**바이오 데이터의 이해**

-선형 회귀 분석



|  |  |
| --- | --- |
| **과목** | 바이오 데이터의 이해 |
| **담당교수** | 손대순 교수님 |
| **학과,이름** | 데이터테크전공 20183226신희진 |
| **제출일** | 2019.12.13 |

**목 차**

**Ⅰ. 서론**

ⅰ. 선형 회귀분석

-단순 선형 회귀분석

-다중 선형 회귀분석

ⅱ. 선형 회귀분석을 한다는 것

**Ⅱ. 본론**

ⅰ. 선형 회귀분석의 4가지 기본가정의 선형모델

1. 선형성
2. 독립성
3. 등분산성
4. 정규성

ⅱ. 선형 회귀분석의 가설검정

-왜 우측검정인가?

**Ⅲ. 결론**

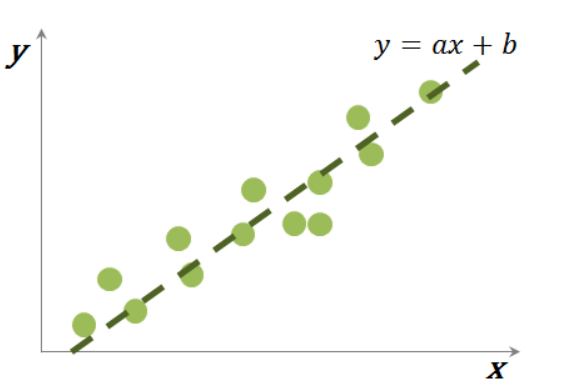
-회귀분석 결과의 해석

**Ⅳ. 출처**

**I. 서 론**

**ⅰ-ⅰ. 단순 선형 회귀분석**

선형 회귀분석은 주어진 데이터를 대표하는 하나의 직선을 찾는 것인데, 이 직선을 회귀선이라고 부르고 이 선을 함수로 표현한 것을 회귀식이라고 부른다. 아래 그래프는 단순 선형 회귀분석을 시각적으로 표현한 그림인데, 이처럼 단순 선형 회귀분석은 x변수와 y변수 간의 관계를 y = ax + b 와 같은 하나의 선형 관계식으로 표현한다. 이때 y = ax + b 인 회귀 식에서 x가 1개인 것을 보면 알 수 있듯이 단순 선형 회귀분석에서 Input이 되는 설명변수는 1개이다.



**ⅰ-ⅱ. 다중 선형 회귀분석**

단순 선형 회귀분석이 설명변수 x가 1개, 반응변수 y가 1개인 선형 회귀분석이라면, 다중 선형 회귀분석은 설명변수 x가 2개 이상, 반응변수 y가 1개인 선형 회귀분석이다. 따라서 다중 선형 회귀분석은 아래와 같은 회귀 식을 갖는다.



다중 선형 회귀분석에서는 최소제곱 법을 이용하여 각 설명변수 x의 가중치 역할을 하는 회귀계수(a, b, ... )와 y절편(c)을 구하게 된다.

**ⅱ. 선형 회귀분석을 한다는 것**

회귀분석을 한다는 것은 분석가가 데이터, x 변수, y 변수를 지정하고 통계 툴이 최소제곱 법을 이용해 회귀 식 y = ax + b에서 a와 b를 구하는 과정을 말한다. 회귀 식에서 a를 회귀계수, b를 y절편이라고 부른다. (큰 의미에서 b도 회귀계수라고 부르는 경우도 있다.) a는 가중치의 역할을 하기 때문에 x가 y에 얼마나 영향을 주는지 그 크기와 방향을 알 수 있게 해주고, b는 절편의 역할을 하기 때문에 y = ax 라는 회귀선을 얼마나 위 또는 아래로 평행이동 시키는지를 정한다.

**Ⅱ. 본 론**

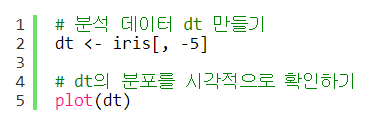
**ⅰ. 선형 회귀분석의 4가지 기본가정**

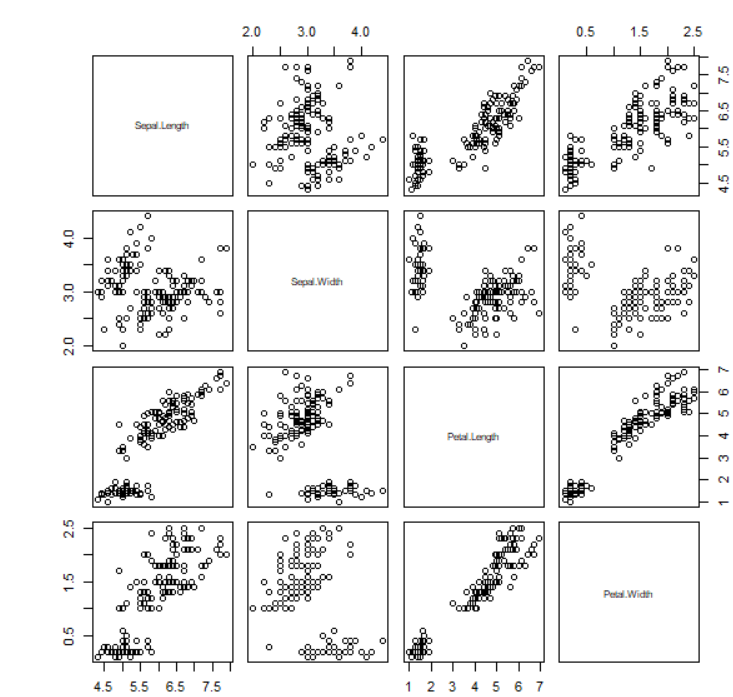
선형 회귀분석을 통해 좋은 모델을 만들기 위해서는 분석 데이터가 4가지 기본가정을 만족해야 한다. 4가지 기본가정을 만족하지 않으면 제대로 된 선형회귀모델이 생성될 수 없다.

1. 선형성
2. 독립성
3. 등분산성
4. 정규성

1 ) 선형성

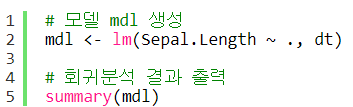
선형성은 “선형” 회귀분석에서 중요한 기본가정으로, 당연히 “비선형”회귀분석에서는 해당하지 않는다. 선형성이란, 예측하고자 하는 종속변수y와 독립변수 x 간에 선형성을 만족하는 특성을 의미한다.

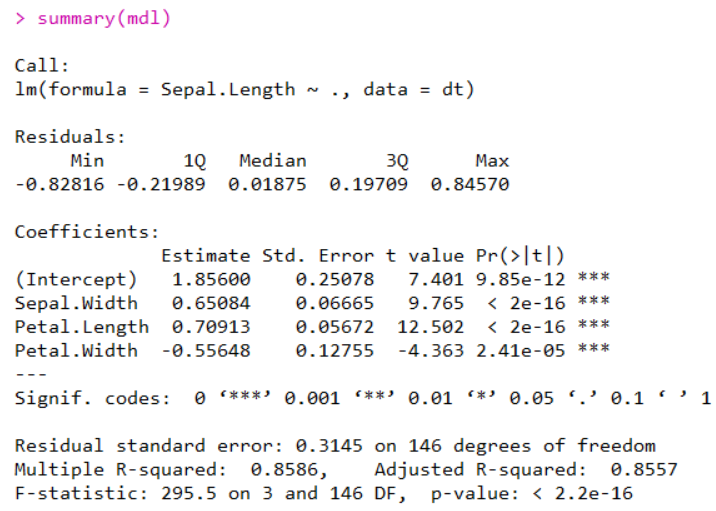




만약 Sepal.Length를 우리가 예측하려고 하는 종속변수 y라고 한다면, 위 그래프를 보았을 때 Sepal.Length와 대략적인 선형관계를 이루고 있는 변수는 Petal.Length와 Petal.Width이고, 선형성을 만족하지 않는 것은 Sepal.Width인 것으로 보인다.

이 데이터로 선형 회귀모델을 만들면





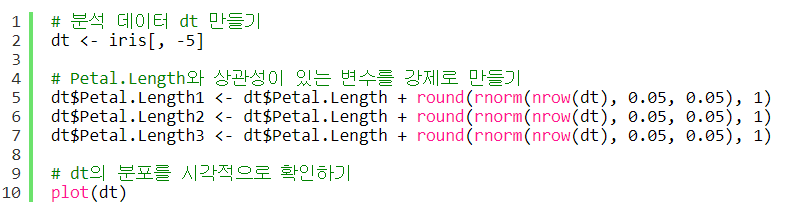
회귀분석 결과, 모든 변수가 유의한 모델이 만들어졌음을 확인할 수 있다.

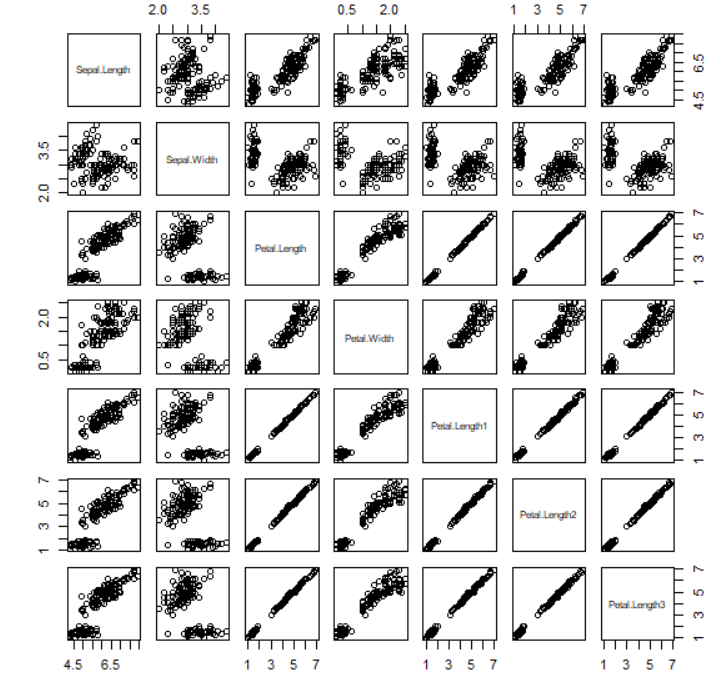
선형성을 만족하지 않는 Sepal.Width까지 유의 하다고 결과가 나왔는데 왜 이런 결과가 나온 이유는가장 선형성을 만족하는 Petal.Length와 Petal.Width의 영향 도를 뺀 나머지 값들이 종속변수 y인 Sepal.Width와 선형성을 이루기 때문이다

2) 독립성

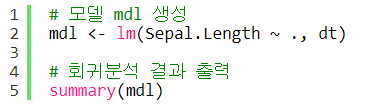
독립성은 "다중" 회귀분석에서 중요한 기본가정으로, 당연히 "단순" 회귀분석에서는 해당하지 않는다. 독립성이란, 독립변수 x 간에 상관관계가 없이 독립성을 만족하는 특성을 의미한다. (다중 회귀분석은 x변수가 2개 이상인 회귀분석이다.)

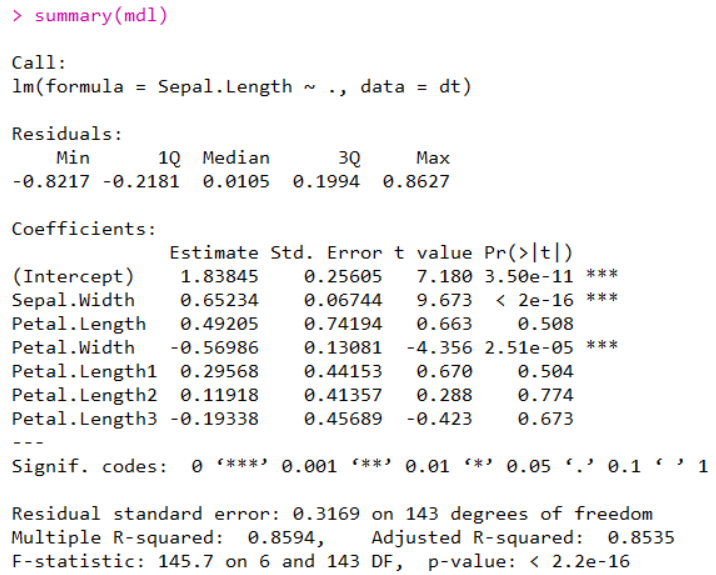
만약 서로 상관관계가 있는 독립변수 x들이 여러 개 들어간다면 회귀분석 결과는 어떻게 바뀔지 알아본다. 위에서 분석한 바로는, Sepal.Length와 Petal.Length는 선형성을 잘 만족하고 있는 유의한 변수이다. 이 Petal.Length와 유사한 변수를 3개 더 만든 경우



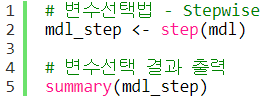


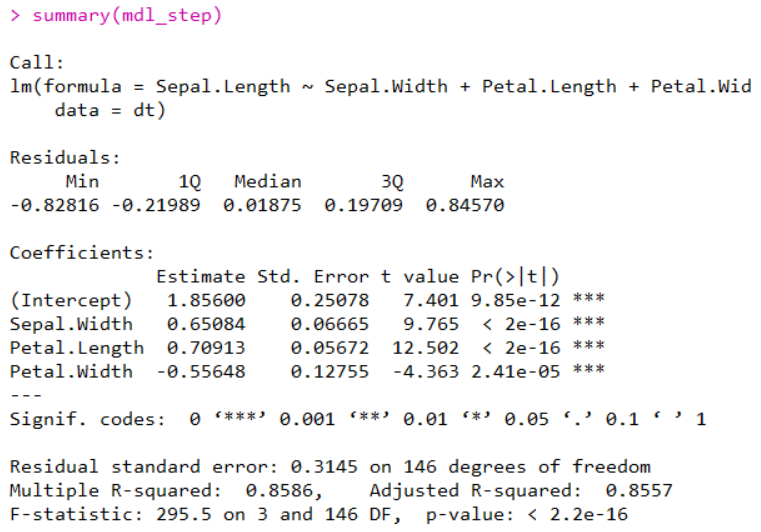
강제로 만든 Petal.Length1 ~ Petal.Length3 변수들이 Petal.Length까지 해서 강한 상관관계를 띄고 있으므로 당연히 이 데이터는 독립성을 만족하지 못하고 있는 상태이다. 이 상태의 데이터를 바로 회귀 분석하면 했을 경우





회귀분석 결과를 살펴보면 Petal.Length는 원래 유의했던 변수임에도 불구하고 독립성을 위배하는 비슷한 변수가 많이 들어가게 되니 마치 유의하지 않는 것과 같은 결과를 나타내게 된다. 그렇기 때문에 다중공선성, 즉 서로 상관관계가 있다면 이를 제거해주어야 한다. 제거하는 방법은 다중공선성을 일으키는 변수를 제거하거나, 이것들을 모아서 다른 변수로 치환을 해주면 된다. 변수선택법 중 Stepwise를 사용해서 다중공선성을 일으키는 변수들을 제거해보려고 한다.

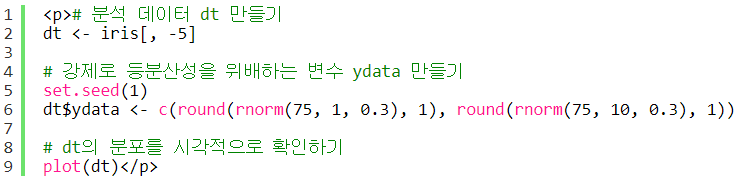


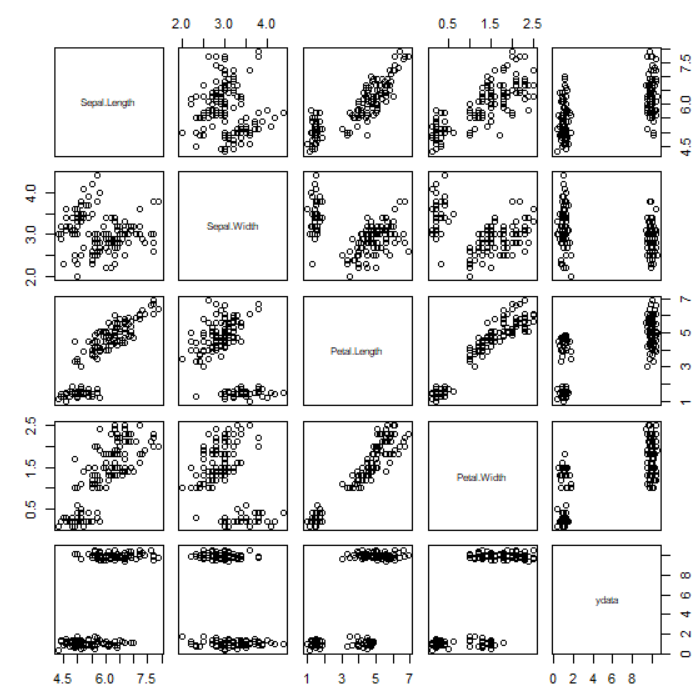


Stepwise를 사용하니, 독립성을 위배하는 강제로 만들었던 변수들이 제거되고 유의한 모델이 만들어진 것을 확인할 수 있다.

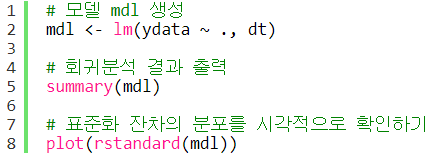
3) 등분산성

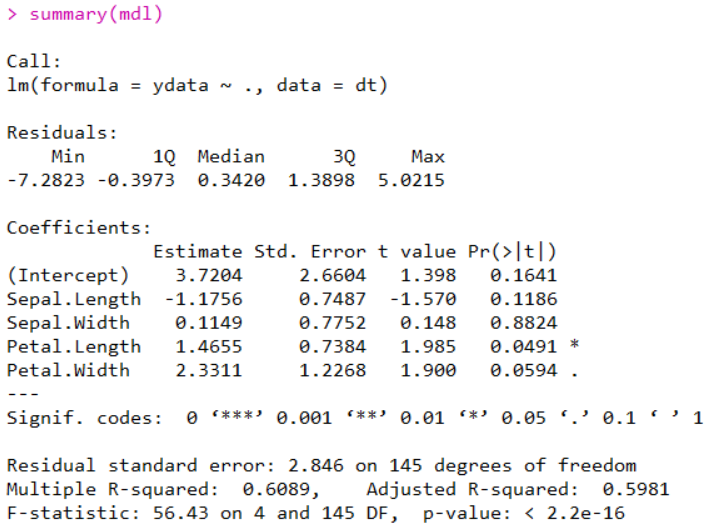
등분산성이란 분산이 같다는 것이고, 분산이 같다는 것은 특정한 패턴 없이 고르게 분포했다는 의미이다. 여기서 등분산성의 주체는 잔차이다. (여기서는 표준화 잔차를 볼 것이다.)



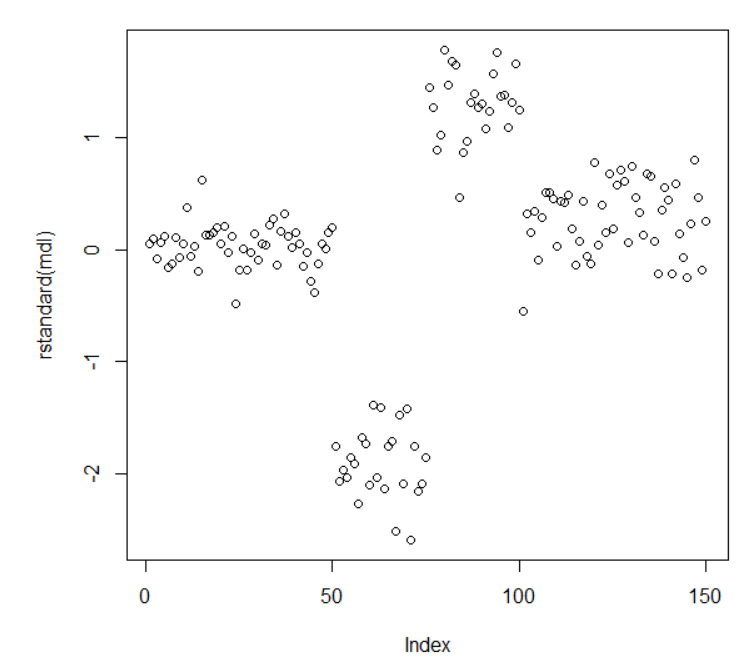


위 그래프에서 볼 수 있듯이 매우 이상한 ydata라는 변수를 만들었다. 앞에서 했던 것과 다르게 Sepal.Length 대신 ydata를 종속변수 y로 두고 예측해 볼 때





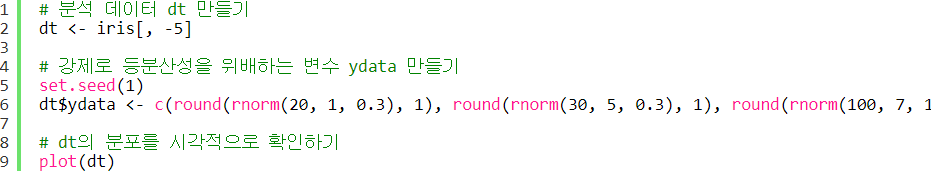
회귀분석 결과 유의한 모델이 나오지 않았는데, 표준화 잔차의 분포를 살펴보면 그 이유를 알 수 있다.

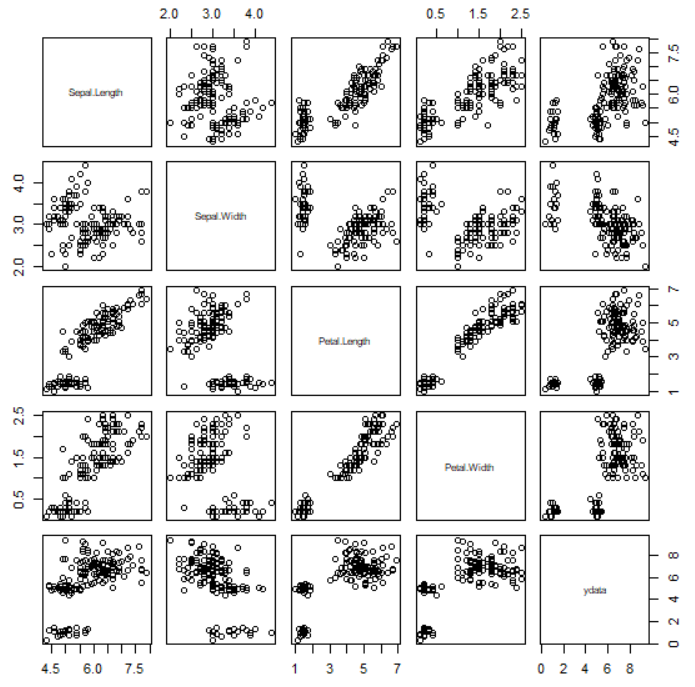


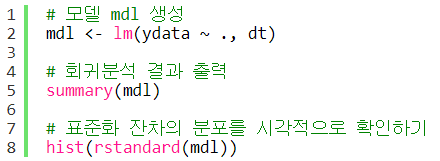
표준화 잔차가 등분산성을 만족하지 않고 4개의 덩어리가 있는 특정 패턴을 가지고 있기 때문이다. 표준화 잔차가 이러한 분포를 가지고 있는 경우, 매우 중요한 변수가 분석 데이터에 추가되지 않고 빠져있다는 의미로 해석할 수 있다. 앞에서 우리가 만든 ydata는 위에 75행은 평균이 1인 랜덤 값이고, 아래 75행은 평균이 10인 랜덤 값이므로 1~75행과 76~150행을 나누어줄 수 있는 추가적인 변수가 필요하다.

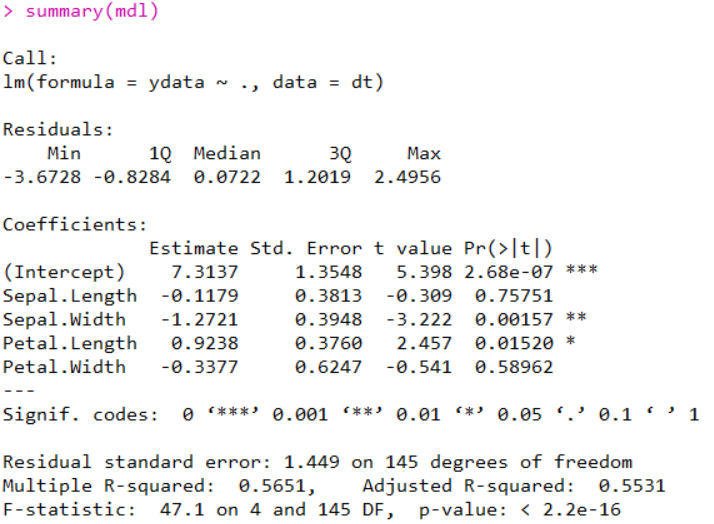
4) 정규성

마지막 정규성은 또한 잔차가 정규성을 만족하는지 여부로, 정규분포를 띄는지 여부를 의미한다.

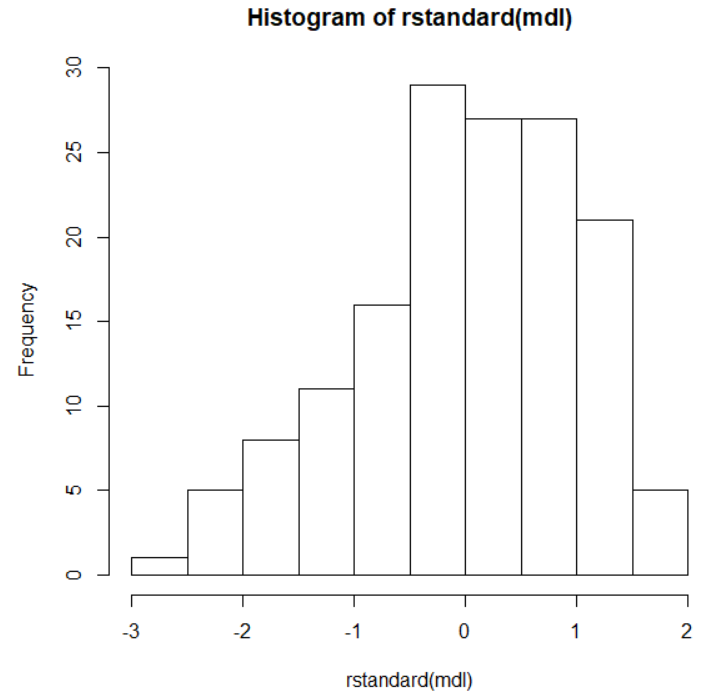


  
강제로 한쪽에 쏠려있어 정규분포가 아닌 변수 ydata를 만들어보았다. 이 변수로 회귀분석을 해볼 때

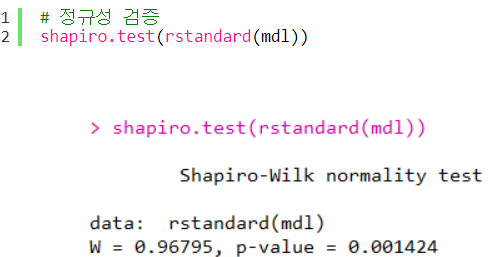




유의하지 않은 변수들이 섞여있는 것을 확인할 수 있다. 표준화 잔차의 히스토그램을 살펴보면 정규분포가 아님을 시각적으로 확인할 수 있다.



시각적으로 말고 수치적으로 확인하기 위해 Shapiro-Wilk Test로 정규성 검정을 해보았을 때.



정규성 검정의 귀무가설은 "정규분포와 차이가 없다" 이다. p-value가 0.001 이므로 유의수준 0.05보다 작으므로 귀무가설을 기각한다. 즉, 이 데이터는 정규분포와 차이가 있다 = 정규성을 만족하지 않는다. 라는 해석을 할 수 있다. 정규성을 만족하기 위해서는 (3) 등분산성을 해결하기 위한 방식와 유사한 방법들이 필요하다.

**ⅱ.** 선형 회귀분석의 가설검정

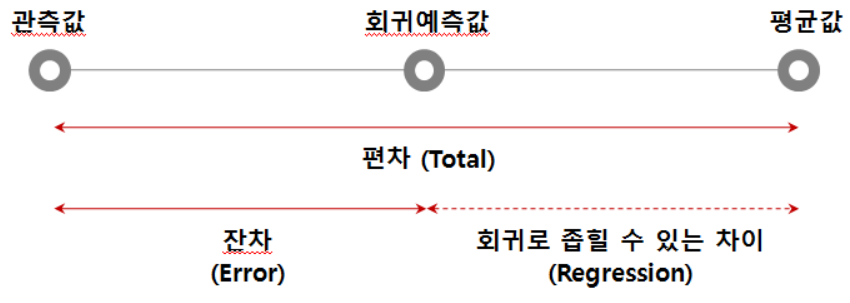
-왜 우측검정인가?

선형 회귀분석의 가설은 흔히 "y = ax + b 에서 회귀계수 a는 0이 아니다"라는 대립가설을 가정한다고 한다. 즉 x의 기울기가 0이 아니기 때문에 "x는 y에 영향을 준다"가 선형 회귀분석의 대립가설이 된다고 한다. 하지만 이 가설은 개별 설명변수에 대한 가설이고 회귀모델 자체에 대한 가설은 아니다.

실질적으로 회귀모델에 대한 유의성 검증은 F-Test로 이루어지는데 이를 위해 분산분석(ANOVA, Analysis of variance) 표를 생성하게 된다. 분산분석표는 아래와 같은 구조를 가진 표이다.



위 표에 따르면 F-통계량은 회귀제곱평균(MSR) / 잔차제곱평균(MSE) 이라고 정의되는데, 그렇게 정의하는 이유는 아래 예시와 그림으로 이해해볼 수 있다. 예를 들어 대한민국 남성의 키 데이터를 가지고 어떤 남성의 키를 예측한다고 하자. 가장 간단하게 아무런 부가정보 없이 예측 키를 말하는 방법은 데이터의 평균 키를 말하는 방법이다. 그래서 분산분석표에서도 기준을 '평균값'으로 잡고 평균으로 예측한 것 대비, 회귀로 예측한 것이 얼마나 더 유용한지 그 비율을 계산한 것이 F-통계량(F-Value)이라고 할 수 있다. 아래 그림과 위 표를 함께 보면 조금 더 이해하기 쉬울 것 같다. 관측 값은 우리가 예측하려는 목표치, 실제 y값이라고 보면 되고, 회귀 예측값은 회귀모델로 예측했을 때의 값, 그리고 평균값은 관측 값 y들의 평균이라고 생각하면 된다.



회귀제곱평균(MSR)은 회귀로 설명할 수 있는 편차라고도 부른다.

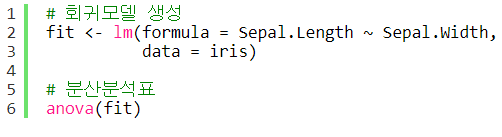
회귀제곱평균은 회귀제곱합(SSR) / 자유도(DF) 인데 회귀제곱합(SSR)은 회귀예측값과 평균의 차이에서 기인한다. 즉, 평균으로 말하는 것과 회귀로 예측해서 말하는 것의 차이가 얼마나 있는지를 말해주는 수치라고 할 수 있다.

잔차제곱평균(MSE)은 회귀로도 설명할 수 없는 잔차라고도 부른다.

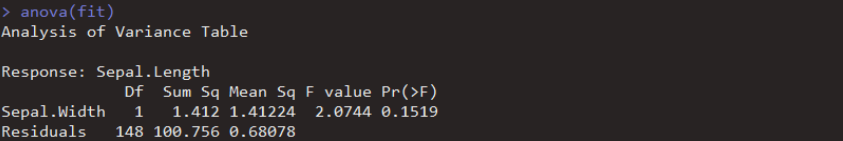
잔차제곱평균은 잔차제곱합(SSE) / 자유도(DF) 인데 잔차제곱합(SSE)은 관측값과 회귀예측값의 차이에서 기인한다. 즉, 회귀로 예측해도 좁힐 수 없는 관측값과의 차이가 얼마나 있는지 말해주는 수치라고 할 수 있다.

정리하자면 회귀로도 좁힐 수 없는 예측 오차 대비, 회귀로 인해 좁혀진 예측 오차가 얼마나 큰가를 표현한 수치가 F-통계량이다. 이 통계량이 크면 클 수록 (우리가 증명하고 싶은) 이 회귀모델이 유용하다는 것을 증명하기에 좋다. 그래서 회귀분석은 F-통계량에 대한 우측검정으로 가설을 검정하게 된다.

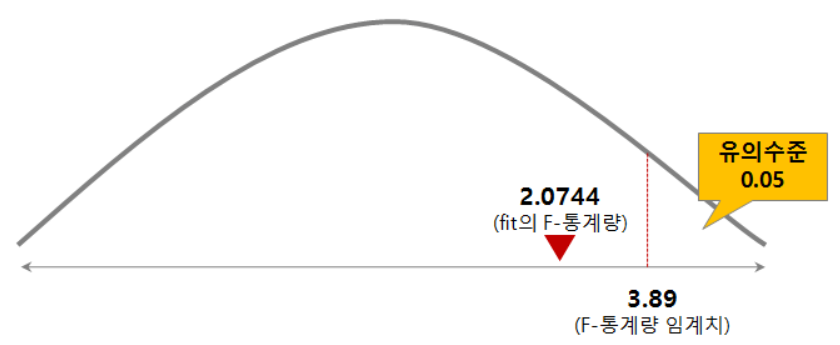
예를 들어 iris 데이터 중 Sepal.Length를 Sepal.Width로 예측한다고 했을 때



출력되는 분산분석표는 아래와 같다.



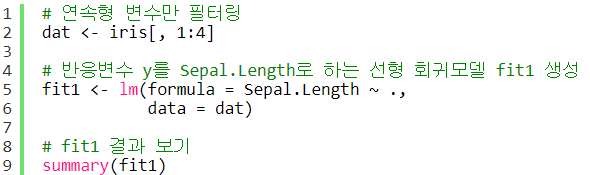
회귀모델 fit에서 F-통계량(F-Value)은 2.0744이다. 유의수준(P-Value)은 0.15로 유의확률 0.05보다 작지 못하므로 Sepal.Width 만으로 이루어진 fit 모델의 유의하지 못하다.



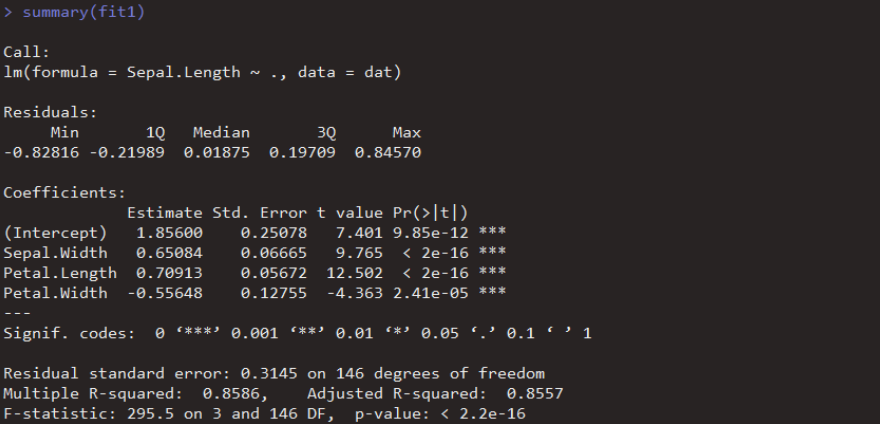
**Ⅲ. 결론**

-회귀분석 결과에 대한 해석

ris 데이터에서 Species 변수를 뺀 나머지 4개 변수로 이루어진 dat 데이터를 만들고 이 중 Sepal.Length 변수를 종속변수 y 값으로 두고 회귀분석을 돌리는 R 코드이다.



위 코드를 돌려 출력된 내용을 보면서 결과를 해석해보면



이 결과에서 밑에서 두번째 줄 Adjusted R-squared, 조정된 결정계수는 0.8557이므로 회귀모델 fit1의 설명력은 85.57%라고 이야기할 수 있다. 회귀모델의 p-value는 < 2.2e-16으로 유의수준 0.05보다 작으므로 회귀모델은 유의미하고 해석할 수 있다. 또한 중간에 Coefficients 부분에 모든 변수의 Pr(>|t|), p-value가 유의수준 0.05보다 작기 때문에 모든 변수가 유의미하다고 해석할 수 있다.

각 변수의 Coefficients Estimate, 회귀계수에 따라 회귀식을 구성해보면 아래와 같다.



이 식을 간단히 해석해보면

- Sepal.Width가 1 증가할 때마다 Sepal.Length는 0.65084 만큼 증가한다.

- 모든 x변수가 0이어도 기본적으로 존재하는 Sepal.Length는 y절편인 1.856 이다.

- Petal.Width가 클수록 오히려 Sepal.Length는 작아진다.

**Ⅳ. 출처**

<https://kkokkilkon.tistory.com/77>

https://kkokkilkon.tistory.com/175