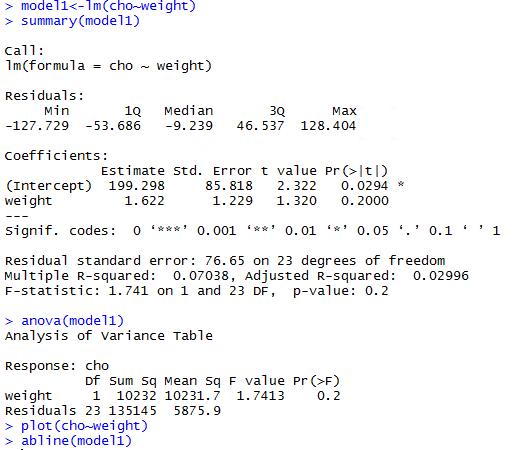
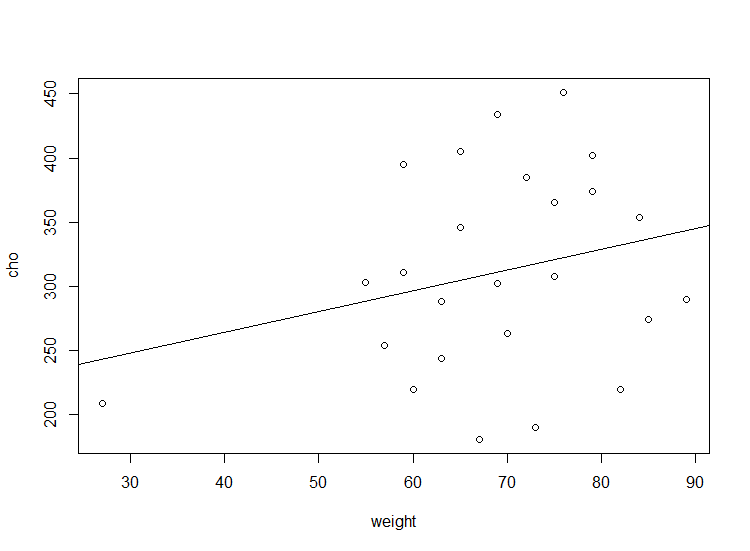
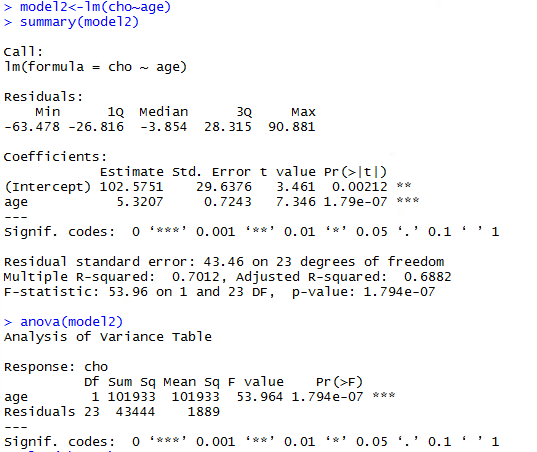
**3. HLPE**

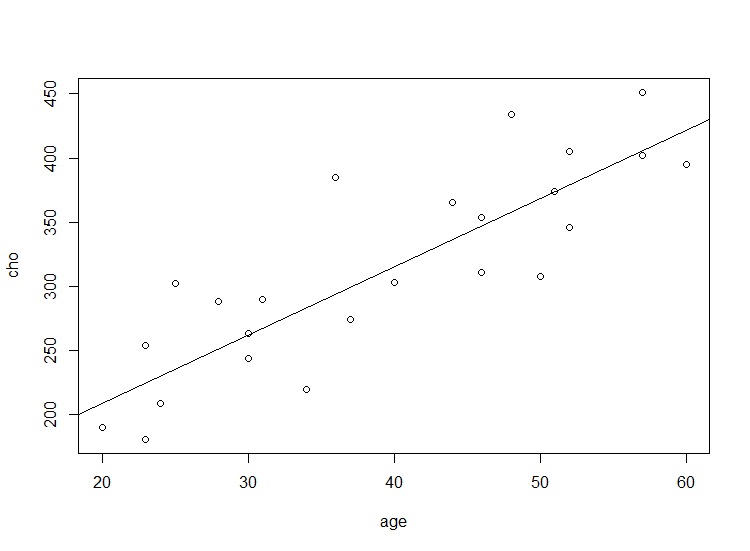
**(a) model 1:**

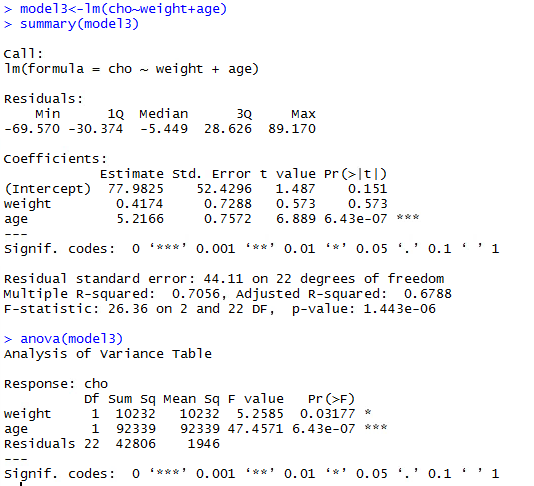




**model 2:**





**model 3:**

Summary of regressions: with unit “mg/100 ml”

model 1:

= 199.298 + 1.622 × (Weight)

Total SS is 10232+135145 = 14537; error SS is 135145.

model 2:

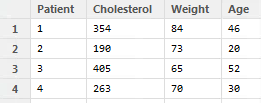
= 102.5751 + 5.3207 × (Age)

Total SS is 101933+43444 = 145377; error SS is 43444.

model 3:

= 77.9825 + 0.4174 × (Weight) + 5.2166 × (Age)

Total SS is 10232+92339+42806 = 145377; error SS is 42806.

**(b) Predicted value:**

The observed value is 263(mg/100 ml).

Model 1:

= 199.298 + 1.622 × (70) = 312.838 (mg/100 ml)

Model 2:

= 102.5751 + 5.3207 × (30) = 262.1961(mg/100 ml)

Model 3:

= 77.9825 + 0.4174 × (70) + 5.2166 × (30) = 263.6985(mg/100 ml)

In general, the 2nd and 3rd model are good fits for predict the cholesterol. The diffrences between the obs. value and the predicted value are 0.8039 and 0.6985, which are very small value. For the 1st model, the predicted value is 49.838 higher than the obs. value, which indicates that the 1st model is not a good fit.

**(c) R2:**

Model 1: R2 = 0.07038 = 7.04%

Model 2: R2 = 0.7012 = 70.12%

Model 3: R2 = 0.7056 = 70.56%

R-squared value represents the percentage of the variability in cholesterol is accounted for by variation in the explanatory variables. To get a good fit of regression, we are looking for a higher percentage of variability. In the above three models, the 2nd and the 3rd models are having relatively higher variability. Meanwhile, since we need to take the simplicities of the models into concern, we say that the 2nd model is the best fit of our regression model because it has a high variability (70.12%) in cholesterol by difference in age and it only need one explanatory variable (age).

**4. Cancer mortality rate**

**(a) ANOVA table:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Degree of freedom | Sum of Squares | Mean Square |
| Regression | 3 | 551.723 | 183.9077 |
| Residual | 19 | 1835.93 | 96.6279 |
| Total | 22 | 2387.653 | - |

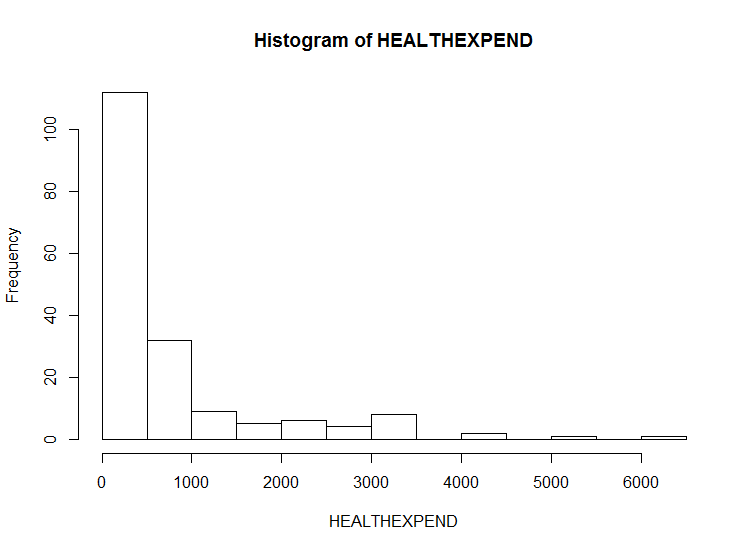
**(b) R2:**

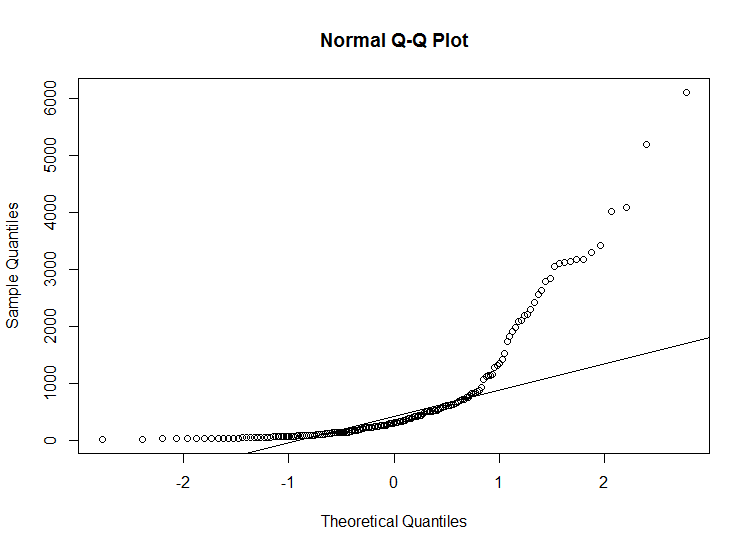
R2 = 551.723/2387.653 = 0.23107

The R-squared value indicates that 23.107% of the variability in the cancer mortality rate is accounted for by variation in the pollution lv., mean age, and percentage of people working in the particular industry. This coefficient determination is low, and therefore, weak evidences are provided to say that this model is a good fit. In all, the explanatory variables are not linearly related to the cancer mortality rate.

**5. Life Expectancy**

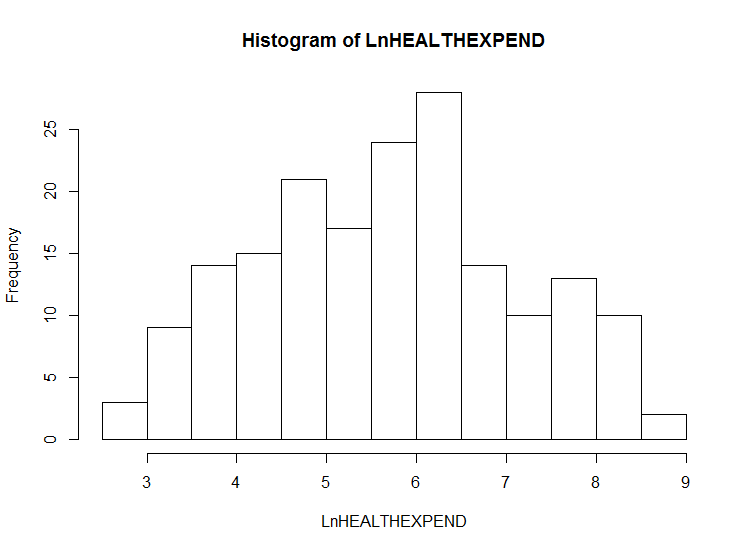
**(a) Histogram & qq-plot:**

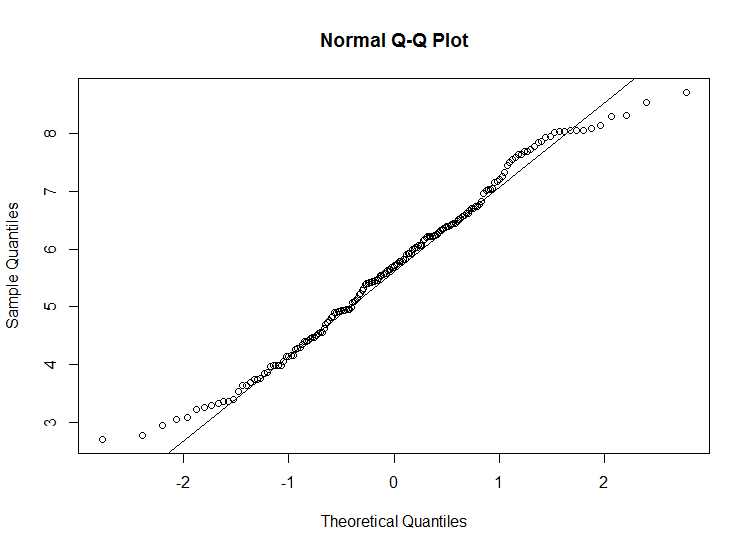




The histogram of HEALTHEXPEND shows a distribution that is skewed to the right. The distribution has a higher concentration on the small values, especially on the [0, 500] interval. The qq-plot shows that there’re many points away from the line, which means that the distribution is not normal.

Both of the plots imply that the distribution of HEALTHEXPEND is not symmetric and is not normal distributed.

**(b) Log transformation:**



The histogram of LnHEALTHEXPEND shows a fairly normal and symmetric distribution. The distribution has higher concentration in the middle of the histogram plot. The qq-plot agrees with the previous statement. The points are close to the line, which indicates the normality and the symmetry. We can say that log transformation does a good job to normalize the distribution.