Homework 3 Solutions:
Chapter 2: #18, 29bcde, 33bc, 56a, 57a
Chapter 3: #1, 18, 20

Assigned Mon, Oct 14:

2.18 You can test one-sided hypotheses with a t test, but not with the F test.

2.29 **b.** The ANOVA Table can be read directly from the R output (using "anova(Model), where "Model" is the name you gave your model) except that it does not provide the "Total" row.

```
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
TOTAL     59 15501.9
```

c. From the anova table, F = 174.06, p-value = 2.2×10^{-16}. Clearly the p-value is small enough to reject the null hypothesis at the 0.05 level.
d. Proportion of total variation in muscle mass that remains "unexplained" after age is introduced into the analysis is SSE / SSTO = 3874.4 / 15501.93 = 0.2499332 ≈ 25%
That means that 75% of the variation is explained, which is relatively high when dealing with human measurements. Therefore, the proportion that remains unexplained is relatively small.
e. R^2 = SSR / SSTO = 11627.5 / 15501.93 = 0.7500668 ≈ 0.75 or 75%. It can also be read directly from the R output as .7501.

r = $-\sqrt{R^2} = -0.866064$ (negative, based on negative slope)

2.33 **b.** Full model is $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$; Reduced model is $Y_i = 7.5 + \beta_1 X_i + \varepsilon_i$
c. Yes, the difference in number of parameters estimated is 1, which is $df_R - df_F$. You could also write them out, as (n – 1) – (n – 2) = 1.

2.56 **a.** $\bar{X} = 8, SSX = \sum_{i=1}^{5}(X_i - \bar{X})^2 = (49+16+4+9+36) = 114$, so

$E\{MSR\} = \sigma^2 + \beta_1^2 SSX = .36 + 9(114) = 1026.36$ (Don't confuse this with MSR!)
$E\{MSE\} = \sigma^2 = .36$

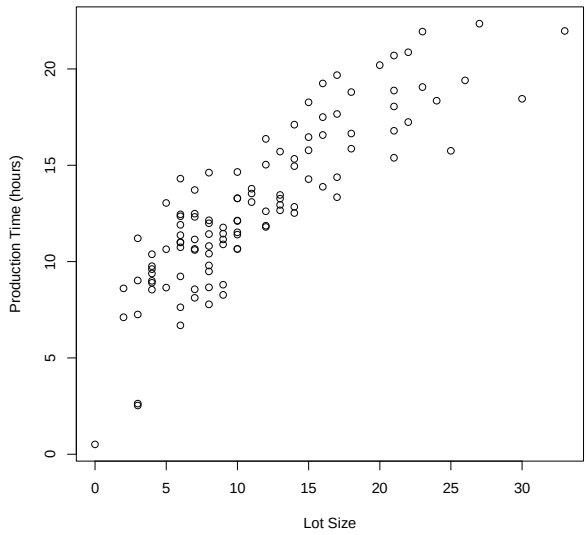
2.57 **a.** The reduced model is $Y_i = \beta_0 + 5X_i + \varepsilon_i$ or equivalently (and necessary if you wanted to run this in a software package) $Y_i - 5X_i = = \beta_0 + \varepsilon_i$. The df = n – 1 because only one parameter is being estimated.

Assigned Wed, Oct 16

3.1 (1) Residuals are $e_i = Y_i - \hat{Y}_i$; semi-studentized residuals are the residual divided by $\sqrt{MSE}$.
(2) $E\{\varepsilon_i\} = 0$ says that the mean of the error terms for the entire population is 0, whereas $\bar{e} = 0$ says that the mean of the residuals for the *n* data points in the sample have mean 0. (See (3) for definitions of error and residual.)
(3) The error term is the difference between the actual Y value and the population regression line $E\{Y\} = \beta_0 + \beta_1 X$. The residual is the difference between the actual Y value and the sample regression line, $\hat{Y} = b_0 + b_1 X$.

3.18 **a.** Here is a scatter plot of Y = production time in hours versus X = lot size:

```
prod <- read.table('CH03PR18.txt',col.names=c('Y','X'))
plot(prod$X,prod$Y,xlab='Lot Size',ylab='Production Time (hours)')
```



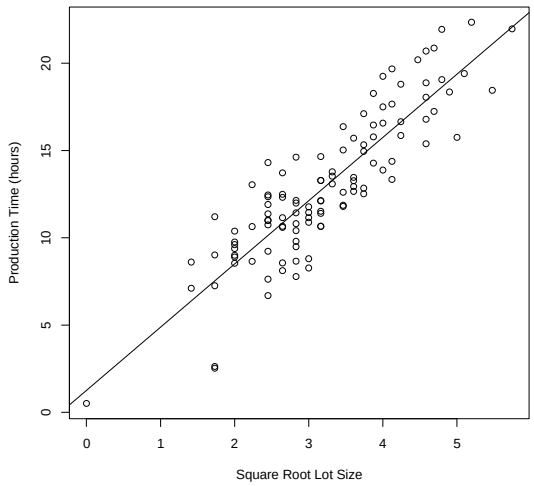
A linear relation does not appear to be adequate here. The regression relation in the scatter-plot appears to be curvilinear. The variability across different X levels appears to be fairly constant, thus a transformation to X is more suitable.

b. $\hat{Y} = 1.2547 + 3.6235\ X'$, where $X' = \sqrt{X}$, produced by the following R commands:

```
prod$rtX <- sqrt(prod$X)
LRYrtX = lm(Y~rtX,data=prod)
summary(LRYrtX)
```

c. Plot the estimated regression line and the data after the transformation.

```
plot(prod$rtX,prod$Y,xlab='Square Root Lot Size',ylab='Production Time
 (hours)')
abline(LRYrtX)
```



Yes, the scatter-plot shows a reasonably linear relation, the estimated regression line appears to be a good fit to the transformed data.

d. Plot the residuals versus the fitted values.

```
plot(LRYrtX$fitted.values, LRYrtX$residuals, main="Residuals
vs. Fitted", xlab="Fitted values", ylab="Residuals", pch=19)
```

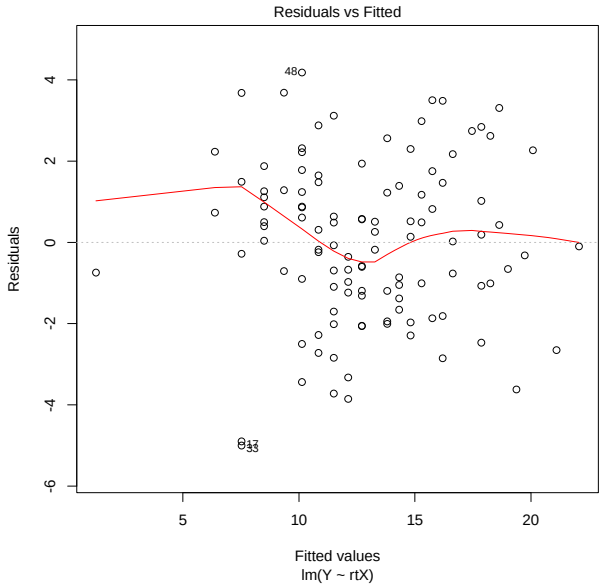
or you can use:

```
plot(LRYrtX,1)
```

where LRYrtX is the fitted linear regression model. Actually, `plot(MyModel, #)` can produce four diagnostic plots for the model called MyModel; plug in a number 1 to 4 in place of #: The plots are 1.Residuals vs. Fitted, 2.QQ-norm, 3.Scale-Location, and 4.Residuals vs. Leverage, indexed by numbers.

See the plot below. The residuals vs. fitted values plot shows that points are spread out without a systematic pattern, implying that the model fit is reasonable; points fall within a horizontal band centered around 0, so the error variances appear to be stable. Two possible outliers are detected on the bottom left side of the plot, between fitted values of 5 and 10. The other apparent outlier near fitted value of 1 isn't a problem because it's residual is small.

NOTE: Both plots on the next page were made using the shorter version of the command, `plot(LRYrtX, #)` and may have features that aren't on your plot.

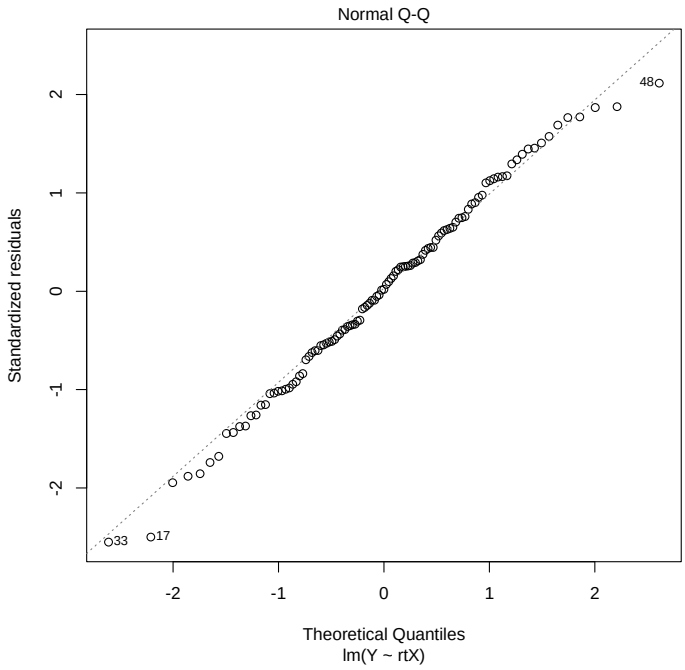


Prepare a normal probability plot. You do this using:

```
qqnorm(LRYrtX$residual, main="Normal Probability Plot", pch=19)
```

or

```
plot(LRYrtX,2)
```



The normal probability plot shows that points fall reasonably close to a straight line, with very small tails deviating from the line, suggesting that the distribution of the error terms is approximately normal.

e. The estimated regression line expressed in original units is $\hat{Y} = 1.2547 + 3.6235\ \sqrt{X}$

3.20 The error terms after the transformation $X' = 1/X$ will still be normally distributed because changing the X-axis doesn't change the vertical spread. The error terms after the transformation $Y' = 1/Y$ will not be normally distributed.