STA206 ASS2

Zhen Zhang

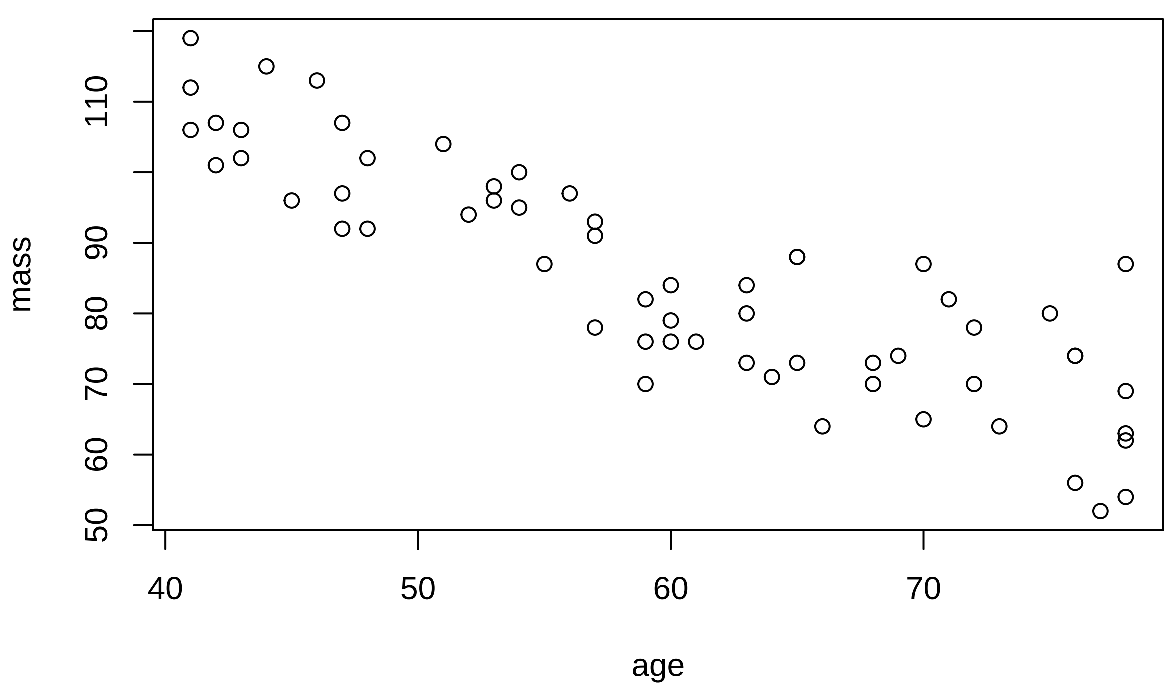
October 9, 2015

(a) First read the data into R, and then give the data the column names

```
muscle <- read.table("~/Github/UCDavis/STA206/Data/muscle.txt")
names(muscle) <- c("mass", "age")
```

Next I use plot function to draw the scatterplot:

```
with(muscle, plot(age, mass))
```



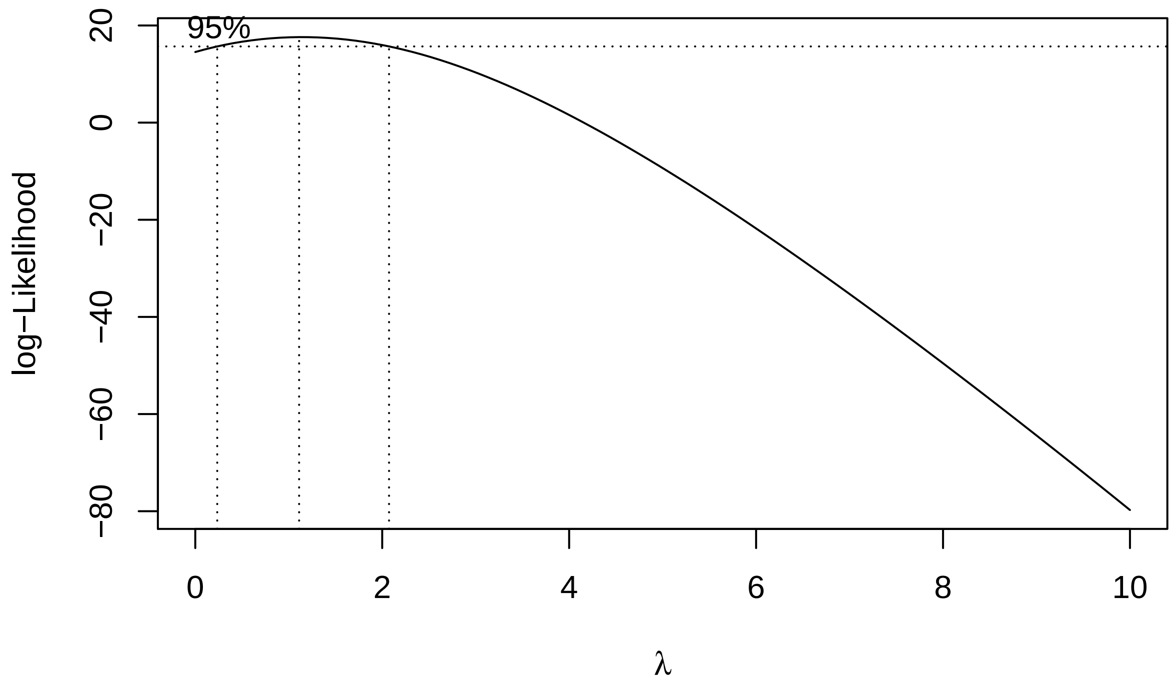
From the scatterplot, I can see a roughly tendancy that the amount of muscle mass decreases with age.

(b) First I need to load the MASS package which includes the box-cox function.

```
library(MASS)
```

Then use the box-cox procedure

```r
boxcox(mass ~ age, data = muscle, lambda = seq(0, 10, 1))
```



From the plot, I see that a transformation is not needed: order 1, which is linear, is best to fit the data.

(c) Let me run the linear regression:

```r
lmfit <- lm(mass ~ age, data = muscle)
```

Now display the summary

```r
summary(lmfit)
```

```
##
## Call:
## lm(formula = mass ~ age, data = muscle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## age          -1.1900     0.0902  -13.19   <2e-16 ***
```

```
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

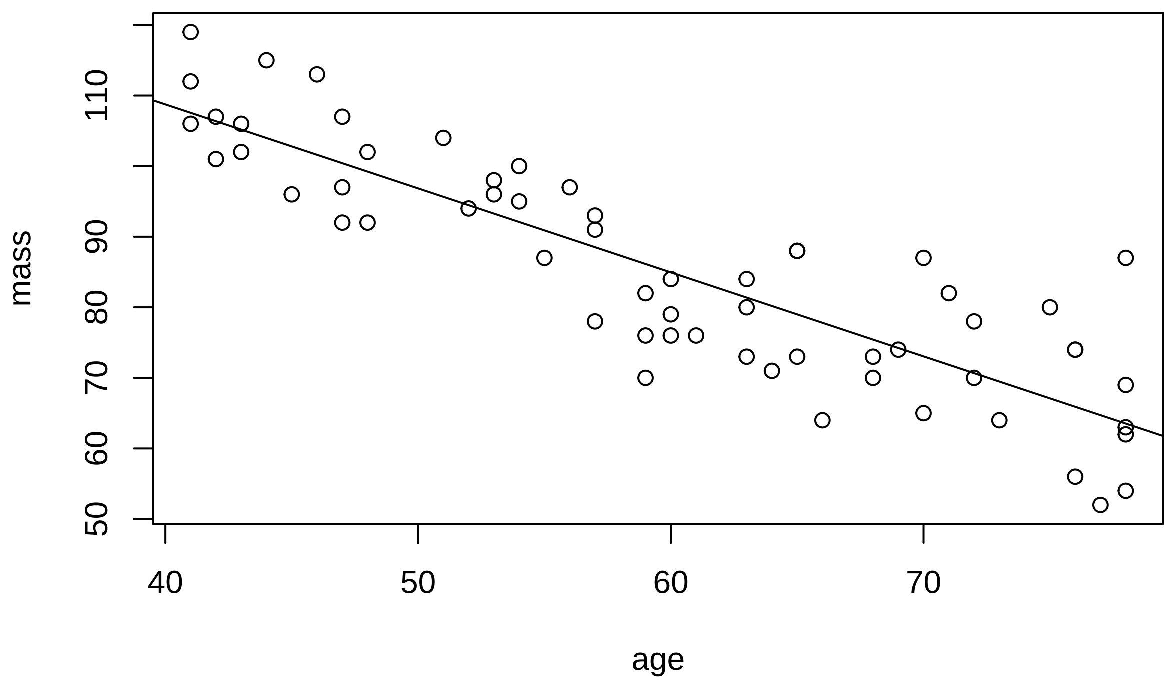
The coefficients are $\hat{\beta}_0$ = 156.3466, $\hat{\beta}_1$ = -1.1900. The standard error for $\hat{\beta}_0$ is 5.5123, for $\hat{\beta}_1$ is 0.0902. MSE is 8.173, degree of freedom is 58.

(d) The regression line is:

$$mass_i = 156.3466 - 1.1900 * age_i$$

Now add it to the scatterplot:

```r
with(muscle, plot(age, mass))
abline(lmfit)
```



The line fits the data well.

(e) The fitted values and residuals for the 6th and 16th cases in the data set are:

```r
lmfit$fitted.values[c(6, 16)]
```

```
##         6        16
## 107.55675  90.89681
```

```r
lmfit$residuals[c(6, 16)]
```

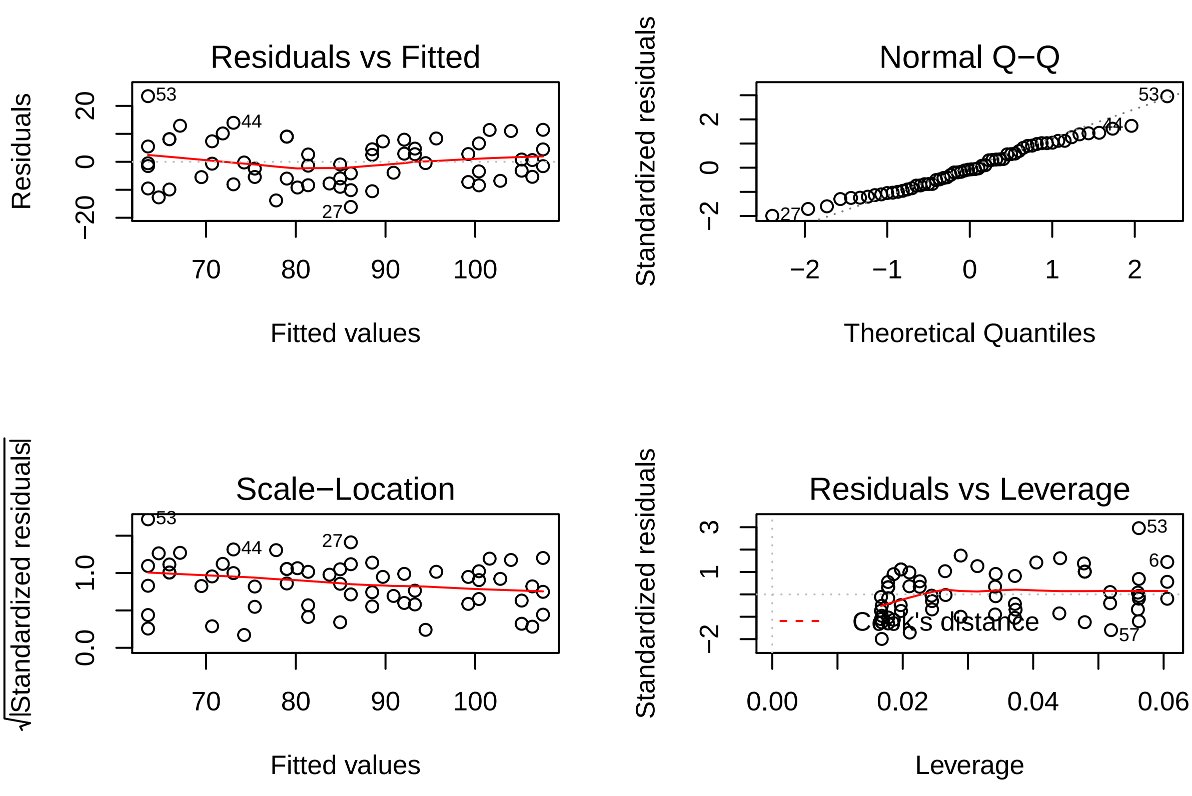
```
##         6        16
## 11.443252 -3.896811
```

The fitted value and residual for the 6th case are 107.55675 and 11.443252 respectively.

The fitted value and residual for the 6th case are 90.89681 and -3.896811 respectively.

(f) Here we will see four plots, and the first one is the residuals vs. fitted values plot, the second one is the residuals Normal Q-Q plot.

```r
par(mfrow = c(2, 2))
plot(lmfit)
```



The simple linear regression model with Normal errors:

$$mass_i = 156.3466 - 1.1900 * age_i + \epsilon_i$$

Assumptions:

- The error terms ϵ_i are independent, identical normal distributed (i.i.d) $N(0, \sigma^2)$ random variables.
- The relationship is linear.

Comment: The linear relationship holds, but the normal distribution is not valid. It is light tailed than normal distribution.

(g) first I need to know the t value for two-tailed 99% with degree of freedom 58:

```
t1 <- qt(0.995, 58)
```

So the intercept is:

$$[\hat{\beta}_0 - t(0.995, 58) * se(\hat{\beta}_0), \hat{\beta}_0 + t(0.995, 58) * se(\hat{\beta}_0)]$$

evaluated at:

$$[141.6657, 171.0274]$$

(h) Null hypothesis (H_0): There is no linear association between the amount of muscle mass and age.

Alternative hypothesis (H_1): There is a negative linear association between the amount of muscle mass and age.

The test statistic: T-statistic: $T^* = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$

The null distribution: Under H_0, $\beta_0 = 0$, $T^* \sim t(58)$

The decision rule: Reject H_0 if and only if $T^* < t(0.01, 58)$

The conclusion: the t statistic here is $t(0.01, 58) =$ -2.392377, and T^* is -13.1929, so reject H_0, meaning there is a negative linear association between the amount of muscle mass and age.

(i) use the function predict in r:

```
predict(lmfit, data.frame(age = 60), interval = "predict")
```

```
##        fit      lwr     upr
## 1 84.94683 68.45067 101.443
```

So the prediction interval is $[68.45067, 101.443]$.

Interpretation: if I sample 60 women from the dataset, then the muscle mass with age 60 is 95% probability falls into the interval [68.45067, 101.443].

(j) Test for anova:

```
anovafit <- anova(lmfit)
```

Null hypothesis (H_0): There is no linear association between the amount of muscle mass and age.

Alternative hypothesis (H_1): There is a linear association between the amount of muscle mass and age.

The test statistic: F-statistic: $F^* = \frac{MSR}{MSE}$

The null distribution: Under H_0, $F^* \sim F(0.99, 1, 58)$

The decision rule: Reject H_0 if and only if $F^* > F(0.99, 1, 58)$

The conclusion: the t statistic here is $F(0.99, 1, 58) = 7.093097$, and F^* is 174.06, so reject H_0, meaning there is a linear association between the amount of muscle mass and age.

(k) let's see

```r
summary(lmfit)
```

```
##
## Call:
## lm(formula = mass ~ age, data = muscle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## age          -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

Since the r^2 is 0.7501, then 75.01% of the total variation in muscle mass is "explained" by age.

The correlation coefficient between muscle mass and age is $-\sqrt{0.7501} = -0.866$.