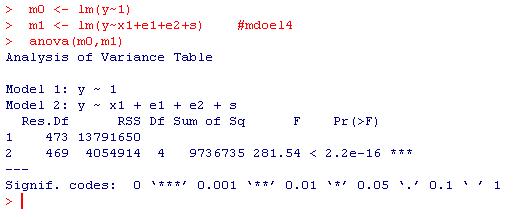
#1.-(1) ANOVA test



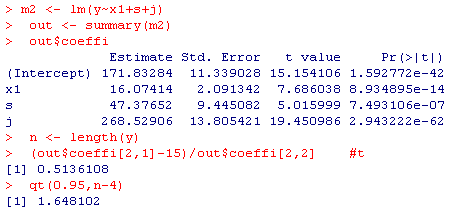
1. FM : model4

RM :

1. H0 : Vs H1 :
2. F = 281.54
3. R : ) = 2.2e-16

따라서 p-value가 0.05보다 작으므로 H0를 기각하여 모델 4는 유의하다 할 수 있다.

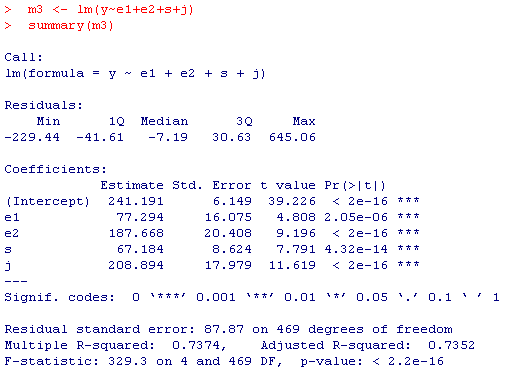
#1-(2) single coeffi test



1. Model 1
2. H0 : vs H1 :
3. t = 0.5136
4. R : =1.648102 로 검정통계량 t보다 크므로 H0를 기각할 수 없다.

따라서 직군과 성별이 같다면, 경력이 증가할 때 평균급여가 15만원보다 더 증가한다고 할만한 충분한 근거가 없다.

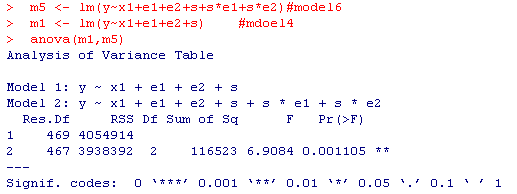
#1-(3) single coeffi test



1. Model 3
2. H0 : 0 vs H1 0
3. t = 7.791
4. R : =4.32e-14 로 p-value가 0.05보다 작으므로 H0를 기각할 수 있다.

따라서 학력과 직군이 같을 때, 성별에 따라 급여에 차이가 있다고 할만한 충분한 근거 있다.

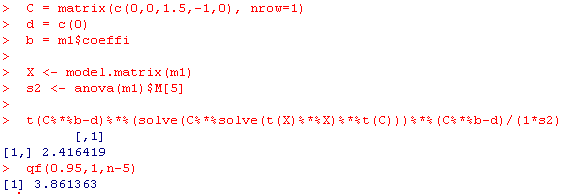
#1-(4) sequencial F-test



1. Model 3
2. H0 : 0 vs H1 0 (for at least one)
3. f = 6.9084
4. R : =0.0011 로 p-value가 0.05보다 작으므로 H0를 기각할 수 있다.

따라서 경력이 같을 때 학력에 의해 생기는 연봉의 차이가 남자와 여자에서 다르게 나타난다고 할만한 충분한 근거가 있다.

#1-(5) general F-test

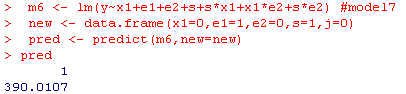


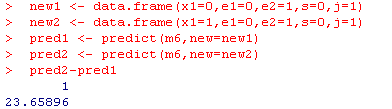
1. Model 4
2. H0 : vs H1 /
3. F=2.4164
4. R : =3.8613 로 검정통계량 f보다 크므로 H0를 기각할 수 없다.

따라서 경력과 성별이 같을 때, 초대졸으로써 더 받는 급여가 대졸으로써 받는 더 급여의 1.5배가 아니라고 할만한 충분한 근거가 없다. (1.5배라고 할만한 충분한 근거가 있다. )

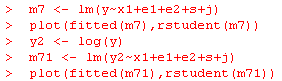
#1-(6)

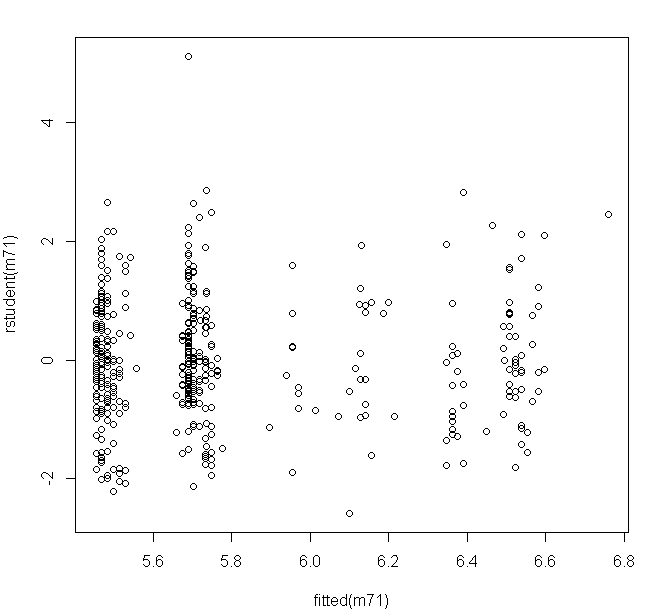
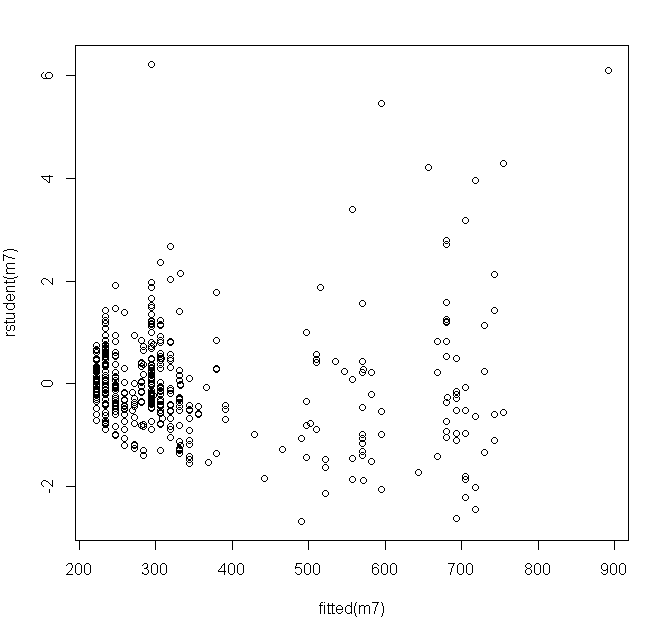
* 1. 신입사원 대졸 남성의 평균 현재급여



* 1. 경력이 1년 증가할 때, 초대졸 임원급 여성의 평균 급여의 변화량 (5점)
  2. 

#1-(7)





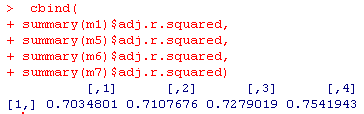
오차항의 대한 가정 중 등분산성이 위배되었다.

따라서 y에 log를 취하여 변수변환 하여 이를 해결할 수 있다.

( y 값에 대한 scale를 줄여 변동폭 또한 작게 만들어 이를 해결할 수 있다.)

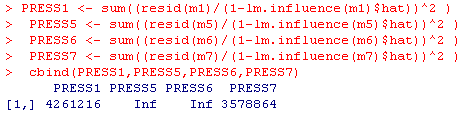
#1-(8)

1. 기준으로 어떤 모델이 가장 설명력이 좋은가? (모델별 제시)



Model7이 기준으로 가장 좋은 모델이다.

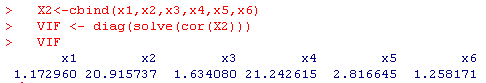
1. 기준으로 어떤 모델이 가장 설명력이 좋은가? (모델별 제시)



Model7이 PRESS기준으로 가장 좋은 모델이다.

#2. [자료 q2.txt] 주어진 자료에 대한 다음 물음에 답하라.

1. 독립변수 모두를 포함하는 모델에서 VIF를 구하여 다중공선성의 문제가 있는지 판단하여라.

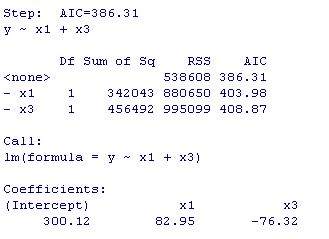


VIF2와 VIF4가 10보다 크므로 다중공선성의 문제가 있다고 판단된다.

1. (1)에서 구한 VIF2(x2에 대한 VIF)가 무슨 뜻을 갖는지 설명하여라.

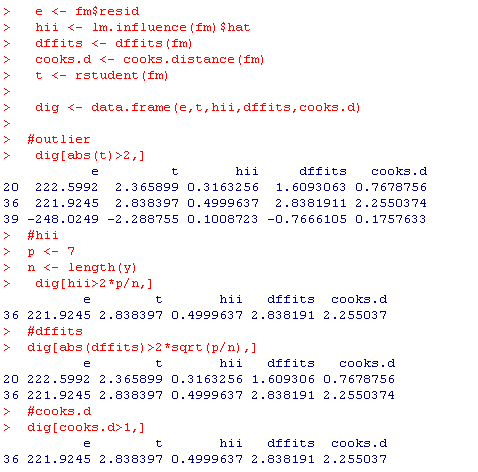
X2가 나머지 다른 변수들과 선형관계를 가지는 것 때문에 var(가 1일 때(변수들간에 linearly independent(독립)할 때) 대비 약 20배 증가(influence)했다는 뜻이다.

1. 후진제거법 (Backward elimination)을 이용하였을 때 선택되는 변수를 쓰시오.



X1,X3

1. (3)번에서 얻은 모델에서 관측치들 중 영향점(influential point)를 찾고자 한다.

다음 값들을 구하여, 영향점을 제시하여라. (ti(r-student residual), hii, dffits, cooks.distance) 

Outlier(|ti|>2,3)이고 high-leverage point(hii>2\*p/n)이면 영향점이라 할 수 있다.(36)

또한 |DFFITSi|가 보다 클 때 또한 영향점으로 판단할 수 있다.(20,36)

마지막으로, cooksi>1일 때 또한 영향점으로 판단된다.(36)

따라서 36번째 obs는 영향점으로 보이며, 20번째 obs또한 영향점으로 의심해 볼만 하다.

#3. 같은 MSE을 갖는 두 모델이 있다. 하나는 bias가 크고, 또 다른 하나는 variance가 크다.

이때 두 모델의 차이를 서술하여라.

같은 MSE를 가지는 모델이므로 예측력에 차이는 없을 것이라 판단된다.

하지만 BIG Bias 인 모델은 true값에서는 벗어나 있지만 예측값들 간에 variance(퍼진정도)가 작아 추청값의 신뢰구간이 작다.(underfitting의 상황)

또한 BIG variance 인 모델은 true 값을 포함하고는 있지만 예측값들 간에 variance(퍼진정도)가 매우 커 추정값의 신뢰구간이 커진다.(overfitting의 상황)

따라서 MSE가 같거나 비슷한 모델 두 개가 있다면, Cp를 통햬 bias가 큰 모델인지 variance가 큰모델인지 볼 필요가 있다.