Homework 3

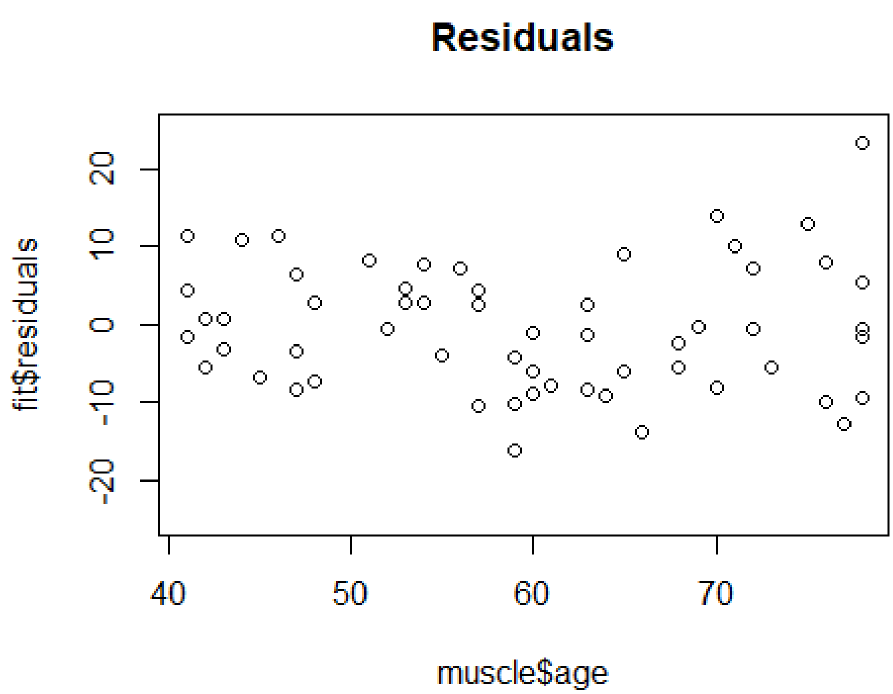
Homework 3 Blake Tindol & Sydney Haase

Q1 a) What is the predicted value and 95% confidence interval for the mean muscle mass for women of age 60?

```
fit <- lm(mass ~ age, data = muscle)
new.data = data.frame(age = 60)
newpred <- predict(fit, newdata =new.data,interval = 'predict',level = 0.95)
newpred

##        fit      lwr     upr
## 1 84.94683 68.45067 101.443

# 95% of predicted values  fall between 68.45 and 101.44
# 84.95 is the mean predicted muscle mass for womens age of 60  95% of time
is going to be between 68.45 and 101.443
```

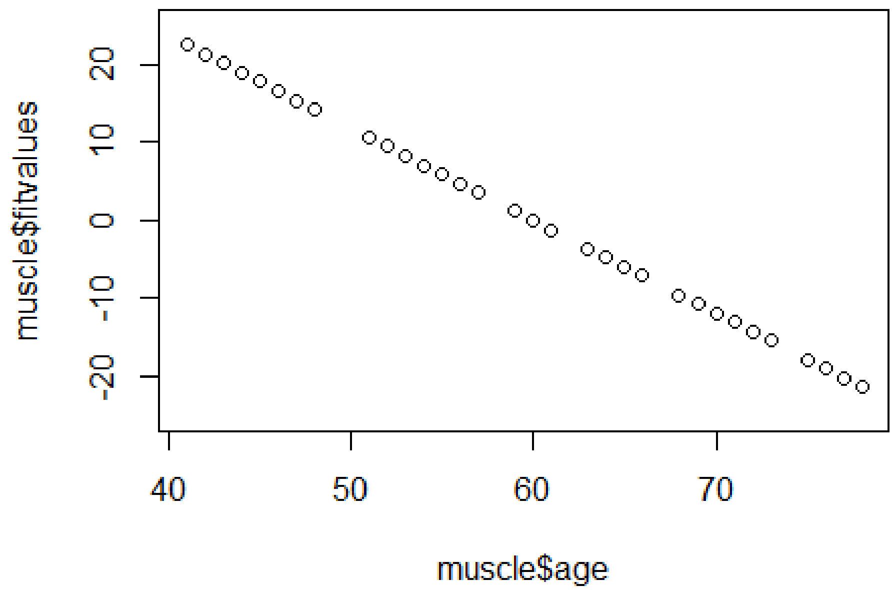
Q1 b) plot the residuals yi=y^i against xi on graph

```
plot(muscle$age, fit$residuals,main = "Residuals",ylim=c(-25, 25))
```



Q1 c) plot the values yi=y^i against xi on graph

```
muscle$fitvalues  <- predict(fit) - mean(muscle$mass)
plot(muscle$age, muscle$fitvalues,ylim=c(-25, 25))
```



Q1 d) From two graphs in part b and c does sse or ssr appear larger component of SSTO? what does this imply of magnitude of R2?

```
# SSE is smaller than SSR that implys that the r squared is larger
# When see is small, SSR large and Rsquare is large
# When SSE large, SSR small R square is small
```

Q1 e) Provide the anova table.

```
anova(fit)

## Analysis of Variance Table
##
## Response: mass
##            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
```

Q1 f) What portion of the total variance in muscle mass remains unexplained when age is added into the model? is this portion relatively small or large?

```
fit.reduced = lm(mass ~age,data=muscle)
summary(fit.reduced)$r.squared
```

```
## [1] 0.7500668

 # 75% of the total variance is explained in the model
# 25% of the model reamins unexplained
```

Q1 g) Conduct a hypothesis test using an F test with a significance level of .05 clearly state the alternatives, test the statistics and conclusion.

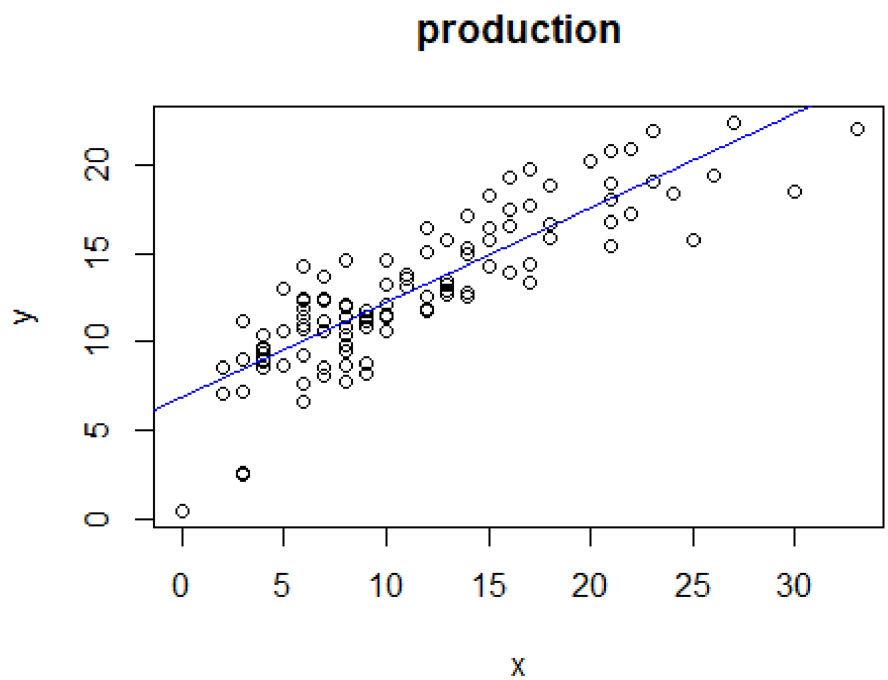
```
anova(fit)

## Analysis of Variance Table
##
## Response: mass
##           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

# since p-value is very small we reject H0 there is enough evidence that
there is a linear association between x and y.
```

Q1 h) Obtain r and R2.

```
cor(muscle$mass,muscle$age) # R again

## [1] -0.866064

cor(muscle$mass,muscle$age)^2 # R squared

## [1] 0.7500668

# slope is negative so r is negative
```

Q2 a) plot a scatter plot of the data. Is a simple linear regression appropirate?

```
plot(production$x,production$y,ylab = "y",xlab = "x",main = "production")
abline(lm(production$y~ production$x, data = production), col = "blue")
```



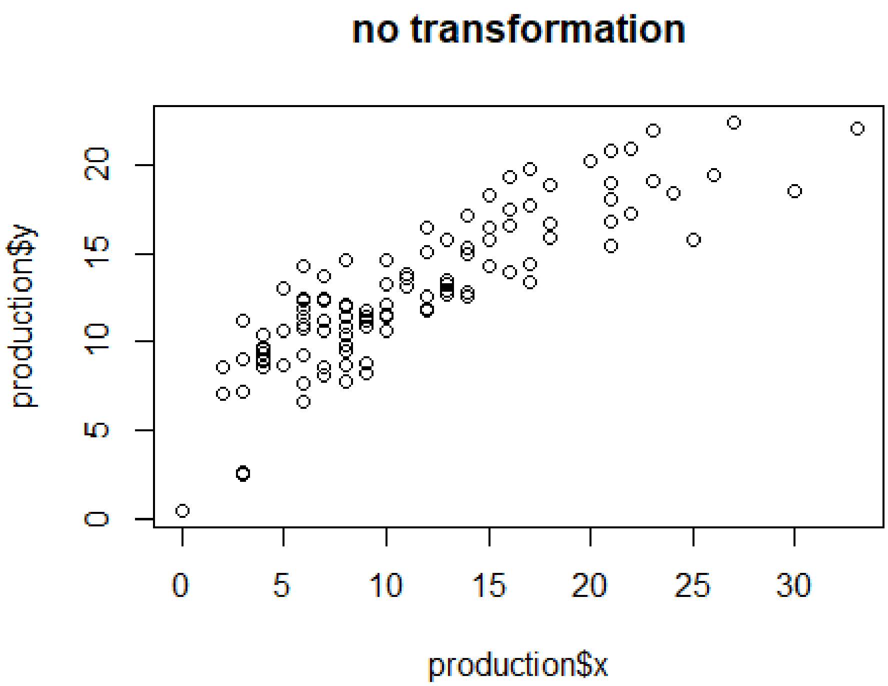

```
# Yes there appears to be no curvilinear trends between x and y
```

Q2 b) Obtain the estimated linear regression function for the data

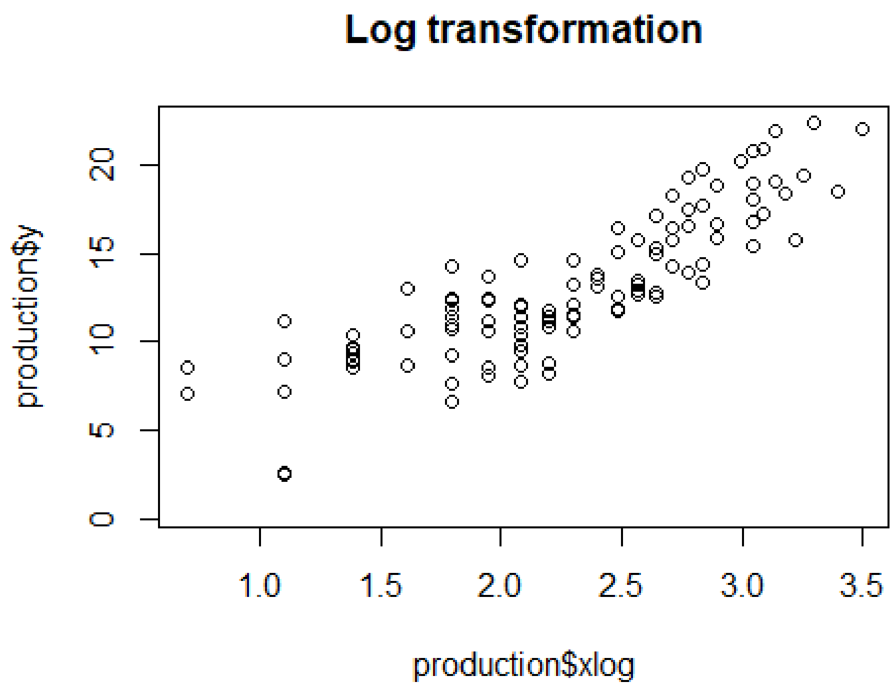
```
fit2 <- lm(y ~ x, data= production)
summary(fit2)

##
## Call:
## lm(formula = y ~ x, data = production)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3535 -1.3154  0.0036  1.2405  4.2469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.86349    0.39863   17.22   <2e-16 ***
## x            0.53327    0.03028   17.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.118 on 109 degrees of freedom
## Multiple R-squared:   0.74,  Adjusted R-squared:  0.7376
## F-statistic: 310.2 on 1 and 109 DF,  p-value: < 2.2e-16

# y^ = 6.86349  + 0.53 xi (fitted regression line no error)
```

Q2 c) Do you consider any transformation on x or y? Explain

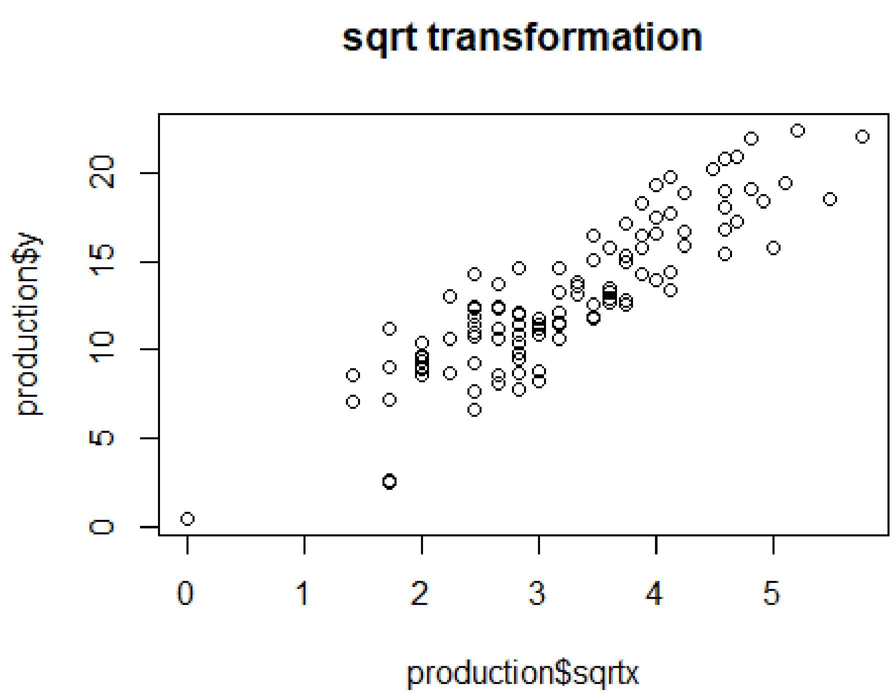
```
production$xlog <- log(production$x)
production$sqrtx <- sqrt(production$x)

plot(production$x,production$y,main = "no transformation")
```




```
plot(production$xlog,production$y,main = "Log transformation")
```




```
plot(production$sqrtx,production$y,main = "sqrt transformation")
```




```
# Yes because when you transform the data it becomes more linear
```

Q2 d) use the transformation sqrt(x) and obtain the estimated linear regression transformation

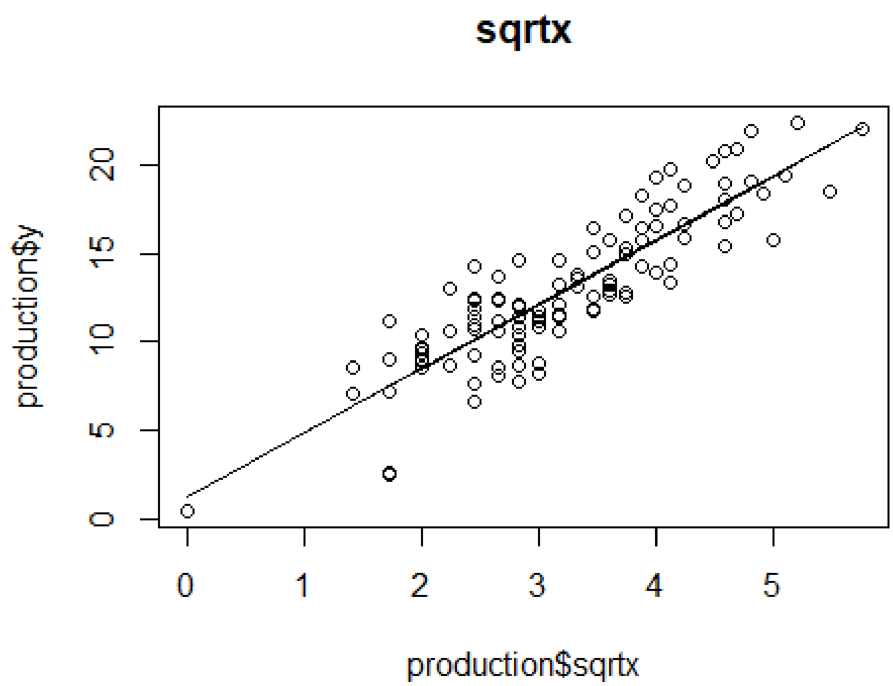
```
production$sqrtx <- sqrt(production$x)
fit3 <- lm(y ~ sqrtx, data= production)
summary(fit3)
```

```
##
## Call:
## lm(formula = y ~ sqrtx, data = production)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0008 -1.2161  0.0383  1.3367  4.1795
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2547     0.6389   1.964   0.0521 .
## sqrtx         3.6235     0.1895  19.124   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.99 on 109 degrees of freedom
## Multiple R-squared:  0.7704, Adjusted R-squared:  0.7683
## F-statistic: 365.7 on 1 and 109 DF,  p-value: < 2.2e-16
```

```
# the estimated transformation equation is y= 1.25 + 3.623xi
```

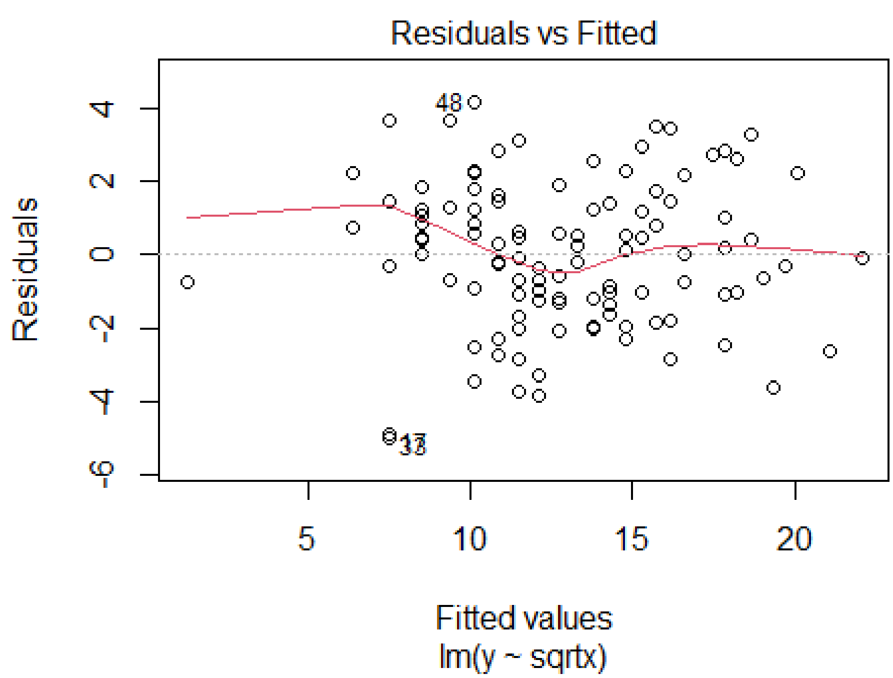
Q2 e) Plot a scatter plot of the transformed data then add the estimated regression line on a graph.

```
production$sqrtx <- sqrt(production$x)
fit3 <- lm(y ~ sqrtx, data= production)
plot(production$sqrtx,production$y,main = "sqrtx") # more linear
lines(production$sqrtx,fit3$fitted.values)
```



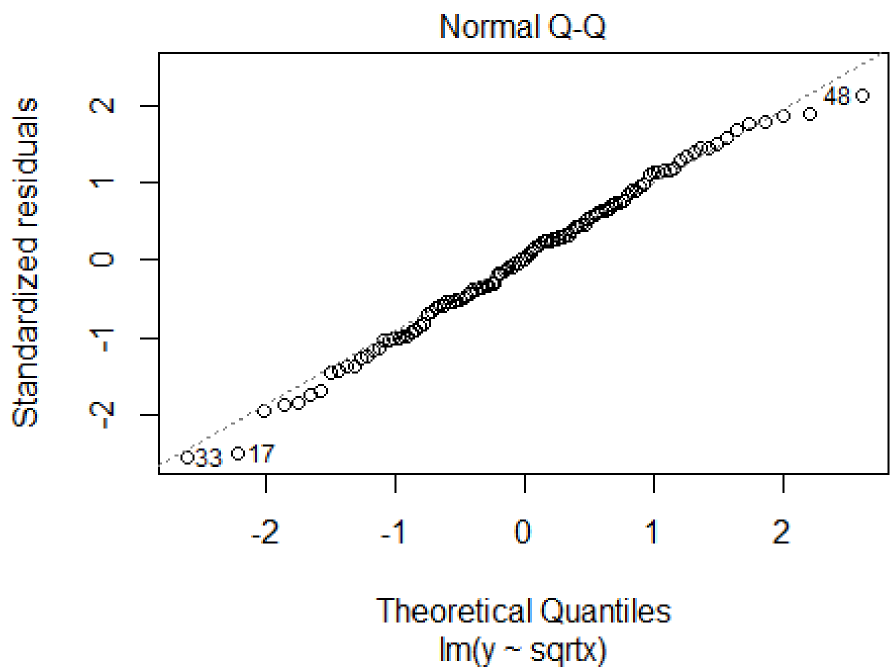
Q2 f) plot residuals against fitted values. What does plot show

```
plot(fit3, which = c(1))
```



Q2 g) plot qq plot

```
plot(fit3, which = c(2))
```



Q3 What is the reduced model? What are the degree of freedom of the reduced model?

The reduced model is: E (yi) = B0 + 5Xi

The degrees of freedom is n-1 because only one parameter is being estimated (the y-intercept)