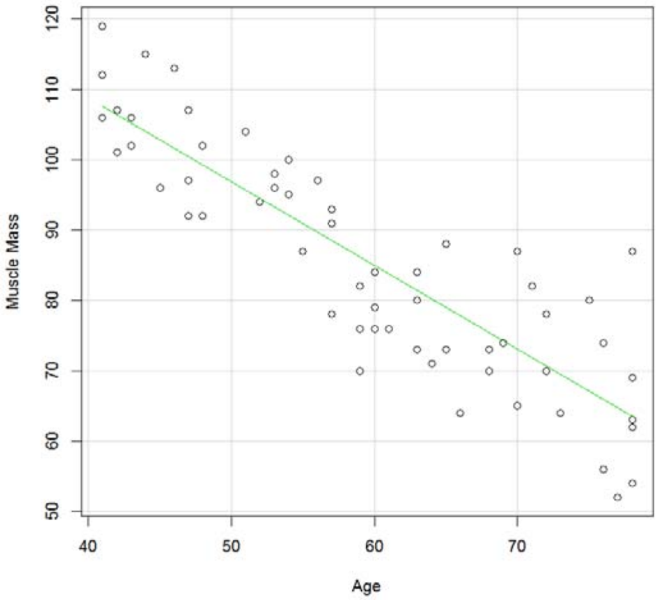
Homework 2 Solutions:
Chapter 1: #27; Chapter 2: #9, 10, 28ab
Chapter 2: #12, 27

NOTE: The R session for all of the parts involving the muscle mass data is shown at the end. Various parts are copied and shown for the relevant exercises. It is okay if you simply copied down the relevant numbers, but as assignments get more involved you should get into the practice of showing the R output. For most of these problems R can do the calculations for you, and you simply had to write down the answers. But I have shown the derivation in many cases, for completeness. You do *not* have to show all those details.

Assigned Mon, Oct 7:

1.27 **a.** From the R output, the estimated regression function is $\hat{Y}_i = 156.3466 - 1.19X_i$. See the plot below**.** The estimated linear regression function fits the data well. The negative slope obvious on the plot supports the anticipation that muscle mass decreases with age.



Plot of the estimated regression function and the data for Problem 1.27a

b. (1) a point estimate of the difference in the mean muscle mass for women differing in age by one year is the slope, which is the coefficient listed in the "Age" row, $b_1 = -1.19$.
(2) a point estimate of the mean muscle mass for women aged X = 60 years can be computed: $\hat{Y}_h = 156.3466 - 1.19(60) = 84.9466$ or read from the output of the "predict" command as 84.94683.
(3) the value of the residual for the eighth case, in which X = 41 years and Y = 112:
$e_8 = Y_8 - \hat{Y}_8 = 112 - 107.5567 = 4.4433$; or can be read from the output as 4.443252.
(4) a point estimate of σ^2 is MSE = 66.8, read from the "Residual" row and MS column in the anova table in the output. You can also find it by squaring the "residual standard error" of 8.173.

2.9 $s\{\hat{Y}_h\}$ is different for each value of X_h. There would be an infinite number of different values, not just one value. There is only one value of $s\{b_1\}$.

2.10 **a.** A prediction interval for a new observation. We are interested in a single day, not a mean for all days when the temperature is set at 31 degrees C.
b. A confidence interval for a mean response. We are interested in the mean for all families with that income, not a prediction for a single family.
c. A prediction interval for a new observation. We are interested in the single value for next month, not a mean across all months with that index of business activity.

2.28 **a.** By hand, 95% confidence interval for the mean muscle mass for women of age 60:

$$\hat{Y}_h \pm t(1-0.05/2)s\{\hat{Y}_h\},\ where\ s^2\{\hat{Y}_h\} = MSE\left(\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right) \text{ and } X_h = 60.$$

Or you can use the R "predict" command directly, to find the interval 82.83471 to 87.05895
Interpretation: We can say with 95% confidence that the mean muscle mass for the population of all women age 60 is between 82.83 and 87.06.

b. By hand, 95% prediction interval for the mean muscle mass for women of age 60:

$$\hat{Y}_h \pm t(1-0.05/2)s\{pred\},\ where\ s^2\{pred\} = MSE\left(1+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right) \text{ and } X_h = 60.$$

Or you can use the R "predict" command to find the interval 68.45067 to 101.443
No, the interval is not very precise. It's much wider than the confidence interval for the mean. Of course there is no way to really be sure because we are not told what units are used to measure muscle mass!

Assigned Wed, Oct 9:

2.12 The value of $\sigma^2\{pred\}$ in (2.37) can never be 0 (unless $\sigma^2 = 0$, in which case the relationship is functional and not statistical) because it always involves σ^2 as a separate term. In contrast, the term $\sigma^2\{\hat{Y}_h\}$ in (2.29b) approaches 0 as the sample size gets larger and larger, because the terms in the denominator approach infinity, so their inverses approach 0. Therefore, eventually it will be σ^2 multiplied by something approaching 0. The implication of this difference is that as the sample size approaches the entire population size, we can estimate $E\{Y_h\}$ precisely, i.e. the confidence interval will have width approaching 0, but we can never get a prediction interval for Y_h with width approaching 0. This makes sense because the prediction interval takes into account the natural variability of the Y values at each X. Even if we had perfect knowledge of the population, there would be a range of possible Y values at each X, not just a single value.

2.27 **a.** Conduct the hypothesis test to decide whether or not there is a *negative* linear association between amount of muscle mass and age. $\alpha = 0.05$.
The hypotheses are: H_0: $\beta_1 = 0$ vs. Ha: $\beta_1 < 0$. (Note, this is a one-sided test.)
The test statistic and p-value can be read directly from the R output, as $t^* = -13.19$. The *p*-value of 2×10^{-16} is for a two-sided test, so we need to divide it in half to get $P(t < -13.19)$ for a t distribution with df $= 60 - 2 = 58$. It's still essentially 0, which is of course less than 0.05. Therefore, we reject H_0. (NOTE: If you use the decision rule approach, reject the null hypothesis if $t^* < -t(.95, 58) = -1.67$. Clearly reject the null hypothesis.) Conclusion: there is sufficient evidence that there is a negative linear association between amount of muscle mass and age.
b. No. Even though the test of non-zero β_0 is significant, b_0 does not provide relevant information on the amount of muscle mass at birth for a female child, because data are not collected in that region; also it does not make sense to directly compare muscle mass of an adult and a newborn child.

c. The difference in expected muscle mass for women whose ages differ by one year is the slope of the regression line: β_1, so we need a 95% CI for β_1, which is provided by R as (–1.370545 to –1.009446). You could also compute it directly as:

$$
\begin{gathered}
b_1 \pm t(1-\alpha/2, n-2)s\{b_1\} \\
b_1 \pm t(1-0.05/2, 60-2)s\{b_1\} \\
-1.19 \pm 2.001717 * 0.0902 \\
(-1.370555, -1.009445)
\end{gathered}
$$

It is not necessary to know the specific ages to make this estimate because the confidence interval depends on the estimated slope of the regression equation, its standard error, and a t multiplier. These don't change as X changes. We assume that the slope stays the same over the range of X values of interest.

R COMMANDS and OUTPUT, with #COMMENTS:

```
> #Read the data; here is the command if it is in the C directory and has
white space separators
> Muscle <- read.table("C:/CH01PR27.txt", header=FALSE, sep="",
col.names=c("Mass","Age"))
> #One way to do a Scatterplot with regression line, for Problem 1.27a
> scatterplot(Mass~Age, reg.line=lm, smooth=FALSE, spread=FALSE,
boxplots=FALSE, span=0.5, xlab="Age", ylab="Muscle Mass", data=Muscle)
> #Get regression model and print results, for Problems 1.27a, b(1), 2.27a
> MuscleModel <- lm(Mass~Age, data=Muscle)
> summary(MuscleModel)
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
> #Get the predicted value for Age 60, for 1.27b(2)
> predict(MuscleModel, list(Age=60))
       1
84.94683
> #Get the residual for case 8, for 1.27b(3)
> MuscleModel$residuals[8]
       8
4.443252
> #Get anova table for 1.27b(4)
> anova(MuscleModel)
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
> #Get a confidence interval for Age 60, for 2.28a
> predict(MuscleModel, list(Age=60),se.fit=T,interval="confidence")
$fit
       fit      lwr      upr
1 84.94683 82.83471 87.05895
$se.fit
[1] 1.055154
$df
[1] 58
$residual.scale
[1] 8.173177
> #Get a prediction interval for Age 60, for 2.28b
> predict(MuscleModel, list(Age=60),se.fit=T,interval="p")
$fit
       fit      lwr     upr
1 84.94683 68.45067 101.443
$se.fit
[1] 1.055154
$df
[1] 58
$residual.scale
[1] 8.173177
> #Get a confidence interval for the slope, for 2.27c
> confint(MuscleModel)
                 2.5 %     97.5 %
(Intercept) 145.312572 167.380556
Age          -1.370545  -1.009446
```