|  |
| --- |
| **데이터사이언스기초 기말고사 프로젝트** |
|  |

|  |  |
| --- | --- |
| **이름** | **김수현** |
| **학번** | **20215124** |
| **소속**  **학과/대학** | **빅데이터** |

## <주의사항>

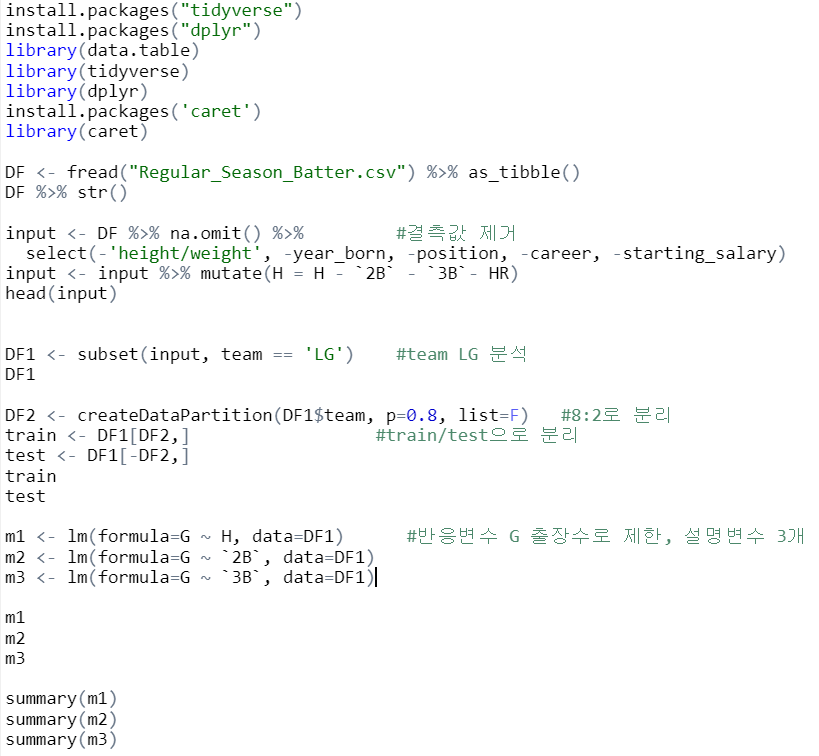
* 개별 프로젝트 입니다. (팀으로 진행하는 과제가 아니며, 모든 학생이 제출해야 함)
* 각각의 문제 바로 아래에 답을 작성 후 제출해 주세요.
  + 소스코드/스크립트 등을 작성 한 경우, 해당 파일의 이름도 적어주세요.
* SMARTLEAD 제출 데드라인: 2022. 06. 06(월요일) ~ 06. 12.(일요일) 23:59 까지
  + 부정행위 적발 시, 원본(보여준 사람)과 복사본(베낀 사람) 모두 0점 처리함
  + 예외 없음
* SMARTLEAD에 아래의 파일을 제출 해 주세요
  + 보고서 형식의 문서와 스크립트(R코드)를 압축하여 제출해야 합니다.
  + 압축 파일명에 이름과 학번을 입력 해 주세요.

## <문제 1번>

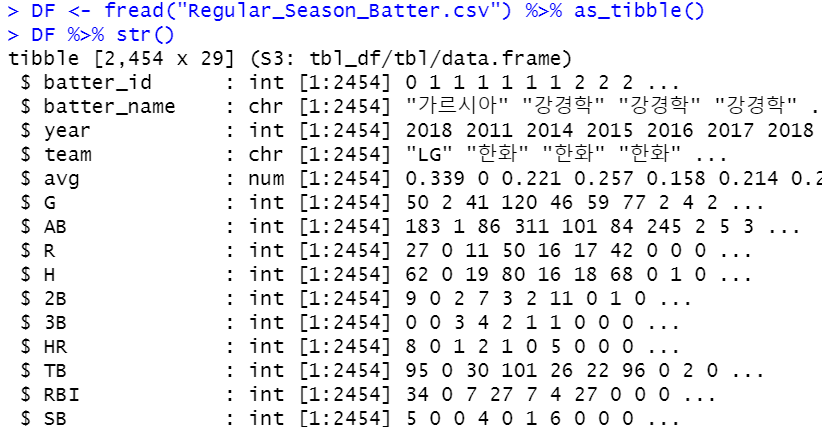
|  |
| --- |
| [Q 1] **[30점] – lm을 이용한 다중회귀분석 적용**  **<사용되는 데이터 – Regular\_Season\_Batter.csv >**  **Regular\_Season\_Batter.csv 데이터를 가지고 다중회귀분석을 적용 및 예측하자!**  **Regular\_Season\_Batter.csv는 결측값을 가지고 있어 시작 전에 아래의 코드를 적용하고 진행하세요.**    **<참고사항>**   1. **데이터는 Regular\_Season\_Batter.csv 데이터를 사용한다.** 2. **분석하고자 하는 데이터 중 자유롭게 한팀만 선택하는 것으로 제한한다.**  * **NC, 넥센, 키움, 현대, 히어로즈 팀은 선택하지 않는다(데이터가 너무 적음)**  1. **선택한 team 데이터를 caret 패키지의 createDataPartition을 이용하여 train / test 으로 분리할 것 (분리 비율은 본인이 지정)**  * **Train은 모델 생성할 때 사용하며, test는 모델 예측에 사용한다.**  1. **반응변수(종속변수)는 G출장수로 제한하고 설명변수(독립변수)들은 자유롭게 선택한다.** 2. **분석은 다중 회귀분석을 한다.** 3. **설명변수의 개수는 3개 이상으로 설정한다.** 4. **사용하는 설명변수에 대한설명, 모델적합(summary 결과 해석), 모델을 활용한 예측에 대한 설명 포함되어야 함** 5. **다중회귀분석 : 모델 생성, 사용하는 변수 및 모델적합에 대한 설명(20점), 예측(10점)** |

**<문제 1번>**

**코드 :**



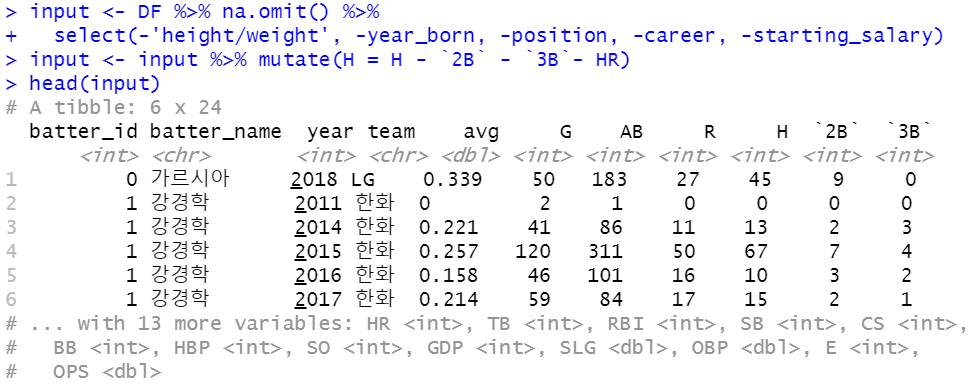
**결과 :**

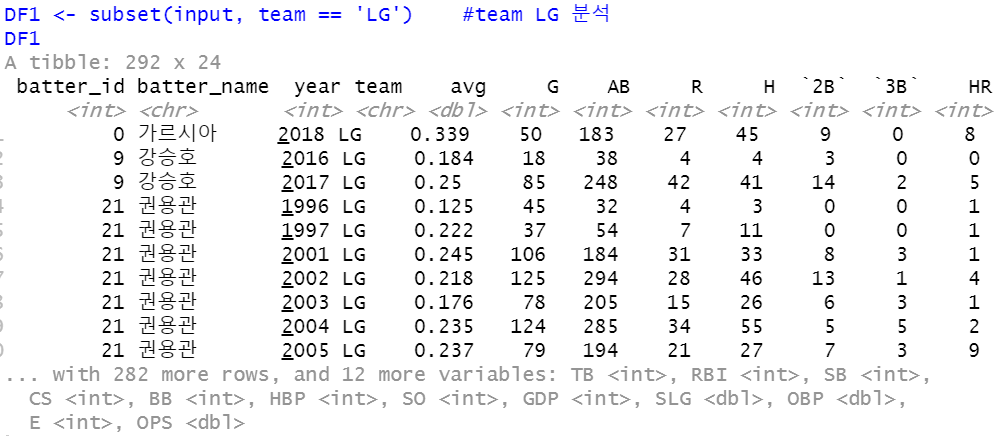
****

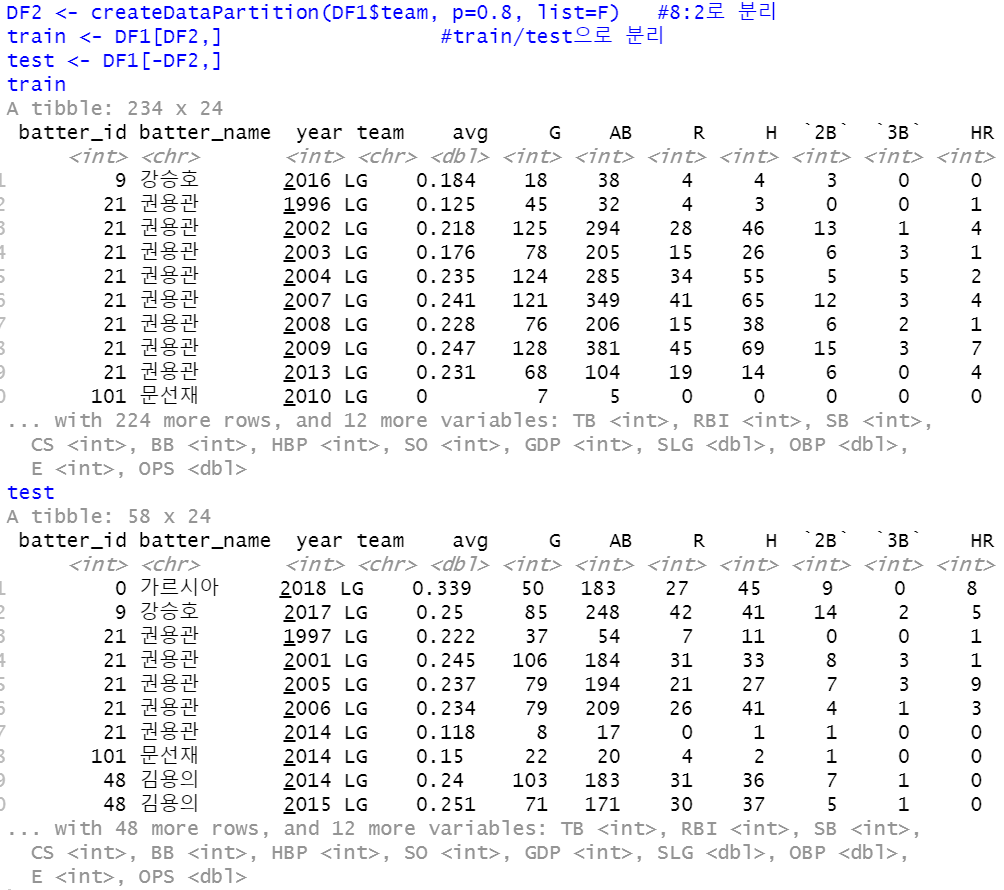
**.**

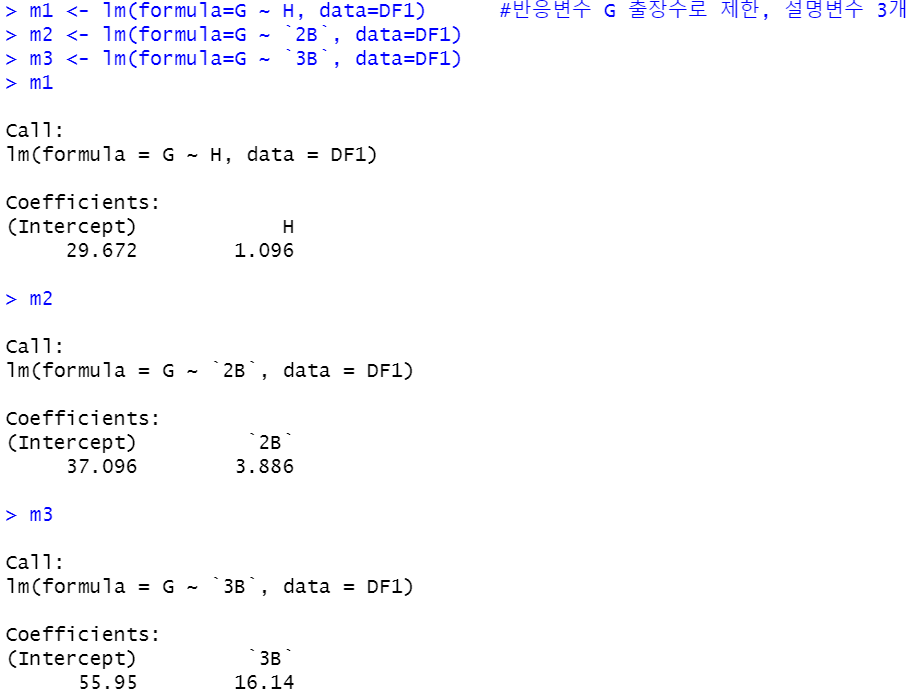
**.**

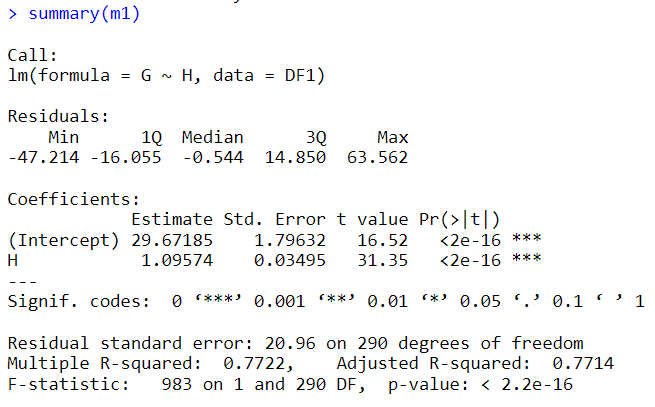
**.**

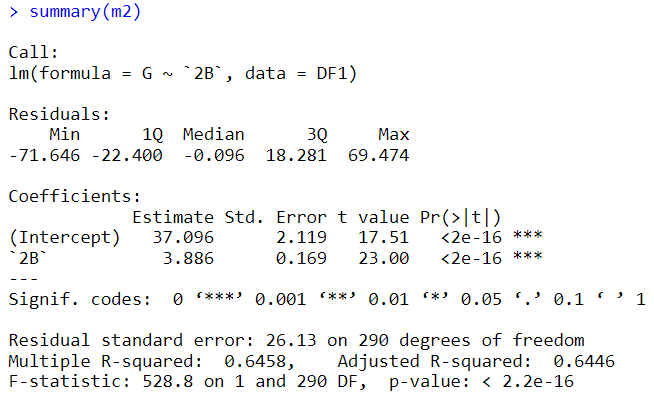
****

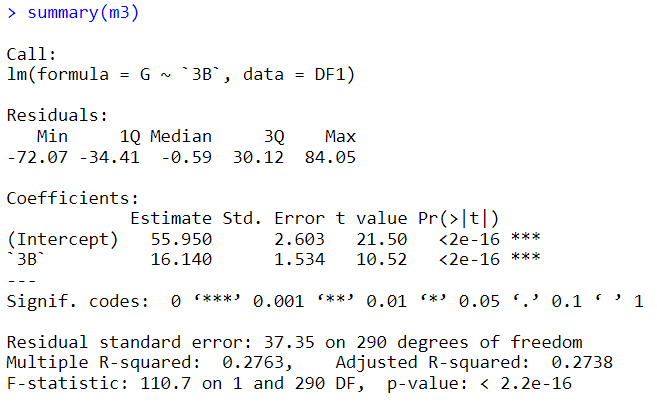
****

****

****

****

****

****

**문제 1 보고서 :**

**1) 반응변수 G와 설명변수 H의 회귀분석**

**#코드의 결과**

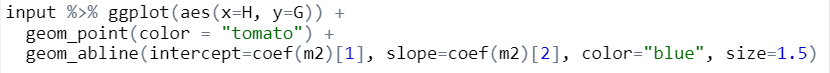
m1에서 B0=29.672, B1=1.096으로부터 G=29.672+1.096\*H 라는 회귀식이 도출된다.

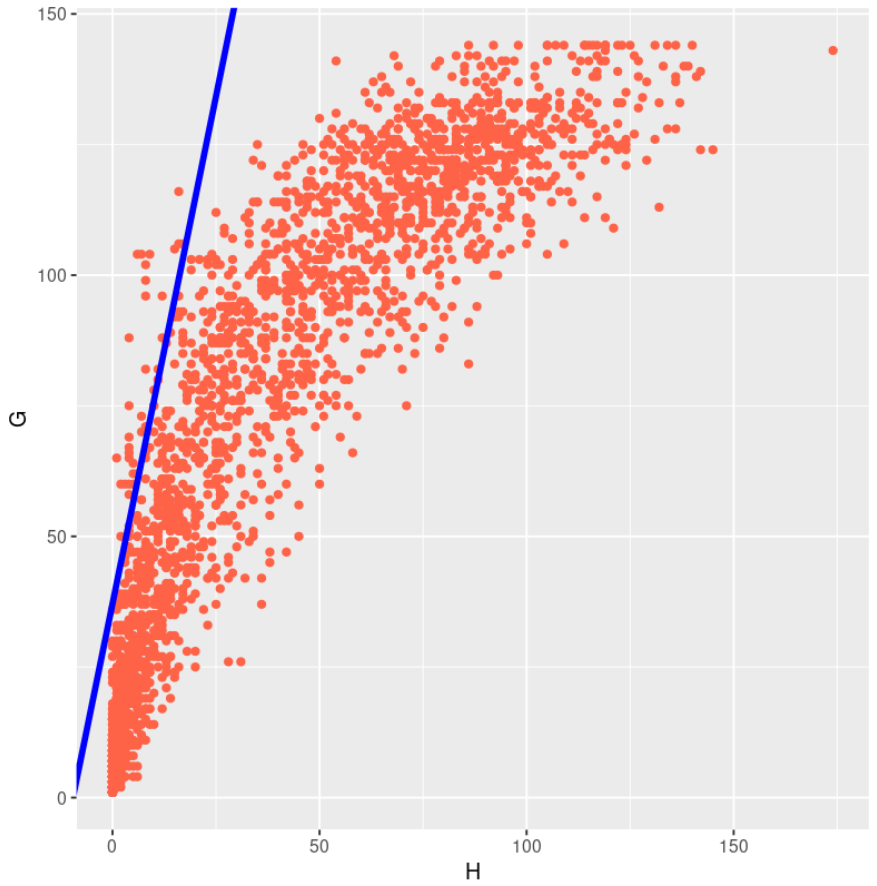
따라서 H(안타)가 1 증가할 때마다 G(출장수)는 1.096만큼 증가한다는 것을 예측할 수 있다.

summary(m1)의 출력결과에서 Residuals(잔차)는 잔차의 최솟값, 제1사분위수, 중앙값, 제3사분위수, 최댓값을 보여준다. 결과에서는 잔차 중앙값이 음수이므로 왼쪽으로 일정 부분이 기울어졌음을 알 수 있고, 제1사분위수의 값이 제3사분위수의 값보다 크기 때문에 데이터가 왼쪽으로 기울어졌다는 것을 알 수 있다.

**# 그래프로 나타내기**

그래프로 보면 G(출장수)가 증가하면 H(안타)도 증가한다는 사실을 알 수 있다.

****

****

**2) 반응변수 G와 설명변수 2B의 회귀분석**

**#코드의 결과**

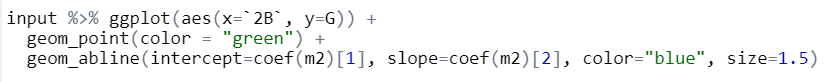
m2에서 B0=37.096, B1=3.886으로부터 G=37.096+3.886\*H 라는 회귀식이 도출된다.

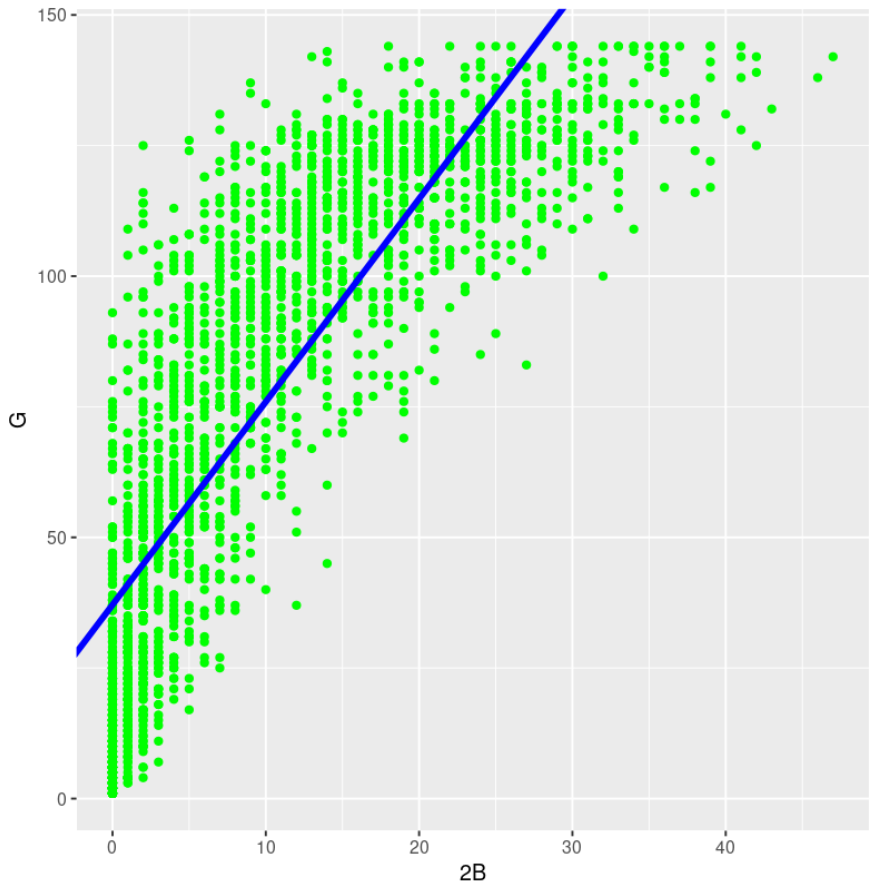
따라서 2B(2루타)가 1 증가할 때마다 G(출장수)는 3.886만큼 증가한다는 것을 예측할 수 있다.

summary(m2)의 출력결과에서 Residuals(잔차)는 잔차의 최솟값, 제1사분위수, 중앙값, 제3사분위수, 최댓값을 보여준다. 결과에서는 잔차 중앙값이 음수이므로 왼쪽으로 일정 부분이 기울어졌음을 알 수 있고, 제1사분위수의 값이 제3사분위수의 값보다 크기 때문에 데이터가 왼쪽으로 기울어졌다는 것을 알 수 있다.

**# 그래프로 나타내기**

그래프로 보면 G(출장수)가 증가하면 2B(2루타)도 증가한다는 사실을 알 수 있다.



****

**3) 반응변수 G와 설명변수 3B의 회귀분석**

**#코드의 결과**

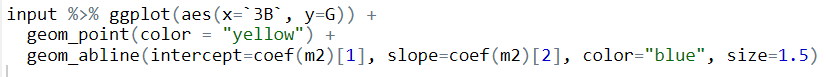
m3에서 B0=55.95, B1=16.14으로부터 G=55.95+16.14\*H 라는 회귀식이 도출된다.

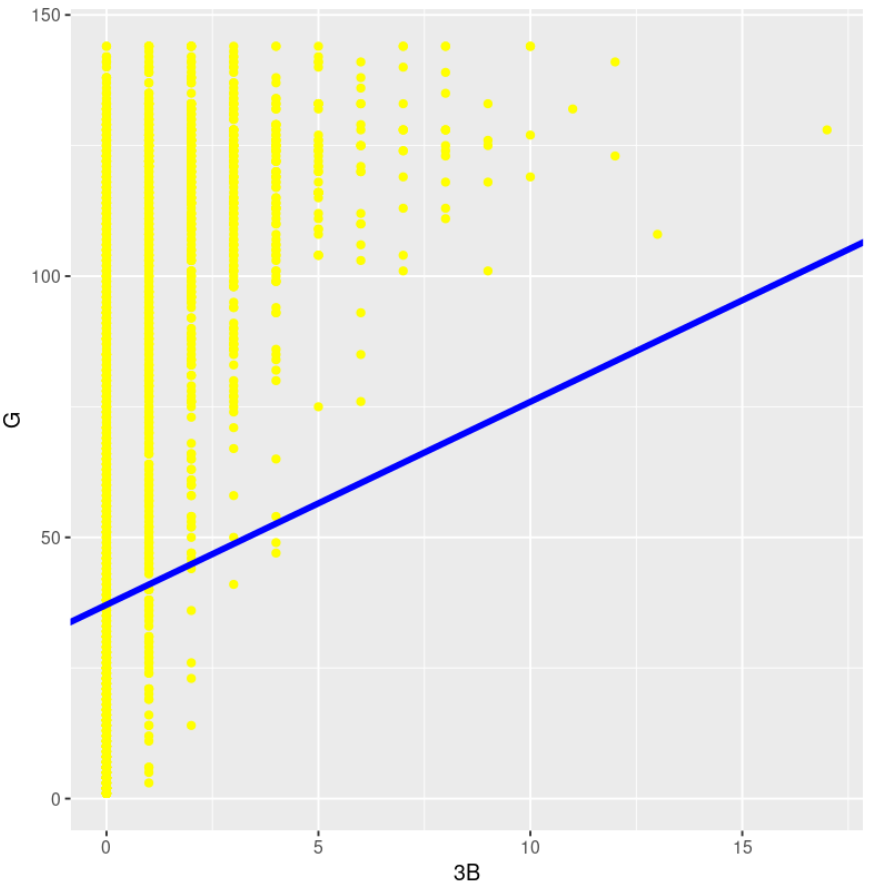
따라서 3B(3루타)가 1 증가할 때마다 G(출장수)는 16.14만큼 증가한다는 것을 예측할 수 있다.

summary(m3)의 출력결과에서 Residuals(잔차)는 잔차의 최솟값, 제1사분위수, 중앙값, 제3사분위수, 최댓값을 보여준다. 결과에서는 잔차 중앙값이 음수이므로 왼쪽으로 일정 부분이 기울어졌음을 알 수 있고, 제1사분위수의 값이 제3사분위수의 값보다 크기 때문에 데이터가 왼쪽으로 기울어졌다는 것을 알 수 있다.

**# 그래프로 나타내기**

그래프로 보면 G(출장수)가 증가하면 3B(3루타)도 증가한다는 사실을 알 수 있다.

****

****

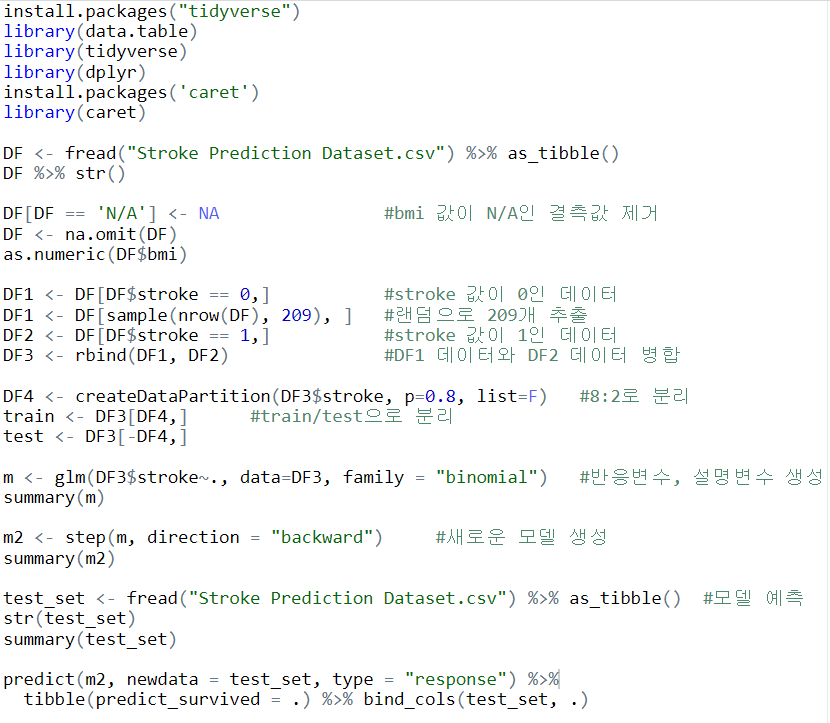
## <문제 2번>

|  |
| --- |
| [Q 2] **[70점] – glm을 이용한 로지스틱 회귀분석**  **<사용되는 데이터 – Stroke Prediction Dataset.csv>**  **Kaggle에서 제공하는 뇌졸증 예측 데이터셋을 이용하여 로지스틱 회귀분석을 해보자!**  https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset  **<참고사항>**   1. **Stroke Prediction 데이터를 로드 후 다음과 같은 전처리 과정을 거친다.**  * **결측값 제거(bmi 컬럼값이 N/A로 작성되어 있는 값을 결측값으로 지정함)** * **결측값 제거 후 bmi 데이터 타입 변경** * **Stroke 값이 0인 데이터를 랜덤으로 209개 추출** * **Stroke 값이 1인 데이터는 모두 사용** * **Stroke 값이 0인 추출된 데이터와 stroke 값이 1인 데이터 병합**  1. **병합한 Stroke Prediction 데이터를 caret 패키지의 createDataPartition을 이용하여 train / test 으로 분리할 것 (분리 비율은 본인이 지정한다)** 2. **Train dataset을 사용하여glm 함수를 사용하여 모델을 생성한다.**  * **반응변수는 stroke로 설정하고, 설명변수는 stroke를 제외한 모든 것을 사용하여 모델을 생성한다.** * **다음, Step(), update() 함수 등을 이용하여 생성한 모델의 적합도를 높인 새로운 모델을 생성해본다.** * **처음 생성된 모델과 새롭게 생성된 모델의 summary결과를 확인한다.**  1. **Test set을 사용해서 모델을 예측한다. 예측 결과와 원본 데이터와 비교하여 모델의 적합성에 대하여 기술한다.** 2. **모델 생성 및 모델 적합도에 대한 설명(50점), 예측(20점) – 각 과정은 이론, 실습 영상을 참고할 것.** |

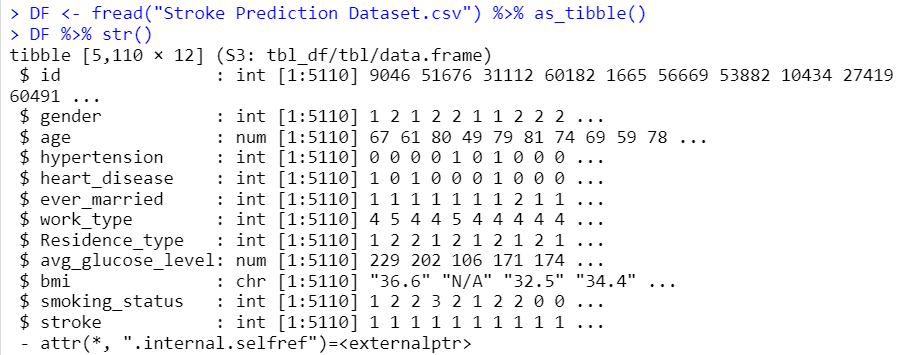
**답변: 출력 결과 및 설명(자유형식) 보고서 문서와 스크립트 파일 압축하여 제출.**

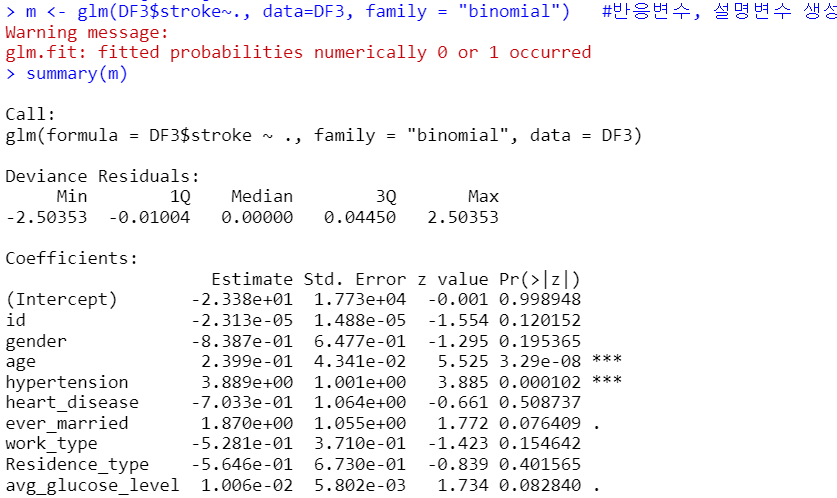
**<문제 2번>**

**코드 :**

****

**결과 :**

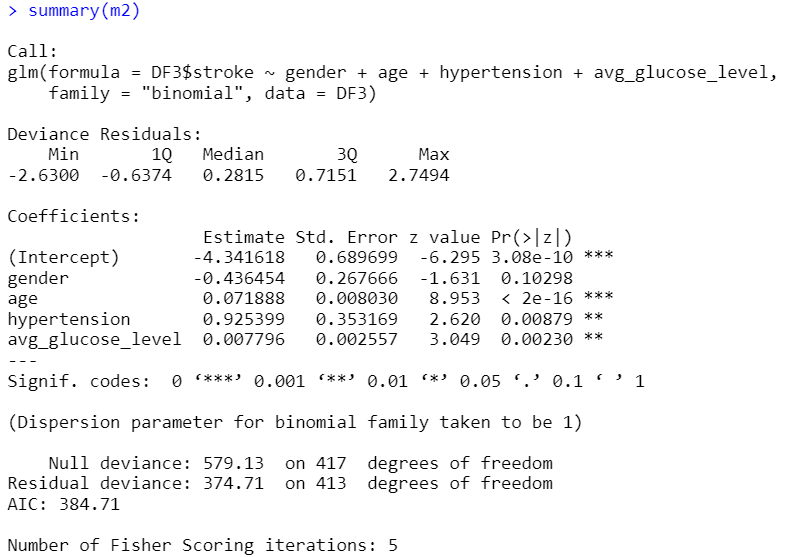
****

****

**.**

**.**

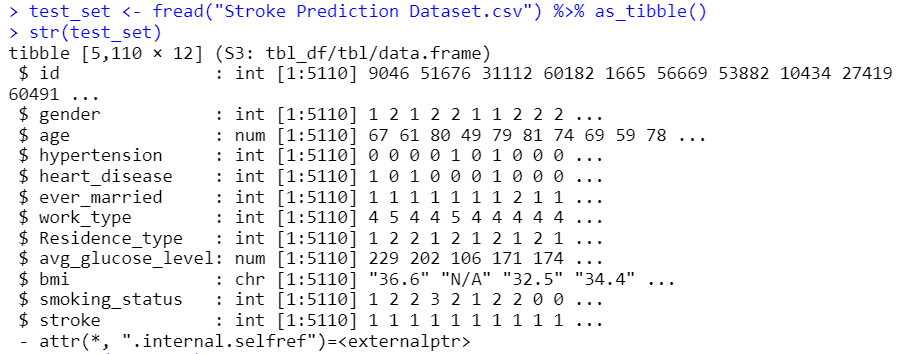
**.**

****

**.**

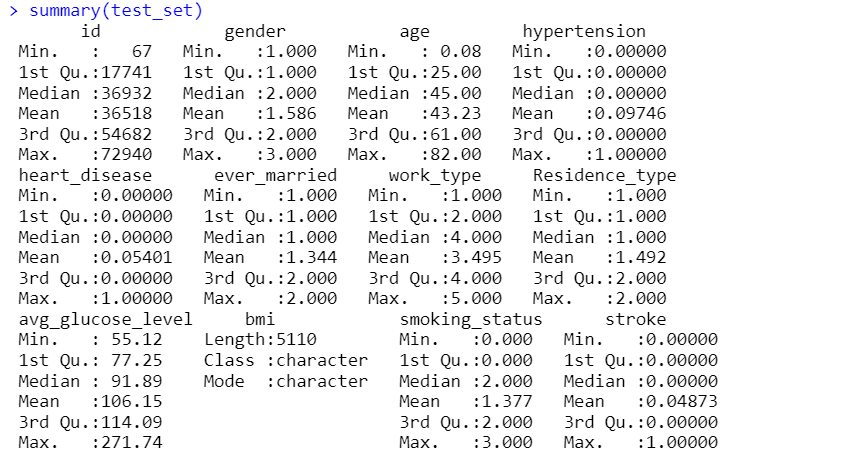
**.**

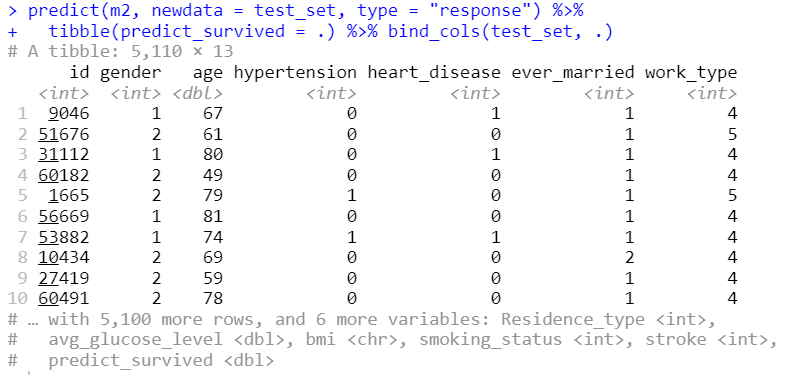
**.**

****

**.**

**.**

****

****

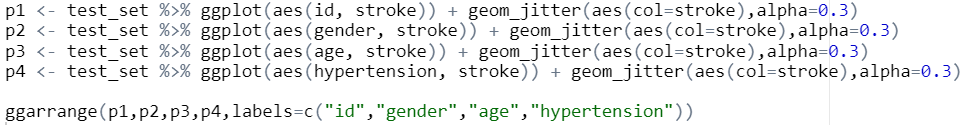
**문제 2 보고서 :**

**# m2 출력**

Pr부분에서 각 변수별로 P-value를 확인할 수 있고, 이에 대한 통계적인 유의성은 P-value의 오른쪽에 별표(\*)로 표시된다. 별표가 많을수록 통계적으로 더욱 유의하다는 것을 의미한다. 따라서, 별표가 없는 변수들은 보통의 유의성에 대해 크게 관여하지 않는다고 해석할 수 있다. 또한, 가장 유의한 변수는 age(나이)라는 것을 볼 수 있다.

R에서 type='response'를 설정함으로써, 특정한 입력 변수를 넣었을 때의 출력 변수가 될 확률을 계산해준다. 위 코드처럼 predict() 함수를 사용해서 결과를 출력했을 경우 **'id가 9046, age가 67이고 gender는 1(남자)는 hypertension(고혈압 바이너리 기능)은 0이다'** 라는 것을 예측할 수 있다. 또한, **'id가 1665 age가 79이고 gender는 2(여자)는 hypertension(고혈압 바이너리 기능)은 1이다'** 라는 것을 예측할 수 있다.

**# 그래프로 나타내기**

****

