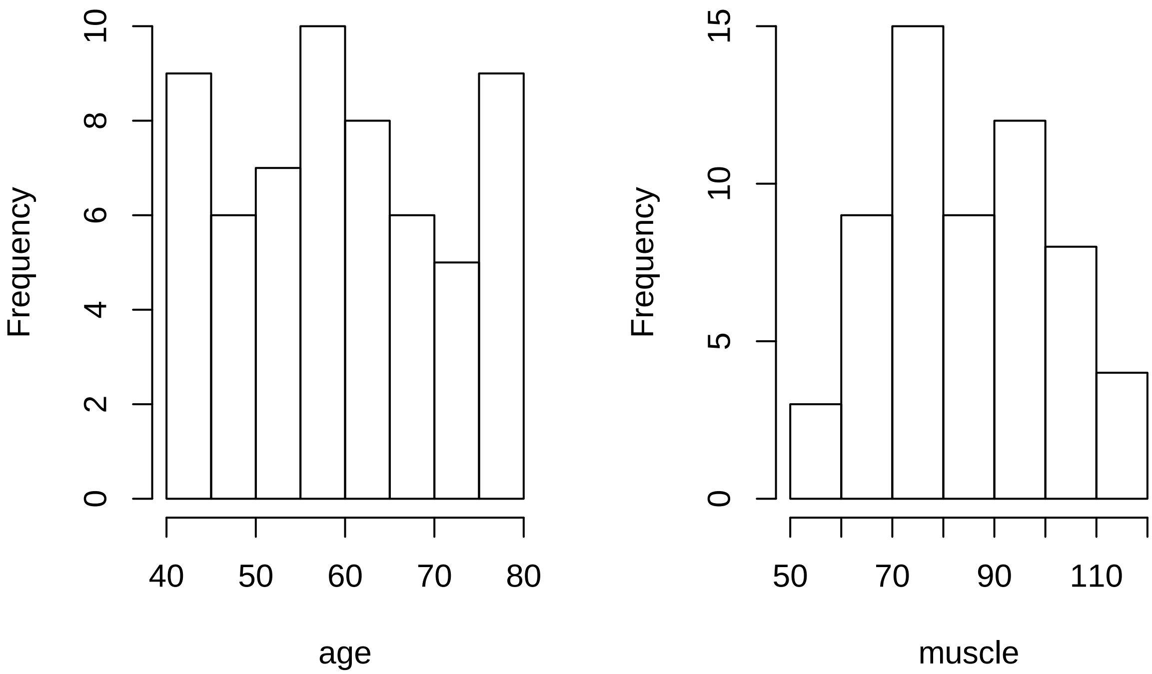
Stats 206 Homework 2

Clark Fitzgerald

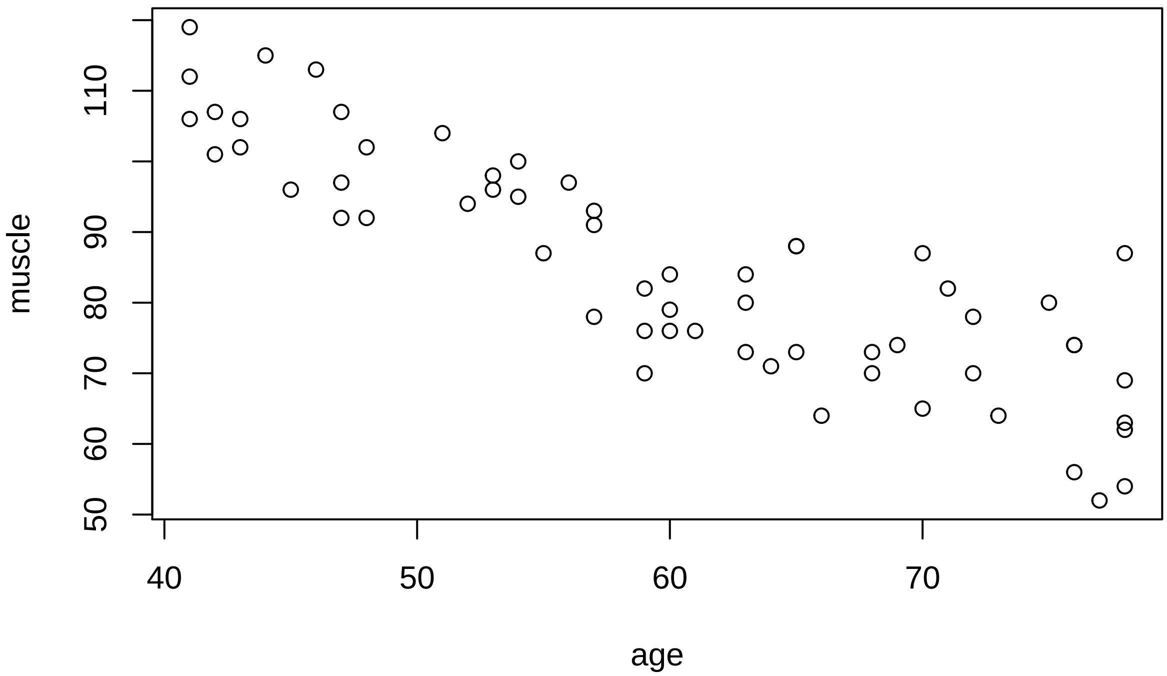
October 16, 2014

1a

The histograms of muscle and age are unsurprising. The ranges are as expected. Age appears to be roughly uniform and muscle has a bell shape.



The relation between age and muscle appears to be linear, and it looks like muscle mass decreases with age.



1b

After fitting the linear model we extract regression coefficients with their standard errors, the mean squared error (MSE), and its degrees of freedom.

Regression coefficients:

```
simple = lm(muscle ~ age, data=women)
b0 = simple$coefficients[1]
b1 = simple$coefficients[2]
c(b0, b1)
```

```
## (Intercept)         age
##  156.346564   -1.189996
```

Standard errors for regression coefficients:

```
ss = summary(simple)
ss$coefficients[, 'Std. Error']
```

```
## (Intercept)         age
##  5.51226249  0.09019725
```

The MSE and its degrees of freedom are:

```
a = anova(simple)
mse = a['Residuals', 'Mean Sq']
df = a['Residuals', 'Df']

c(mse, df)
```

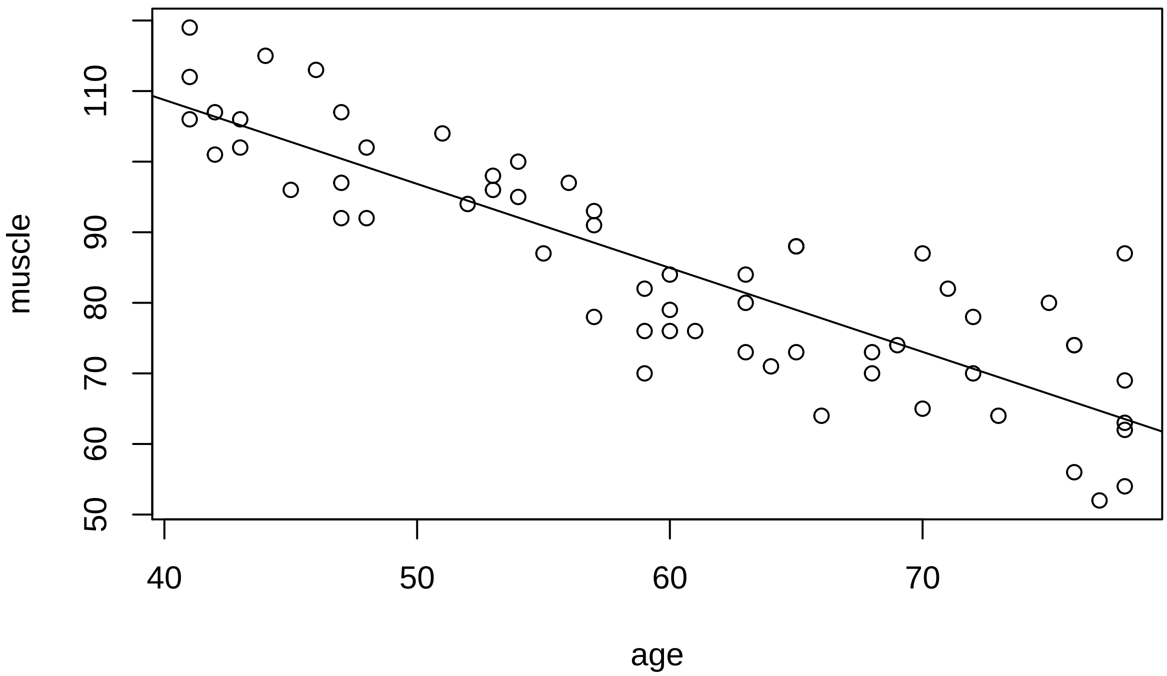
```
## [1] 66.80082 58.00000
```

1c

The fitted regression line is:

```
sprintf('muscle = %.2f + %.2f * age', simple$coefficients[1],
        simple$coefficients['age'])
```

```
## [1] "muscle = 156.35 + -1.19 * age"
```



It looks like the linear regression is a good fit.

1d

The fitted values for the 6th and 16th cases are:

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

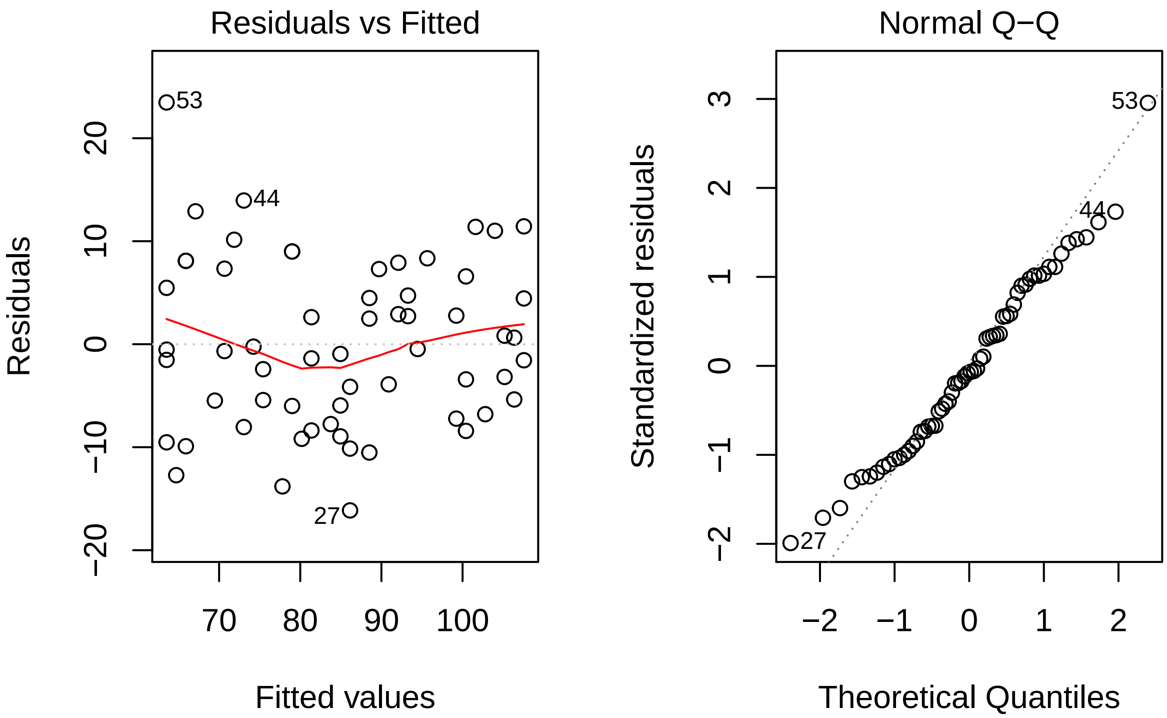
```
##         6        16
## 107.55675  90.89681
```

The residuals for the 6th and 16th cases are:

```
simple$residuals[c(6, 16)]
```

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

1e



Using linear algebra notation, the simple linear regression model with Normal errors assumes

- $y = X\beta + \epsilon$ The response is a linear function of the predictors.
- $\epsilon \sim Normal(0, \sigma^2 I_n)$ The error terms are normally distributed and uncorrelated.

The graphs support the assumptions of the model.

1f

A 99 percent confidence interval for the estimated regression intercept is:

```
##                 0.5 %   99.5 %
## (Intercept) 141.6658 171.0273
```

We are 99 percent confident that the the true parameter lies within this interval.

1g

We test at level 0.01 to see if there is a negative linear association between muscle mass and age. H_0 is $\beta_1 = 0$ and H_1 is the left sided alternative hypothesis $\beta_1 < 0$.

The test statistic is $T^* = \frac{\hat{(\beta_1)}-0}{se(\hat{(\beta_1)})} \sim t(n-2)$.

The decision rule is to reject H_0 if $T^* < t(0.99, n-2)$.

```
## [1] "If -13.193 is less than -2.392 we reject H0"
```

We reject the null hypothesis and conclude that there is a significant negative linear association between amount of muscle mass and age.

1h

A 95% prediction interval for the muscle mass for women of age 60 is:

```
predict(simple, data.frame(age=60), interval='prediction', level=0.95)
```

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

The fit is the expected value. We expect 95% of new observations to fall between the lower and upper bounds.

1i

The limits of a the 95% simultaneous confidence bands for the regression line at $x_h = 60$ are:

```
p60 = predict(simple, data.frame(age=60), se.fit=TRUE,
    interval='prediction', level=0.95)

# Actual fitted value
fit60 = p60$fit[1]
se60 = p60$se.fit

# Working-Hotelling multiplier
W = sqrt(2 * qf(0.95, 2, df))

c(fit60 - W * se60, fit60 + W * se60)
```

```
## [1] 82.29593 87.59774
```

1j

The ANOVA table for this data is:

```
## Analysis of Variance Table
##
## Response: muscle
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## age        1 11627.5 11627.5  174.06 < 2.2e-16 ***
## Residuals 58  3874.4    66.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We use an F test at level 0.01 to see if there is a linear association between muscle mass and age. H_0 is $\beta_1 = 0$ and H_1 is the two sided alternative hypothesis $\beta_1 \neq 0$.

The test statistic is $F^* = \frac{\hat{(\beta_1)}-0}{se(\hat{(\beta_1)})}\ t(n-2)$.

The decision rule is to reject H_0 if $F^* > F(0.99; 1, n-2)$.

```r
f99 = qf(0.99, 1, n-2)
msr = sum(simple$residuals ** 2) / n
fstar = msr / mse

sprintf('If %.3f is less than %.3f we reject H0', fstar, f99)
```

```
## [1] "If 0.967 is less than 7.093 we reject H0"
```

We reject the null hypothesis and conclude that there is a significant linear association between amount of muscle mass and age.

1k

The proportion of total variation in muscle mass explained by age is R^2:

```r
summary(simple)$r.squared
```

```
## [1] 0.7500668
```

The correlation between muscle mass and age is:

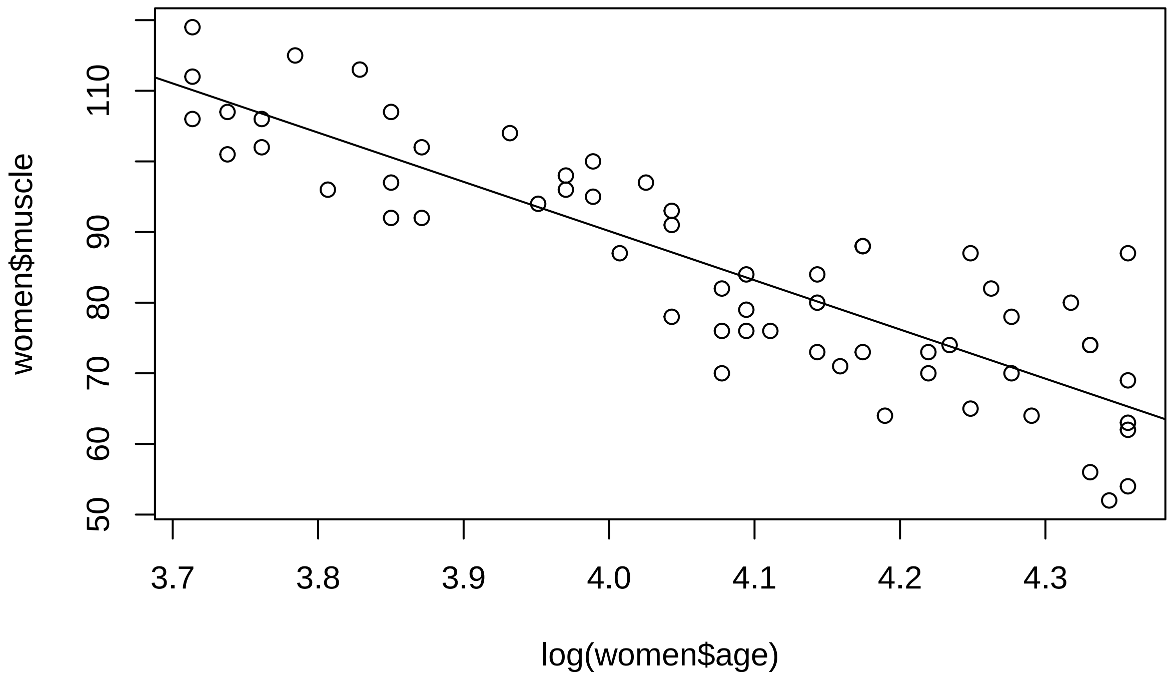
```r
cor(women$muscle, women$age)
```

```
## [1] -0.866064
```

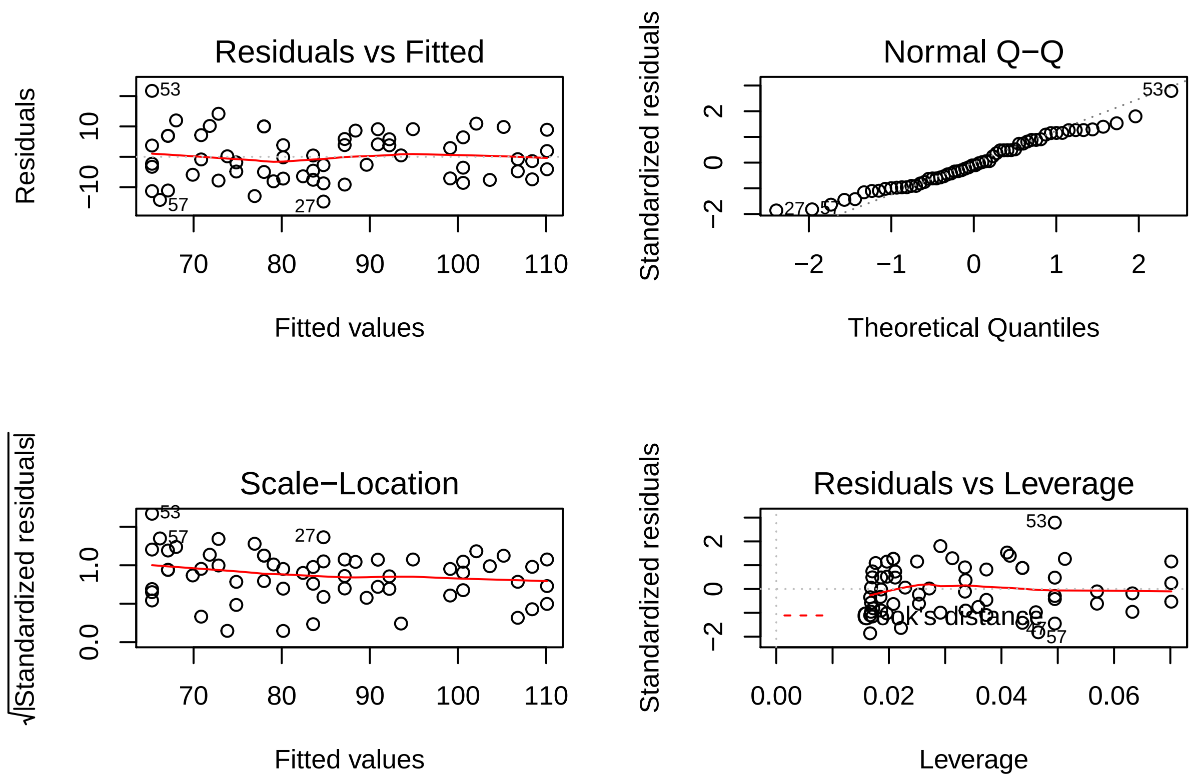
1l

We fit the model using the log of age.

```
##
## Call:
## lm(formula = muscle ~ log(age), data = women)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7382  -6.5901  -0.8211   6.5403  21.7113
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  368.816     20.894   17.65   <2e-16 ***
## log(age)     -69.669      5.122  -13.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.987 on 58 degrees of freedom
## Multiple R-squared:  0.7613, Adjusted R-squared:  0.7572
## F-statistic:   185 on 1 and 58 DF,  p-value: < 2.2e-16
```



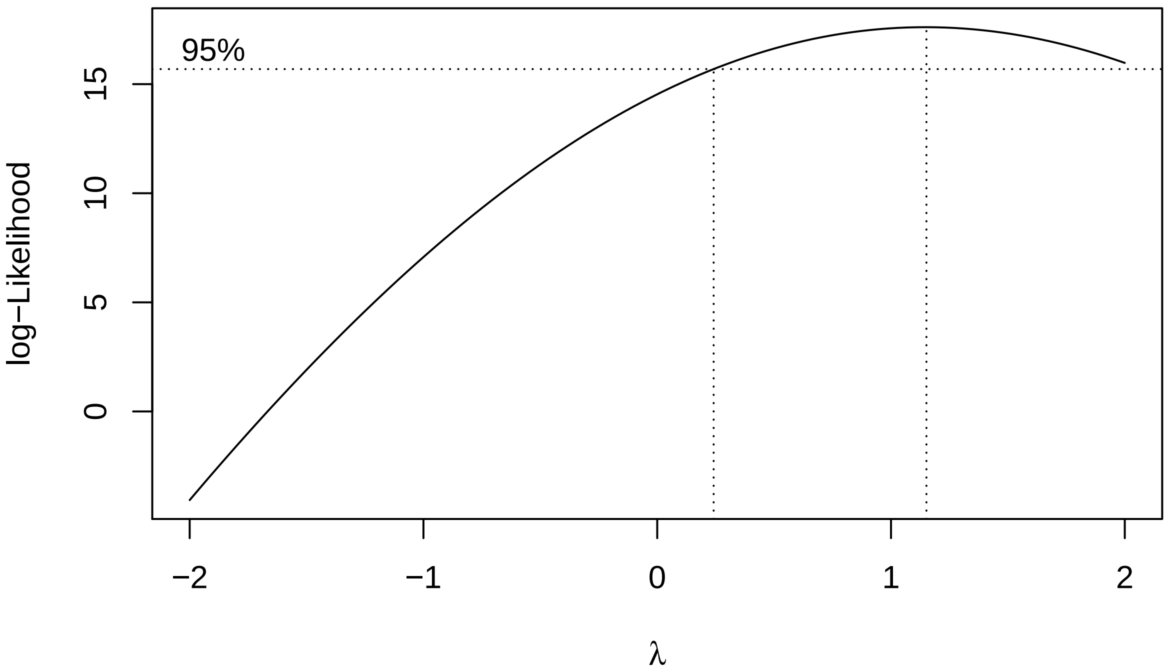
This model has the corresponding residual plots:



The fit looks similar to the first model.

1m

We plot the Box-Cox power transformation.



This suggests that a value of $\lambda = 1$ is appropriate. In other words, no transformation is required.