1. Go through the 5 steps of hypothesis testing given in class for Example 0.6 in the book. (Hint: See the output on page 11 for the numerical values you will need.)

**Step 1:** Let $\mu_1 =$ population mean weight loss if everyone were to go on this weight loss program with no financial incentive, and $\mu_2 =$ population mean weight loss if everyone were to go on it with the financial incentive. Then the hypotheses are $H_0: \mu_1 = \mu_2$ (or $\mu_1 - \mu_2 = 0$) and $H_a: \mu_1 \neq \mu_2$ (or $\mu_1 - \mu_2 \neq 0$).

**Step 2:** Check necessary conditions and compute test statistic. The condition is that the two populations are approximately normal (because the sample sizes are small). The dot plots and normal probability plots shown on page 10 verify that this condition is reasonable. (Note that we did not cover the necessary conditions, so it’s okay if you did not mention them.) The test statistic is $t = -3.80$, from the output (pg. 11).

**Step 3:** Compute the $p$-value. This is also given in the output on page 11, as 0.001.

**Step 4:** Make a decision. Because 0.001 < 0.05, reject the null hypothesis (or accept the alternative hypothesis).

**Step 5:** Make a conclusion in context. There is statistically significant evidence that financial incentives would result in a higher mean weight loss than not having the incentive, for the population of people trying to lose weight who are similar to the volunteers in this study.


**1.14 a and b.** The two plots are shown below, with the R commands.

Comments for part a: This is definitely not a linear relationship, and in fact there may be two separate linear relationships, one positive and one negative.

Comments for part b: The plot shows somewhat more of a linear relationship but looks somewhat curved, and still has many points that don’t fit with the linear relationship.

```
plot(Mass~Intake,data=Caterpillars)  
plot(log10(Mass)~log10(Intake),data=Caterpillars)
```

**1.14 c** No, a linear relationship does not appear to be appropriate for either plot.
3. Do exercise 1.16 (page 59-60), with the following modifications and additions:
   a. Do as stated.
      \[ \text{plot}(x=\text{USstamps}\$\text{Year}, \]
      \[ y=\text{USstamps}\$\text{Price}, \text{xlab =}
      \[ "\text{Year}, \text{ylab = } "\text{Price (cents)}") ]

      The prices are constant across years until about 1958, then increase in a pattern that is slightly curved, but almost linear.

   b. Do as stated, but in addition give an interpretation of the slope and intercept in the context of this situation.
      \[ \text{NewUSstamps} <- \text{USstamps}[-c(1,2,3,4),] \]
      \[ \text{StampModel} <- \text{lm}(\text{Price} \sim \text{Year}, \text{data = NewUSstamps}) \]
      \[ \text{summary(StampModel)} \]

      Here is the relevant part of the output:
      Coefficients:
      \[
      \begin{array}{rcccr}
      \text{(Intercept)} & -1.647e+03 & 4.686e+01 & -35.15 & <2e-16 \text{ ***} \\
      \text{Year} & 8.410e-01 & 2.357e-02 & 35.68 & <2e-16 \text{ ***} \\
      \end{array}
      \]

      The equation is \( \hat{y} = -1647 + 0.841x \), where \( x = \text{Year} \).

      Interpretation of intercept: In the year 0, predicted stamp price is –1647 cents. (Of course this makes no sense.)

      Interpretation of slope: For each year, stamp prices are predicted to rise 0.841 cents.

      \textit{Note:} You could subtract 1958 from all of the \( x \) values and then rerun the regression. The intercept in that case would be predicted price in the year 1958.

   c. Do as stated.
      The regression line appears to be a very good fit, with the possible exception of 1958.
d. Do as stated, but for the plots include a normal probability plot, a plot of residuals versus year, and a histogram of residuals.

Normal probability plot:
```r
> qqnorm(residuals(StampModel))
> qqline(residuals(StampModel))
```

Plot of residuals versus year (with line added at 0)
```r
> plot(x=NewUSstamps$Year, y = residuals(StampModel))
> abline(h=0)
```

```r
> hist(residuals(StampModel))
```

Discussion:
The normality condition appears to be met except for one outlier, at 1958. But the plot of residuals across years seems to show a cyclical pattern. So a linear fit may be a reasonable model, but there appears to be a cycle cause by the fact that usually stamp prices are only increased every few years, not every year. So they remain flat for awhile, then go up, then remain flat again, and so on.

e. Do as stated.

The unusual residual can be identified in the plot of Year versus residuals, and is for 1958. For that year the actual stamp price was 5 cents, but the predicted price (using R) is –53 cents (negative 53 cents), which of course is an impossible price.

f. (Not in book.) The cost of mailing a letter in 2015 is 49 cents. Use the regression equation from part (b) to predict the cost for mailing a letter in 2015, and find the residual for 2015. Interpret the residual.

```r
> predict( StampModel, list(Year = c(2015) ))
1
47.40346
```

Using R, the predicted cost is 47.4 cents. So the residual is \((49 – 47.4) = 1.6\) cents. The interpretation is that it costs 1.6 cents more to buy a stamp in 2015 than would be predicted from the linear model.
4. Do exercise 1.29 (page 65) with the following modifications and additions:

a. Do as stated.

*Fit a simple linear model...*

> RetireModel <- lm(SRA ~ Year, data = Retirement)
> summary(RetireModel)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1732400.2   439864.9  -3.938  0.00148 **
Year             868.0      219.4   3.956  0.00144 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4046 on 14 degrees of freedom
Multiple R-squared:  0.5278, Adjusted R-squared:  0.494
F-statistic: 15.65 on 1 and 14 DF,  p-value: 0.001436
```

The model is \( \hat{y} = -1732400 + 868x \), where \( x = \text{Year} \).

*Identify the two sabbatical years... and compute residuals. Are they outliers?*

A plot of year versus residuals is shown below. The two sabbatical years must have been 2003 and 2011, because they have residuals that are far below 0. The values of the residuals for those years can be found using R, or computed by hand. Here is how to use R to do it, once you recognize that they are in the 7th and 15th rows of the data:

> RetireModel$residuals[c(7,15)]

```
7        15
-5642.725 -12200.955
```

So the residual for 2003 is –$5,642.725, and for 2011 it is –$12,200.955

They appear to be outliers on the residual plot shown below, but one way to check is to find the standardized residuals:

> SRst <- rstandard(RetireModel)
> SRst[c(7,15)]

```
7        15
-1.445406 -3.343753
```

The standardized residual for 2003 does not look so bad, but the one for 2011 is more than 3 standard deviations below 0, so is a clear outlier.

[NOTE: We did not cover how to determine for sure if something is an outlier, so any reasonable method you use is acceptable.]
b. Do as stated, but in addition give an interpretation of the slope in the context of this situation.

Remove sabbatical years and refit the model:

```r
> NewRetire <- Retirement[-c(7, 15),]
> NewModel <- lm(SRA ~ Year, data = NewRetire)
> summary(NewModel)
```

| Coefficients:                        | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------------------------------|----------|------------|---------|----------|
| (Intercept)                          | -2.258e+06 | 7.891e+04  | -28.62  | 2.06e-12 *** |
| Year                                 | 1.131e+03  | 3.937e+01  | 28.72   | 1.97e-12 *** |

Residual standard error: 674.7 on 12 degrees of freedom
Multiple R-squared: 0.9857, Adjusted R-squared: 0.9845
F-statistic: 825.1 on 1 and 12 DF, p-value: 1.97e-12

The model is now \( \hat{y} = -2258000 + 1131x \).

Interpret slope: For each additional year, the faculty member is predicted to increase the amount deposited into the retirement account by $1131.

Does this model provide a better fit? Make graphical and numerical comparisons.
Residual plots for the two models are shown below. The one with the outliers removed (on the right) looks much better. Numerically, the residual standard error is $4046 with the outliers, but only $674.7 without them, so there is much less unexplained variability with the outliers removed.

c. (Not in book.) Three potential explanations for outliers were given in class, and it was noted that the decision about whether to remove outliers should depend on which explanation is most likely. In this example, which of the three explanations is appropriate? Is it appropriate to remove the two outliers?

Answer: The two outliers come from a different “population.” Sabbatical years are different, so it is appropriate to remove those years as long as the model will not be used to predict the investment amount in the future for sabbatical years.

d. (Not in book.) How much would you predict that the faculty member will invest in 2015?

Answer: Using the model without the sabbatical years, the answer given by R is $20,753.8.

```r
> predict( NewModel, list(Year = c(2015) ) )
20753.8
```
5. For this problem, you will use the highway sign data and the applet for guessing and viewing a regression line, both linked to the class website. The links are below as well, in case you need them.

   a. Copy and paste the highway sign data into the data box at the applet website, removing the data that is there when you open the applet. Check the box “Show regression line” and write down the equation it provides, in the form shown.

   Answer: Here is how it looks. \( \text{response}^\wedge = 576.68 + -3.01 \times \text{explanatory} \)

   b. Remove the point that looked like a slight outlier when the example was discussed in class, which had a standardized* residual of 2.3. (You can remove it by just deleting it in the data box.) Now check the box “Show movable line.” Move the line until you think you have found the right place for the regression line. Write down the equation for the line you have placed. (This answer will differ for each student, but must be plausible to get credit.) Now check the box “Show regression line” and write the equation it gives. How well did you do in guessing where the line should go?

   *Updated 10/11 by adding “standardized.” Sorry for the confusion!

   http://www.ics.uci.edu/~jutts/110/HighwaySign.txt
   http://www.rossmanchance.com/applets/RegShuffle.htm

   Answer: The point to remove is the one with age of 75 and distance of 460. The estimated equation will differ for each student. The actual equation is: \( \text{response}^\wedge = 583.23 + -3.21 \times \text{explanatory} \)

6. Do exercise 2.14 (page 82) with the following modifications and additions.

   a. Use the 5 steps for hypothesis testing given in class.

   Answer: First, get the results using R. Here is the relevant output:

   ```r
   > TextModel <- lm(Price ~ Pages, data = TextPrices)
   > summary(TextModel)
   Coefficients:
     Estimate Std. Error t value Pr(>|t|)
     (Intercept) -3.42231   10.46374  -0.327    0.746
     Pages        0.14733    0.01925   7.653 2.45e-08 ***
   Residual standard error: 29.76 on 28 degrees of freedom (Needed for part c)
   ``)

   Step 1: Hypotheses are \( H_0: \beta_1 = 0 \) and \( H_a: \beta_1 \neq 0 \).

   Step 2: Check conditions and calculate the test statistic.

   Two plots for checking conditions are shown below. They both show slight deviations from normality, but nothing too extreme. Read the test statistic from the output: \( t = 7.653 \).
Step 3: Read the $p$-value from the output, as $2.45 \times 10^{-8}$.
Step 4: The $p$-value is clearly much less than .05, so reject the null hypothesis.
Step 5: Conclude that number of pages is useful in predicting the price of a textbook.

b. Include an interpretation of the confidence interval.
Use R to find the interval:

```r
> confint(TextModel)
     2.5 %     97.5 %
(Intercept) -24.8563229 18.011694
Pages        0.1078959  0.186761
```
The 95% confidence interval for the slope is from 0.108 to 0.187. The interpretation is that we are 95% confident that as the number of pages goes up by one, the predicted price goes up by somewhere between $0.108$ (about 11 cents) and $0.187$ (about 18 cents).

c. (Not in book.) Give numerical values for $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}_\varepsilon$, then interpret the value of $\hat{\sigma}_\varepsilon$. In other words, what does that value estimate, in the context of this example?

Answer (see output in part a):

$\hat{\beta}_0 = -3.442$

$\hat{\beta}_1 = 0.147$

$\hat{\sigma}_\varepsilon = 29.76$

Interpretation: For any fixed value of pages, there is a distribution of possible prices for textbooks. The standard deviation of that distribution is assumed to be the same for all values of $x = \text{number of pages}$. The estimated value of that standard deviation is $29.76$. 