1. (a) $M_X = 5, s_X = 2.236, M_Y = 3, s_Y = 1.225, cov_{XY} = 8/4 = 2, \ r = 2/[(2.236)(1.225)] = .73,$
   $b = .73(1.225/2.236) = .40, a = 3 - .40(5) = 1.00$

   Regression equation: $\hat{Y} = 1.00 + .40(X)$

   ![](chart.png)

   (b) $M_X = 5, s_X = 2.236, M_Y = 5, s_Y = 1.225, cov_{XY} = 8/4 = 2, \ r = 2/[(2.236)(1.225)] = .73,$
   $b = .73(1.225/2.236) = .40, a = 5 - .40(5) = 3.00$

   Regression equation: $\hat{Y} = 3.00 + .40(X)$

   ![](chart.png)

   The new regression line has the same slope as the original line, but its intercept has been increased by 2 (the same amount by which the $Y$ scores were increased). Notice how the effect of increasing the $Y$ scores by 2 was simply to raise the entire body of data points up by 2 on the $Y$ axis.

2. Regression equation from problem 1a: $\hat{Y} = 1.00 + .40(X)$

   \[
   \begin{array}{c|c|c|c|c|c}
   X & Y & \hat{Y} & Y - \hat{Y} & Y - \hat{Y} \\
   \hline
   2 & 1 & 1.8 & -0.8 & \text{Mean of } \hat{Y} \text{ scores = 3, which is equal to } M_Y \\
   4 & 4 & 2.6 & 1.4 & \text{Mean of errors of prediction } (Y - \hat{Y}) = 0 \\
   5 & 3 & 3.0 & 0.0 & \\
   6 & 3 & 3.4 & -0.4 & \\
   8 & 4 & 4.2 & -0.2 & \\
   \end{array}
   \]

   $SS\hat{Y} = (1.8-3)^2 + \ldots (4.2-3)^2 = 3.2$ \hspace{1cm} $SS_{\text{error}} = (-0.8)^2 + (1.4)^2 + \ldots + (-0.2)^2 = 2.8$

   $SS_Y = (1-3)^2 + (4-3)^2 + \ldots (4-3)^2 = 6.0$
SS\(\hat{Y}\) + \(SS_{error}\) = 6.0, which is equal to \(SS_Y\)

Recall that \(SS\hat{Y}/SS_Y = r^2\), which is the proportion of variability explained, so we have \(r^2 = 3.2/6.0 = 0.533\). Therefore \(r = \sqrt{0.533} = .73\) (as per question 1).

3. (a) The correlation is 1.0. The regression equation is \(\hat{Y} = 0 + 0.5X\). Note that even when \(r = 1.0\), the slope is not necessarily equal to 1. The slope very much depends on the relative amount of variability in the \(X\) and \(Y\) variables. In this case, there is more variability in \(X\) (\(s_X = 2.58\)) than in \(Y\) (\(s_Y = 1.29\)), so moving a particular distance along the \(X\) axis yields a smaller change along the \(Y\) axis.

(b) Following is one possible set of 3 pairs of scores with \(r = 1.0\). To ensure that the slope is equal to 3, when constructing these values it is necessary to select \(Y\) scores so that each step of 1 along the \(X\) axis corresponds to a step of 3 along the \(Y\) axis. The regression equation is \(\hat{Y} = 2 + 3X\). For any data set you produce, the slope will have to be 3, but the intercept could be any value.

<table>
<thead>
<tr>
<th>(X)</th>
<th>(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>

4. (a) The estimated mean of the \(Y\) scores for all people who have \(X = 30\) is the predicted value of \(Y\) for an \(X\) score of 30. According to the regression equation, \(\hat{Y} = 8.77 + (.10)30 = 5.77\).

(b) The mean of the \(Y\) scores for these people is expected to be 5.77, as shown in part (a). We assume that these \(Y\) scores are normally distributed with a standard deviation equal to the standard error of estimate, which is \(s_{Y-\hat{Y}} = 0.98\). So the cutoff value of 7 can be
converted to a $z$ score and we can find the area corresponding to scores greater than 7. Draw a diagram of this situation. Because 7 is larger than the mean, the area we want would be the smaller portion. So we have $z = (7 - 5.77)/0.98 = 1.26$; area in smaller portion $= .1038$.

5. Test B because the larger correlation means that the standard error of estimate (or measure of prediction error) will be smaller than for Test A.
   For Test A:
   $$s_{y\hat{y}} = 6.4 \sqrt{(1-.38^2) \frac{60-1}{60-2}} = 5.97$$
   For Test B:
   $$s_{y\hat{y}} = 6.4 \sqrt{(1-.72^2) \frac{60-1}{60-2}} = 4.48$$

6. The square of the correlation coefficient indicates the proportion of variability in one variable that is associated with variability in the other variable. For $r = .64$, $r^2 = .41$, so about 41% of the differences between people on environmental concerns are related to differences in their feelings of self-efficacy. We cannot say which 41% of the total variability in environmental concerns can be explained this way, but we can say that of all the differences we see, about 41% is related (not necessarily causally) to self-efficacy.

7. (a) Drawing at random a productivity score of 24 or higher from the normally distributed population can be computed by converting the score of 24 to a $z$ score: $z = (24-18.5)/5.2 = 1.06$. The smaller portion associated with this $z$ score represents the proportion of scores equal to or greater than 24 and therefore corresponds to the probability of randomly drawing a score in that range. So the probability $= .1446$.
   (b) With a motivation score of 5, the expected productivity score is $\hat{Y} = 8 + 3(5) = 23$. Among those faculty who have a motivation score of 5, then, the mean productivity score is estimated to be 23. These scores are assumed to be normally distributed and are estimated to have a standard deviation equal to the standard error of estimate:
   $$s_{y\hat{y}} = 5.2 \sqrt{(1-.58^2) \frac{80-1}{80-2}} = 4.26$$
   The task is to find what proportion of productivity scores, among faculty who have a motivation score of 5, are equal to or greater than 24. Among this subset of faculty members, the $z$ score for 24 is $z = (24-23)/4.26 = 0.23$. The area above this $z$ score corresponds to the proportion of scores in this distribution that equal or exceed 24, so we need the smaller portion, which is .4090. This is a much larger probability than the one found when we knew nothing about the faculty member's motivation score. Knowing that the faculty member has a motivation score of 5, which is noticeably larger than the mean motivation score of 3.5, implies that this person has an increased probability of having a high productivity score relative to when we knew nothing about the faculty member's motivation score. This makes sense because there is a positive correlation between
motivation and productivity, so a high motivation score should be more likely to be associated with a high productivity score.

8. There is a negative correlation between introversion and number of social media accounts, so a person with a below average scores on introversion is most likely to have an above-average number of social media accounts. Remember that a negative correlation implies that low scores on one variable are usually associated with high scores on the other variable.

9. If a group of people in this scenario all have the same X score, then we have no basis for predicting any variation among their Y scores. Any such variation in Y scores for these people would be unexplained – it would have to be due to other factors or variables that we have not measured. With a correlation of .80 between X and Y, we know that .36 of the variability in Y is unexplained. So we can expect that the variance in Y scores for these people will be roughly .36 of the Y variance, or $(1200)(.36) = 432$. Notice how this rough estimate of Y variance for people who share the same X score is very similar to the equation for the standard error of estimate:

$$s_{Y\hat{Y}} = s_Y \sqrt{(1 - r^2) \left(\frac{N-1}{N-2}\right)}$$

Essentially our rough estimate of variance for people who have their X score in common is very close to the square of the standard error of estimate (but without the $(N-1)/(N-2)$ term); that is, we have computed $s_Y^2(1 - r^2) = 1200(.36)$. This is a rough estimate of the variance among Y scores that cannot be explained by variation in X scores.

10. Start by using the `read.table` command to select the `animacy.txt` file you downloaded from the course web site. You will not need to load the `psych` library for this problem. Read the data file into a data variable called `data`. Note that the data file has a header (the columns of data are labeled X and Y), so the header argument is set to `T` (true).

```r
> data = read.table(file.choose(new = T), header = T)
```

Take a look at the raw data by printing it on the screen.

```r
> data
   X  Y
1  33 81
2  38 64
... 
```

Now construct the scatter plot for these data, with the need to belong variable (X) on the horizontal axis.

```r
> plot(data$X, data$Y)
```

This command will generate the following scatter plot. Notice that this is a negative relationship.
You can now obtain the regression equation by using the `lm` command and putting the result into a variable called `model`.

```r
> model = lm(data$Y ~ data$X)
```

To see the regression equation components (intercept and slope), just print the `model` variable.

```r
> model
Call:
  lm(formula = data$Y ~ data$X)
Coefficients:
  (Intercept)       data$X
   103.2599   -0.8084
```

The first value listed is the intercept and the next one is the slope. So the regression equation is $\hat{Y} = 103.26 + (-0.81)X$.

Alternatively, you can apply the `summary` command to the `model` variable, which produces additional information to be discussed later in the course. For now, however, look at the third line from the bottom in the output shown below, and read the value of **Residual standard error**. For these data, this value is 6.545 and it is the standard error of estimate (i.e., the standard deviation of $Y$ scores for people who share the same score on $X$).
> summary(model)
Call:
  lm(formula = data$Y ~ data$X)

Residuals:
       Min        1Q   Median        3Q       Max
-9.54040 -6.50894 -0.46562  5.42223 10.84278

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.2599     9.8090 10.5274 8.04e-13 ***
data$X       -0.8084     0.2564 -3.1530  0.00315 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.545 on 38 degrees of freedom
Multiple R-squared:  0.2074,  Adjusted R-squared:  0.1866
F-statistic: 9.944 on 1 and 38 DF,  p-value: 0.003147

To place the regression line into the scatter plot, use the `abline` command with `model` as the argument. Note that you first have to construct the scatter plot, which we did above with the `plot(data$X, data$Y)` command.

> abline(model)
11. Download the population data file called `pop.txt`, and read it into a data variable called `data`. Compute the correlation and the regression equation, then construct the scatter plot and the regression line.

```r
> data = read.table(file.choose(new = T), header = T)
> cor(data$X, data$Y)
[1] 0.6787551
> model = lm(data$Y ~ data$X)
> model

Call:
lm(formula = data$Y ~ data$X)

Coefficients:
(Intercept) data$X
1.5225 0.9823

So in the population, the regression equation is $\hat{Y} = 1.52 + 0.98X$.

```r
> plot(data)
> abline(model)
```

Now draw a random sample of 10 pairs of scores and estimate the population regression line. The scatter plot and the line are also produced.

```r
> data10 = data[sample(nrow(data), 10), ]
> model = lm(data10$Y ~ data10$X)
> model

Call:
lm(formula = data10$Y ~ data10$X)

Coefficients:
(Intercept) data10$X
11.088 0.743

In this case the regression equation from the sample is $\hat{Y} = 11.09 + 0.74X$. Yours will differ due to differences in the outcome of the random sampling process. Notice how different the intercept and slope are from the true population values.
> plot(data10)
> abline(model)

Now for a sample of 100.
> data100 = data[sample(nrow(data), 100), ]
> model = lm(data100$Y ~ data100$X)
> model

Call:
  lm(formula = data100$Y ~ data100$X)

Coefficients:
   (Intercept)    data100$X
      4.8485      0.9318

The regression equation is \( \hat{Y} = 4.85 + 0.93X \). Notice that the intercept and slope in this equation are much closer to the population values than was the case for a sample of 10 subjects. The scatter plot and regression line look like this.
> plot(data100)
> abline(model)