**9/3 과제: 상관분석, 회귀분석, 다중회귀, k-인접기법**

**데이터 설명: Heart Disease UCI (heart.csv)**

age : age in years

sex : (1 = male; 0 = female)

cp : chest pain type

trestbps : resting blood pressure (in mm Hg on admission to the hospital)

chol : serum cholestoral in mg/dl

fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg : resting electrocardiographic results

thalach : maximum heart rate achieved

exang : exercise induced angina (1 = yes; 0 = no)

oldpeak : ST depression induced by exercise relative to rest

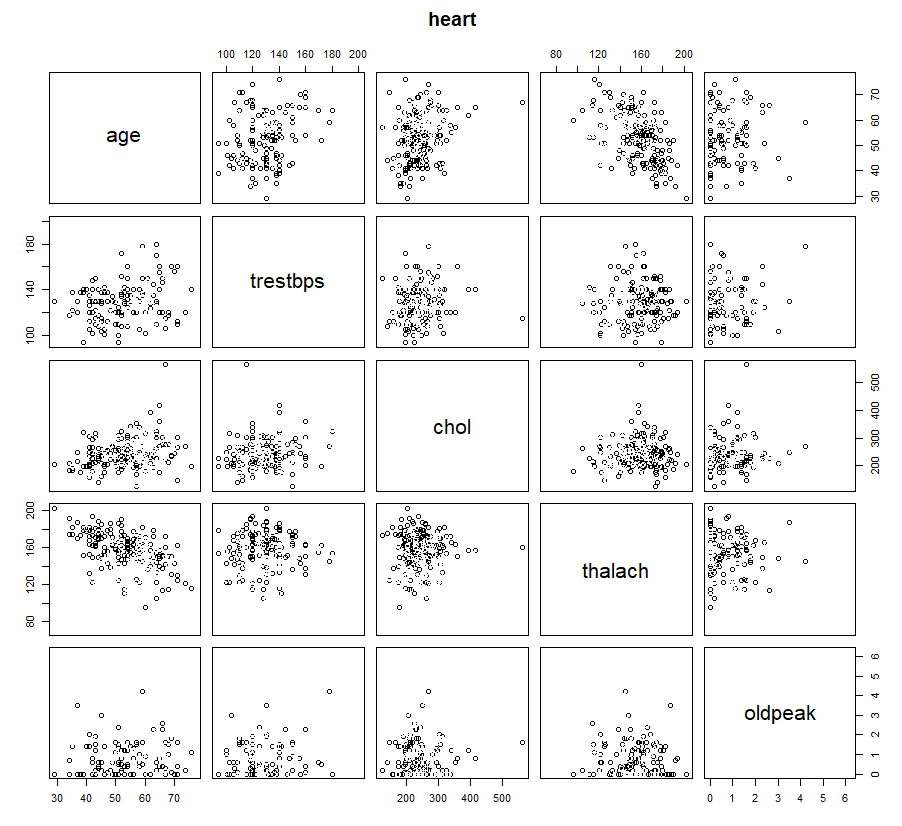
slope : the slope of the peak exercise ST segment

ca : number of major vessels (0-3) colored by flourosopy

thal : 3 = normal; 6 = fixed defect; 7 = reversable defect

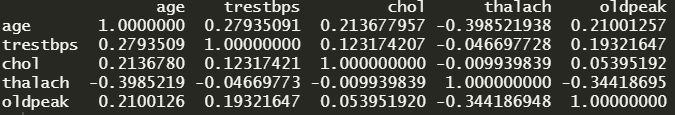
target : have disease or not (1=yes, 0=no)

1. **상관 분석**
   1. **pairs함수를 사용해서 변수 age, trestbhs, chol, thalach, oldpeak에 대해서 산점도 그래프를 그리고 해석하라.**

****

**특별히 강한 상관관계를 보이는 그래프는 보이지 않지만, 그 중에서 age와 thalach이 음의 상관관계를 가지는 것으로 보인다.**

**1.2. 1.1에서 그린 그래프를 통하여 양의 상관계수와 음의 상관관계가 가장 큰 경우에 대해서 말하고 각각 경우에 대해 상관계수 값을 말하라**

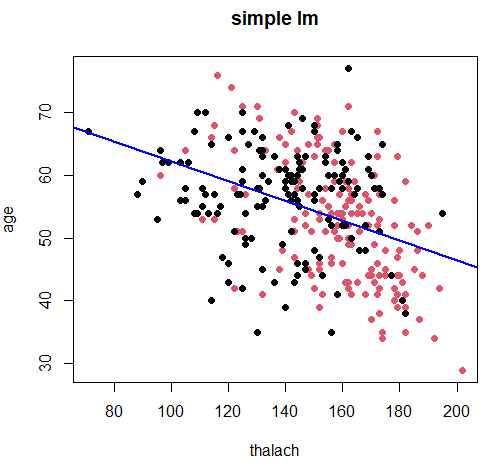
****

**가장 큰 양의 상관계수: age & trestbps 0.27935091**

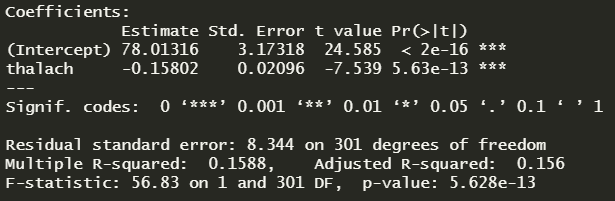
**가장 큰 음의 상관계수: age & thaldch -0.398521938**

**2. 회귀 분석 및 다중회귀**

**2.1 변수 thalach를 x축으로 두고 age에 대하여 단일 회귀 모델을 만들고 그래프로 그려라. (회귀선도 그려라)**

****

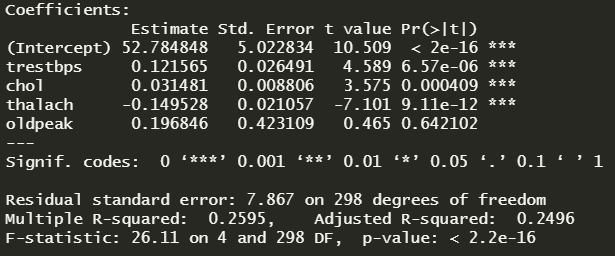
**2.2 단일 회귀 모델의 선형 회귀식과 결정 계수는?**

****

**선형회귀식: age = -0.15802 thalach + 78.01316**

**결정계수: 0.1588**

**2.4 변수 age를 y축으로 두고 1.1에서 사용했던 변수들 중 age를 빼고 나머지를 이용해서 다중 회귀 모형을 만들어라. 선형 회귀식과 결정계수는?**

****

**선형회귀식: age = 52.8 + 0.12 trestbps + 0.03 chol – 0.15 thalach + 0.2 oldpeak**

**결정계수: 0.2595**

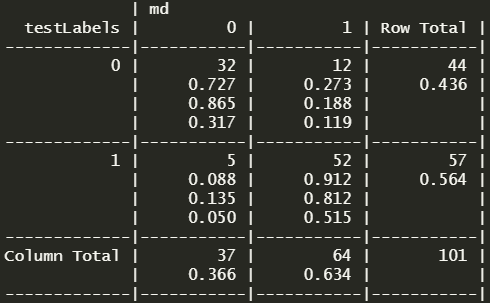
**2.5 단일 회귀 모델과 다중 회귀 모델을 비교 분석하라.**

**결정계수는 당연하게도 다중 회귀 모델이 더 높으며, 더 잘 설명한다고 볼 수 있다.**

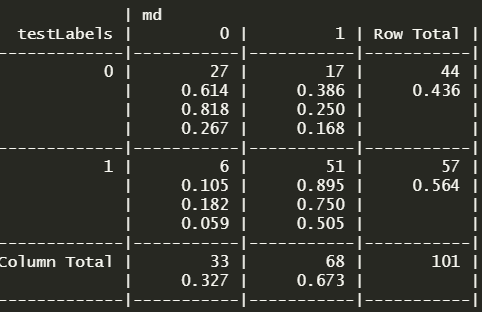
1. **데이터 분할 및 k-인접 기법**
   1. **heart데이터의 2/3을 학습 데이터셋, 1/3을 테스트 데이터 셋으로 분할 하라.(seed값은 1000으로)**
   2. **3.1의 학습 데이터셋을 이용하여 k=3,5,10으로 두고 knn모델을 수행한 뒤 CrossTable을 출력하라. k를 어떤 값을 사용했을 때 가장 좋은 모델인가?(각각의 정확도와 오분율을 가지고 설명해라) 그 이유는?**

**K=3 일 때, 정확도가 가장 높으므로 가장 좋은 모델이다.**

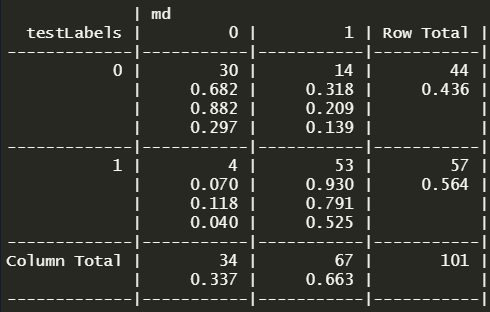
**K=3 (정확도: 83.1%, 오분율: 16.8%)**

****

**K=5 (정확도: 77.2%, 오분율: 22.8%)**

****

**K=10 (정확도: 82.2%, 오분율: 17.8%)**

****