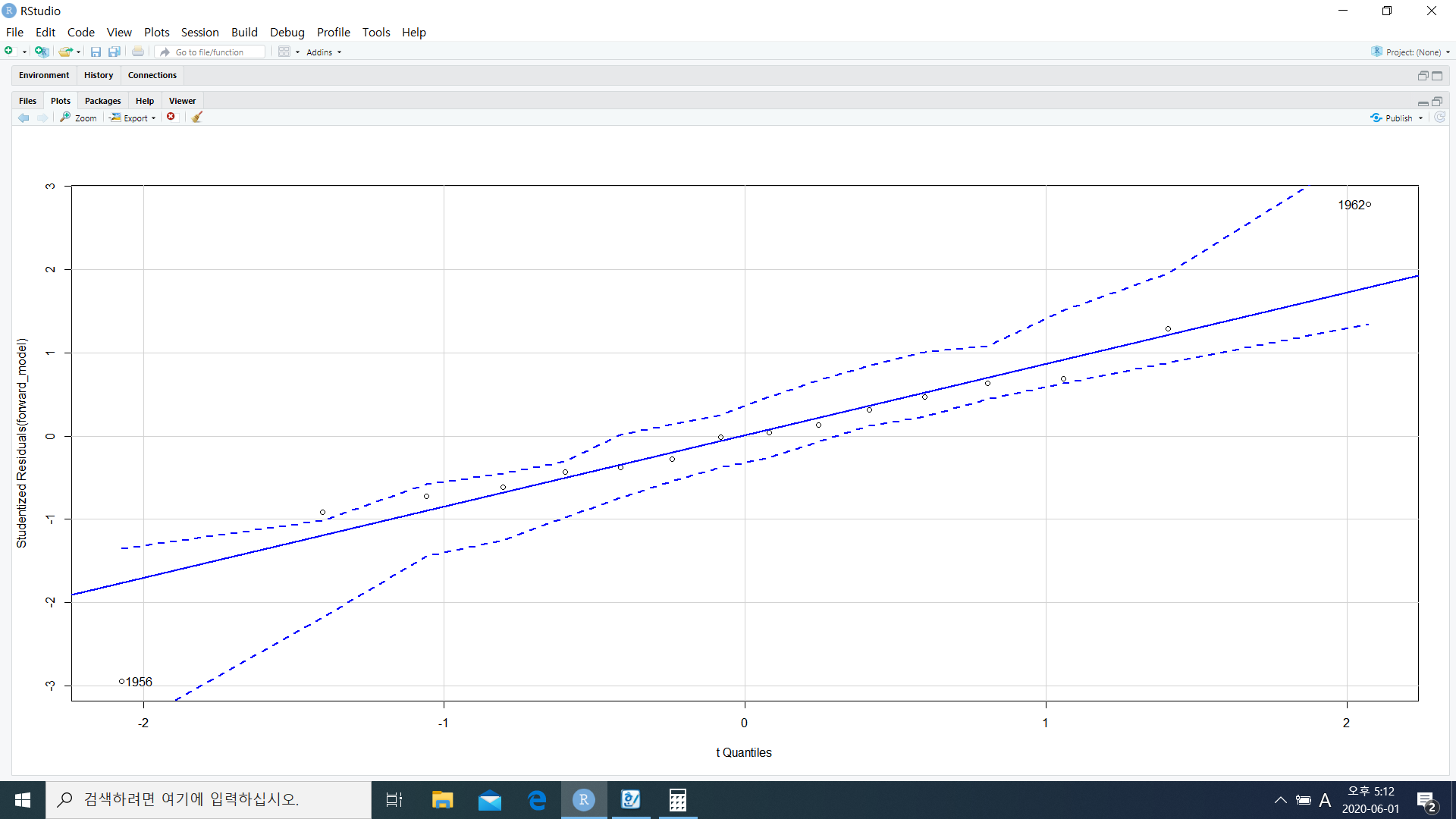




car에서 단계별 전방추가법을 실시한다.

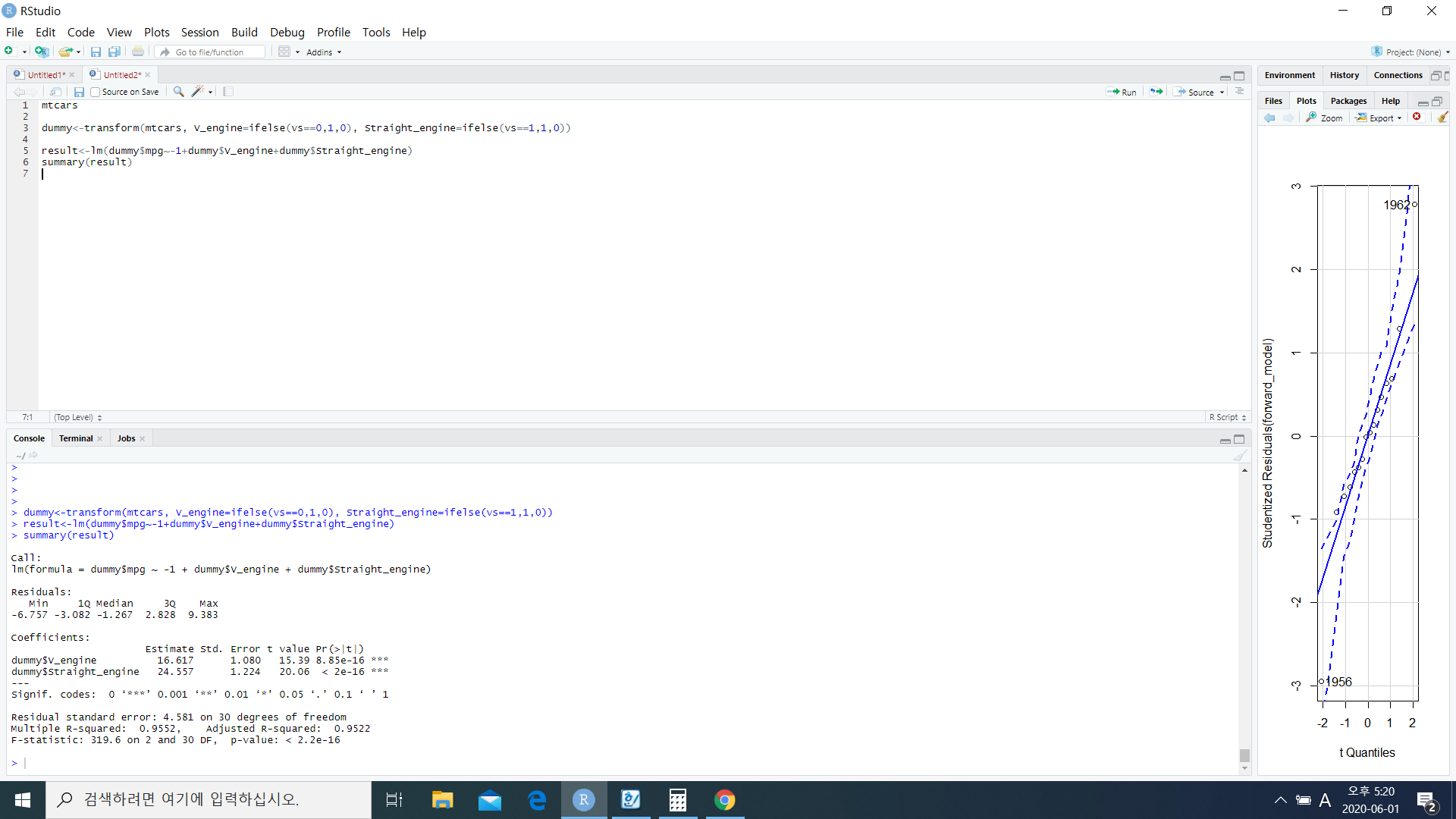
전방추가법에 의한 결과는 모든 변수를 포함한다.

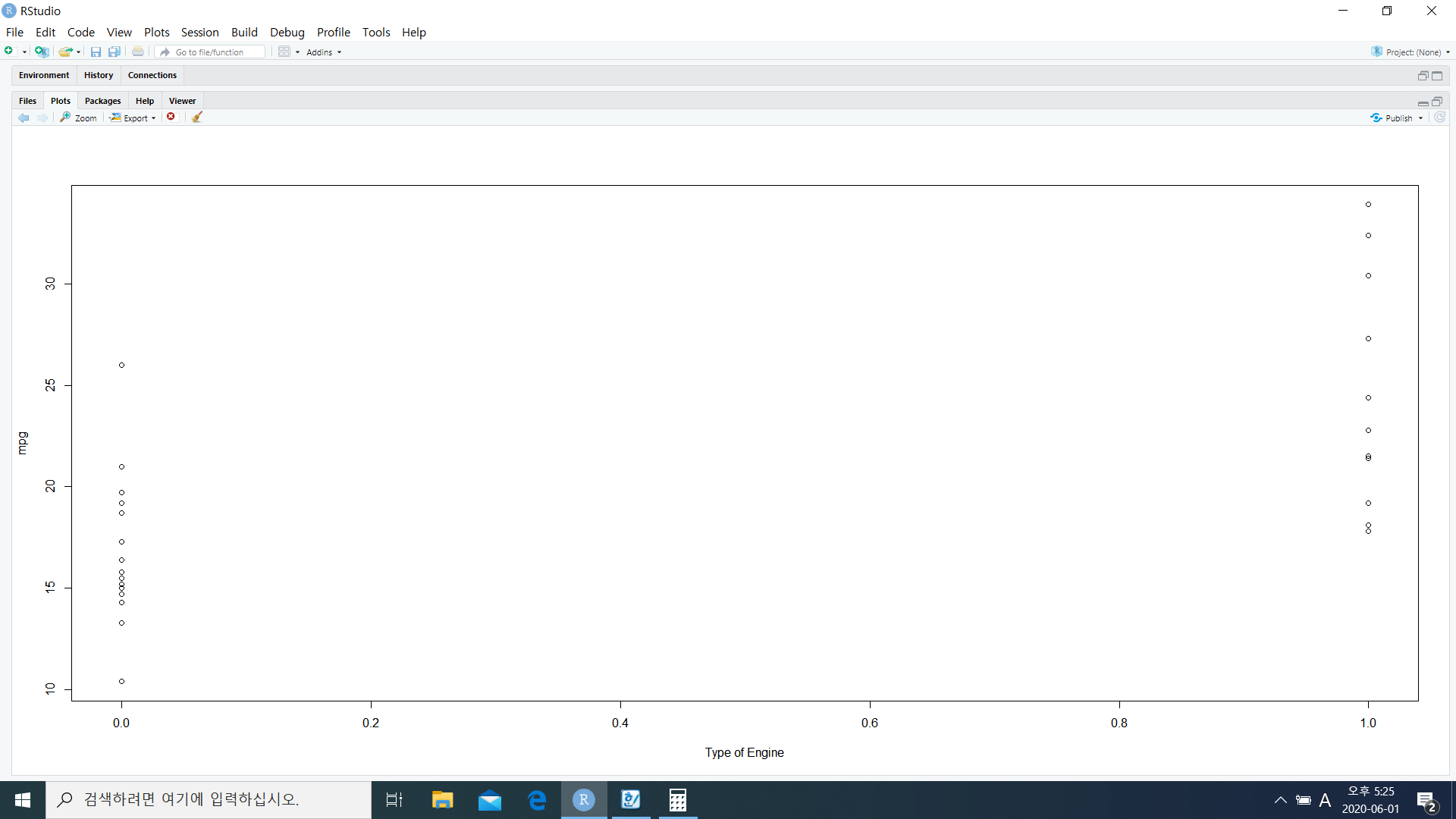
모형의 타당성 검정



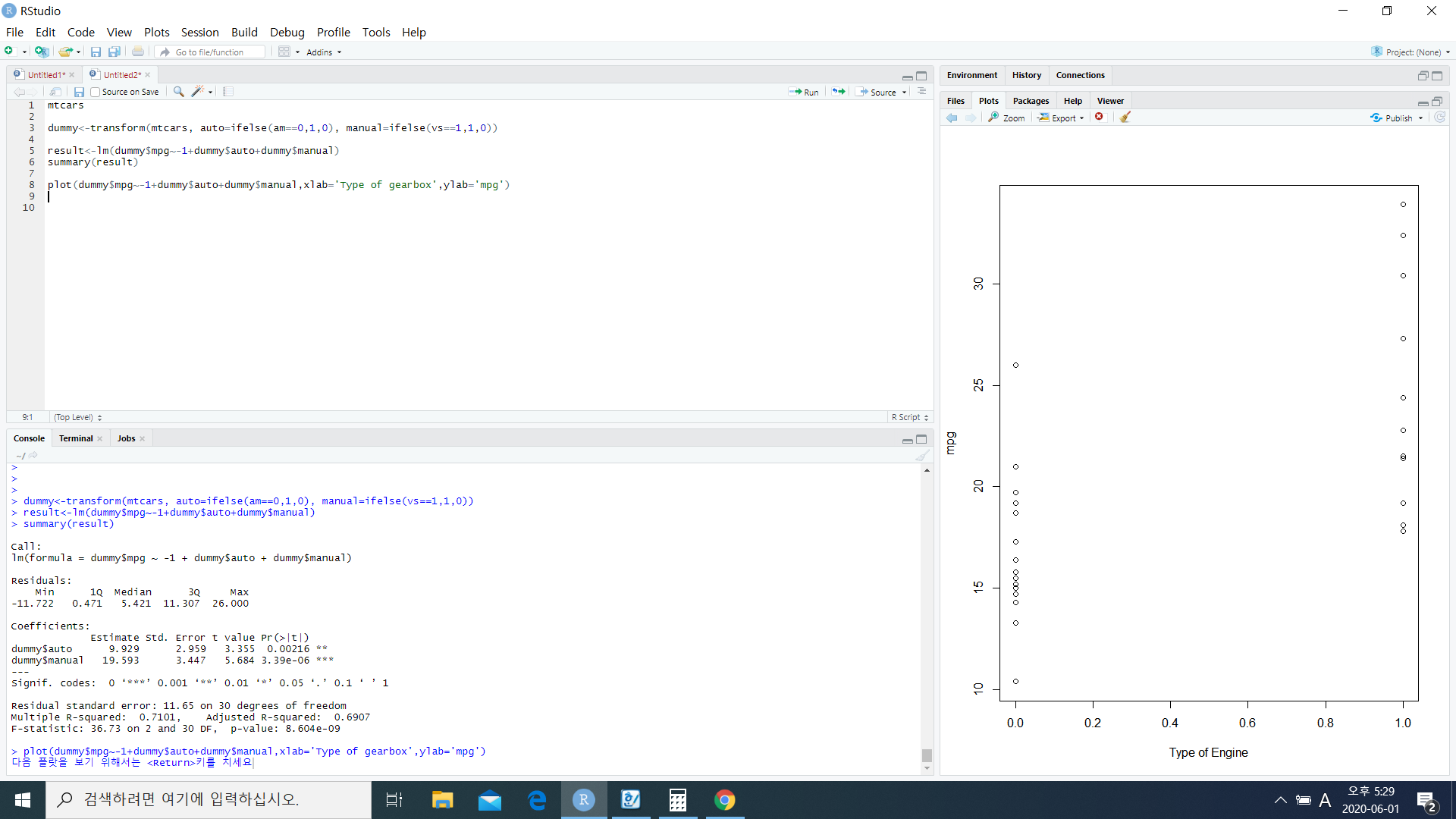
절대치 3을 넘어가는 값이 없으니 이상값이 없다고 판단할 수 있다.

**[10-1번]**

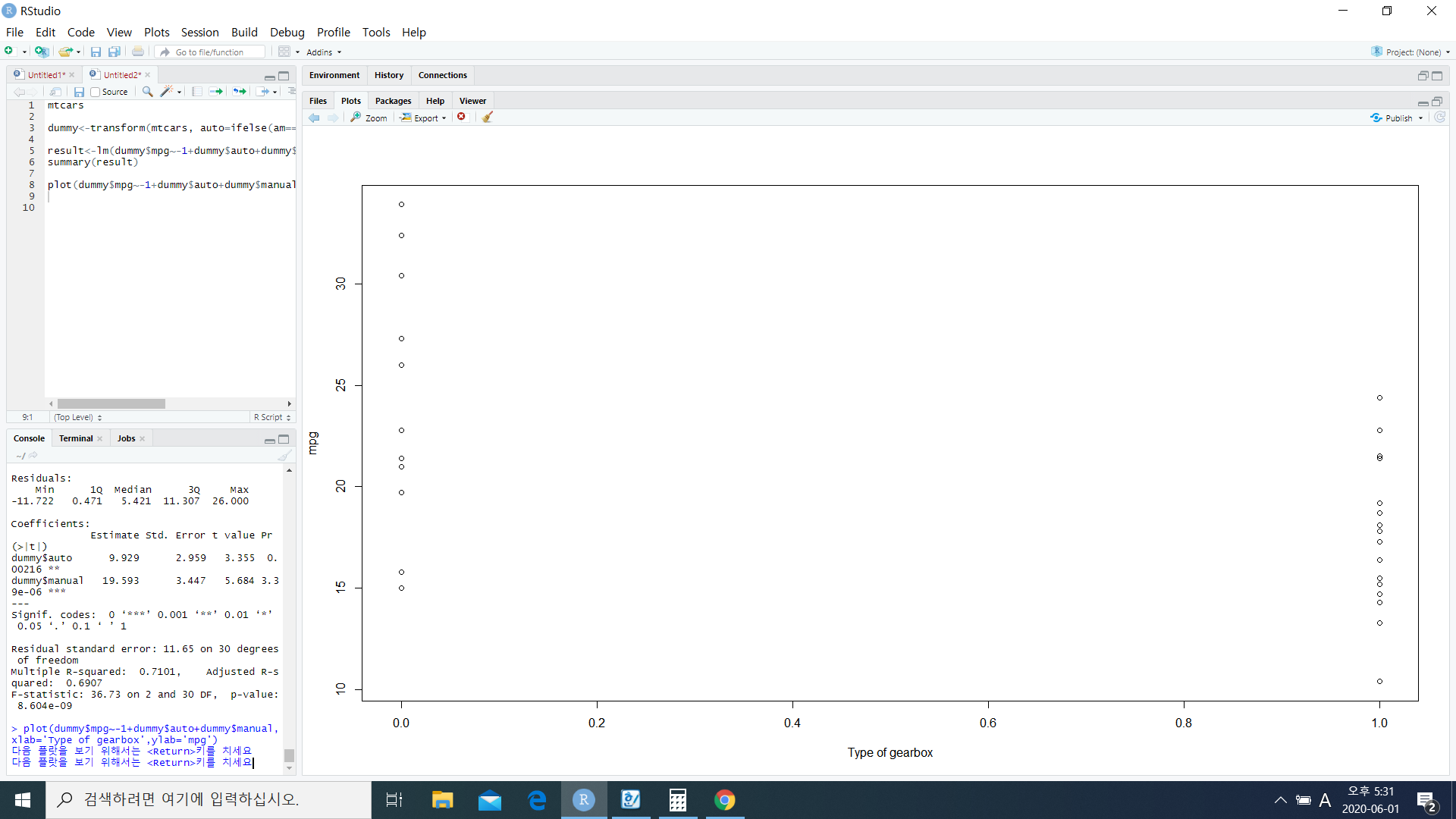
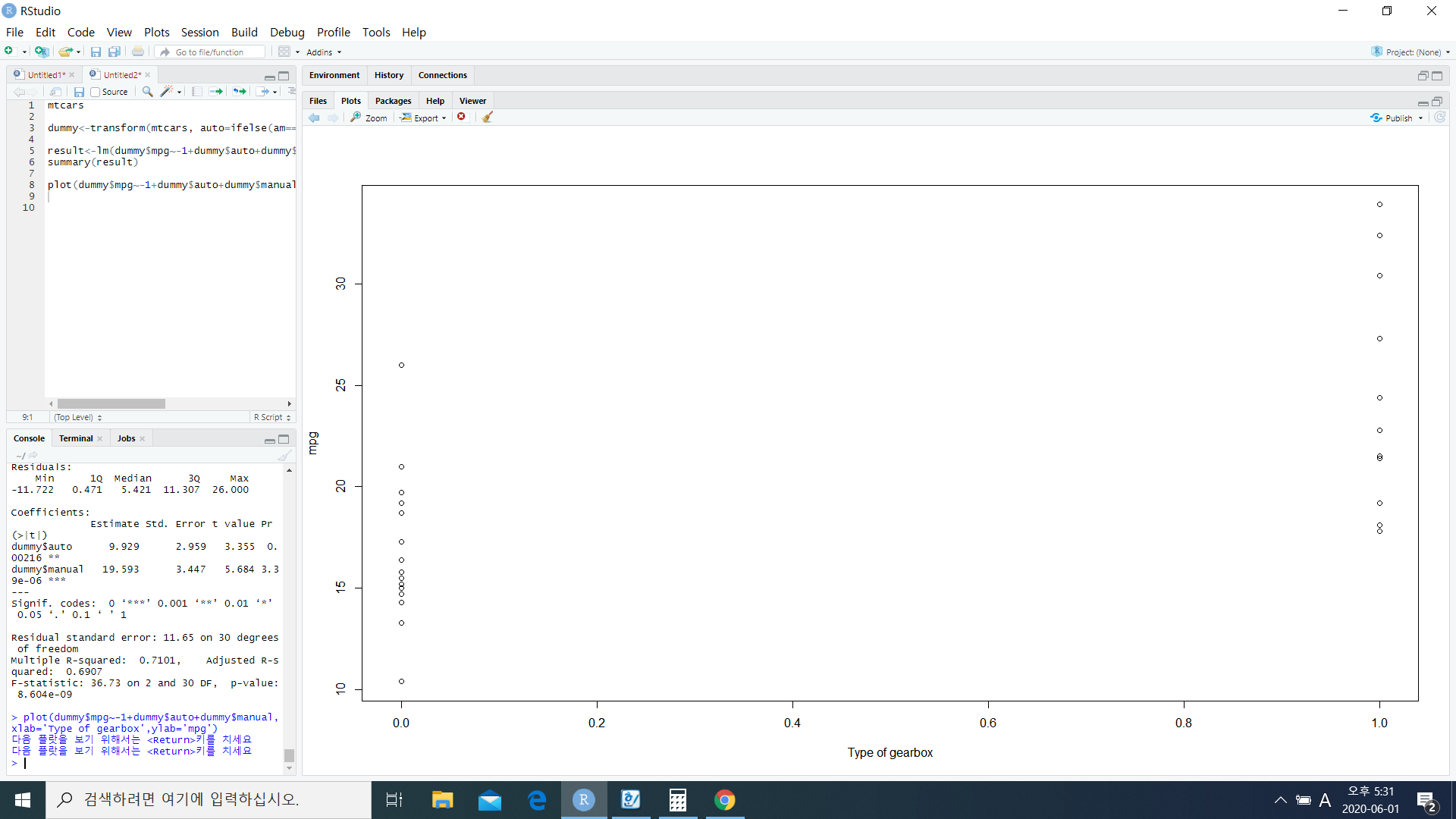
V\_engine의 계수가 16.617이므로 V\_engine은 Straight\_engine에 비해 평균 16.617만큼 연비가 좋다는 사실을 말해준다. 그리고 Straight\_engine은 V\_engine에 비해 평균 24.557만큼 연비가 좋다.



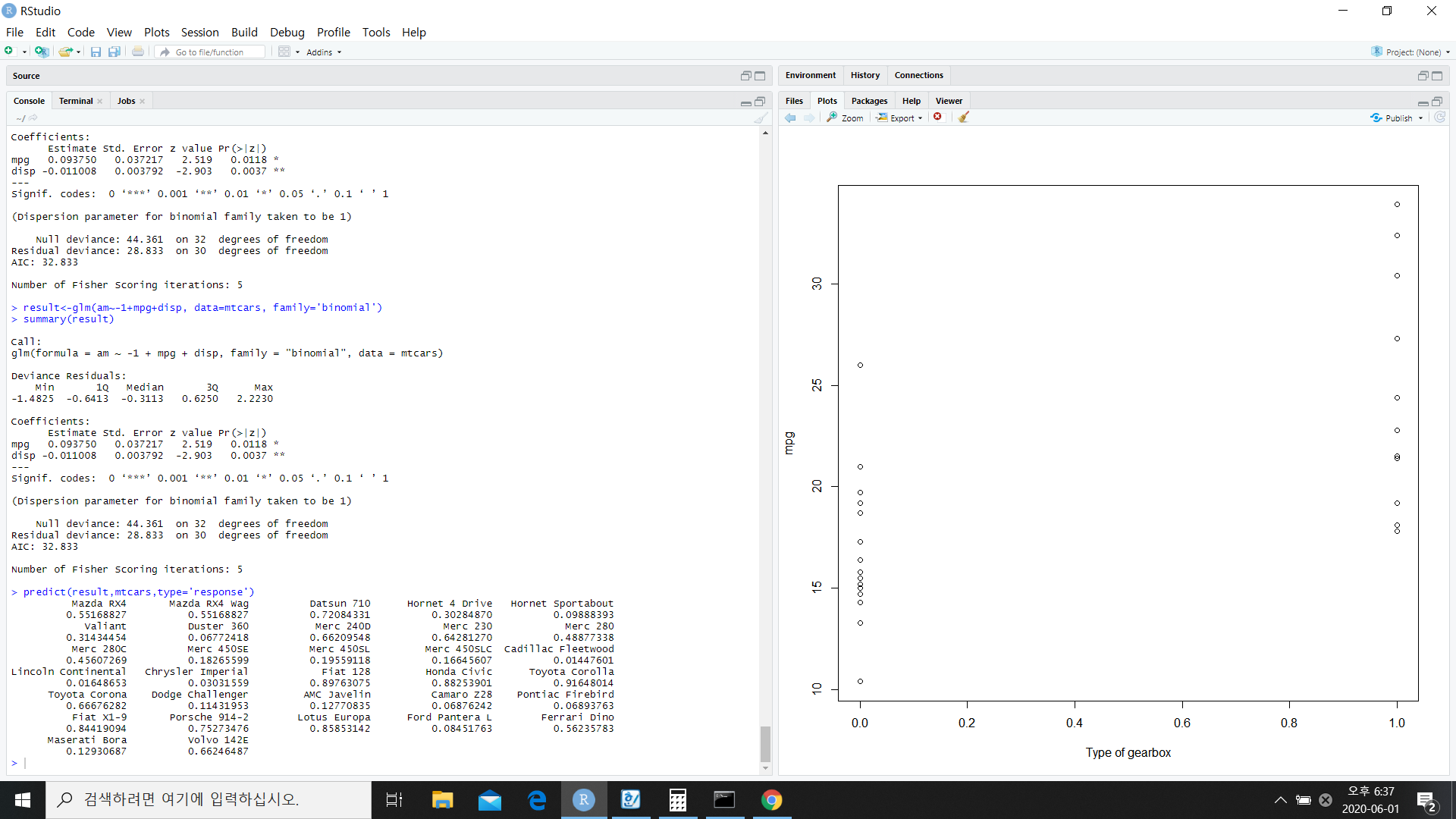
**[10-2번]**



수동변속기의 계수가 19.593이고 자동변속기의 계수가 9.929이므로 수동변속기는 자동변속기에비해 평균19.593만큼 연비가 높다.



**[12번]**



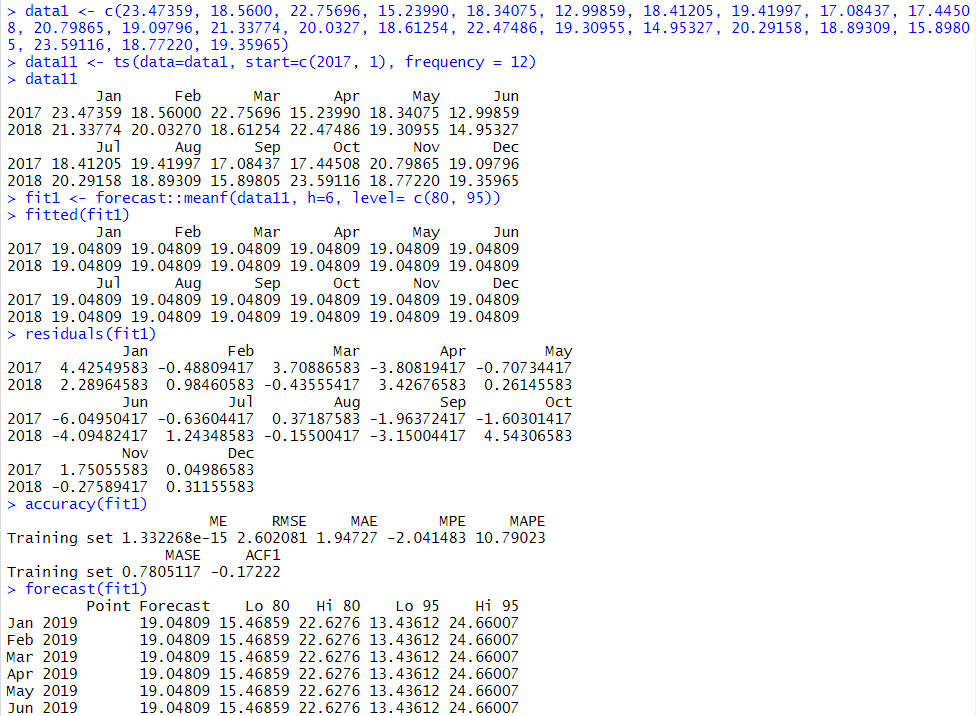
**[실습과제]**

변수가 27개이니 먼저 다중공선성 검정을 통해 종속적인 변수들을 제거하고 난 뒤 전방추가법, 후방제거법, 단계별제거법, 모든 가능한 방법을 적용한다.

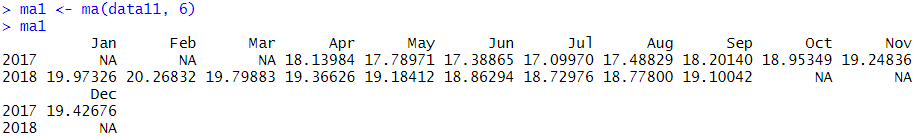
**8장 연습문제**

**[1번]**

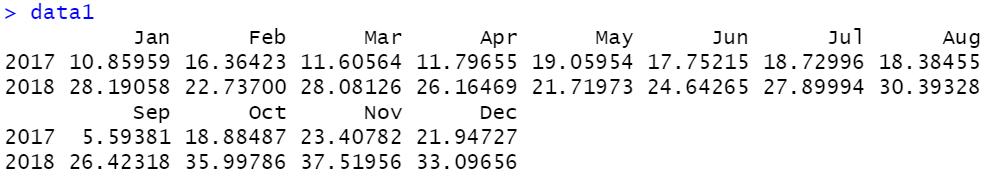
0 전체 데이터 평균으로 예측

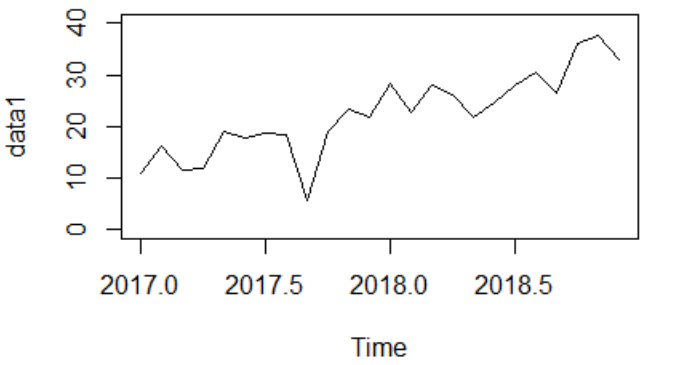


0 이동평균 구하기



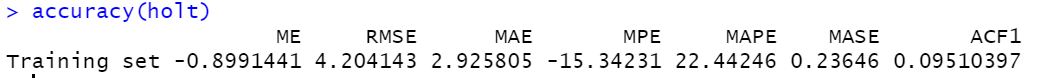
**[2번]**





- 그래프를 보니 추세패턴 시계열임을 확인 할 수 있음.

- 정확도



#예측값과 신뢰구간

