EXERCISE 1

Let \( y = \text{cadmium}, \ x = \text{lead} \). We will need the following table:

<table>
<thead>
<tr>
<th>( y )</th>
<th>( x )</th>
<th>( y^2 )</th>
<th>( x^2 )</th>
<th>( x*y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.7</td>
<td>299</td>
<td>136.89</td>
<td>3498.3</td>
</tr>
<tr>
<td>2</td>
<td>8.6</td>
<td>277</td>
<td>73.96</td>
<td>2267.7</td>
</tr>
<tr>
<td>3</td>
<td>6.5</td>
<td>199</td>
<td>42.25</td>
<td>1349.9</td>
</tr>
<tr>
<td>4</td>
<td>2.6</td>
<td>116</td>
<td>6.76</td>
<td>301.6</td>
</tr>
<tr>
<td>5</td>
<td>2.8</td>
<td>117</td>
<td>7.84</td>
<td>327.6</td>
</tr>
<tr>
<td>6</td>
<td>3.0</td>
<td>137</td>
<td>9.00</td>
<td>411.0</td>
</tr>
</tbody>
</table>

The sum of these columns are:

\[
\sum_{i=1}^{6} y_i = 35.2, \quad \sum_{i=1}^{6} x_i = 114.5, \quad \sum_{i=1}^{6} y_i^2 = 276.7, \quad \sum_{i=1}^{6} x_i^2 = 2516.45 \quad \text{and} \quad \sum_{i=1}^{6} x_i y_i = 821.4.
\]

Using the formulas from the handouts we compute the following:

a. The standard deviation of \( \text{cadmium} \): \( \text{sd}(y) = 3.75 \).

b. The standard deviation of \( \text{lead} \): \( \text{sd}(x) = 81.41 \).

c. The estimates of \( \beta_0 \) and \( \beta_1 \) of the model \( \text{cadmium}_i = \beta_0 + \beta_1 \text{lead}_i + \epsilon_i \):

\[
\hat{\beta}_1 = 0.04517 \quad \text{and} \quad \hat{\beta}_0 = -2.753.
\]

d. The covariance between \( \text{cadmium} \) and \( \text{lead} \): \( \text{cov}(y, x) = 299.37 \).

e. The correlation coefficient between \( \text{cadmium} \) and \( \text{lead} \): \( r = 0.98 \).

EXERCISE 2

Consider the simple regression model:

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]

a. Show that the sum of the residuals is always equal to zero:

\[
\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) = 0.
\]

b. Show that the estimate of \( \beta_1 \) can be computed also using:

\[
\hat{\beta}_1 = r \frac{\text{sd}(y)}{\text{sd}(x)}
\]

From the handout:

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{and} \quad \text{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}.
\]

\[
\hat{\beta}_1 = \frac{\text{cov}(y, x)}{\text{var}(x)} = \frac{\text{cov}(y, x)}{\text{sd}(x) \text{sd}(y)} = r \frac{\text{sd}(y)}{\text{sd}(x)}.
\]

c. Use the result of part (b) to compute again \( \hat{\beta}_1 \) of exercise 1.

\[
\hat{\beta}_1 = r \frac{\text{sd}(y)}{\text{sd}(x)} = 0.98 \frac{3.75}{81.41} = 0.0451.
\]
EXERCISE 3
Let $Y_i = \beta_1 x_i + \epsilon_i$. The $x_i$'s are non-random. To find the estimate of $\beta_1$ we minimize $S = \sum_{i=1}^{n} \epsilon_i^2$ or minimize $S = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$. So, take the derivative w.r.t. to $\beta_1$, set it equal to zero and solve:

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_1 x_i) x_i = 0$$

Solving for $\hat{\beta}$ we get:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$  

EXERCISE 4
We have the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $x_i$ is in inches, and therefore the model in centimeters will be $y_i = \beta_0^* + \beta_1^* x_i + \epsilon_i$.

a. The least squares estimates of $\beta_0^*$ and $\beta_1^*$ are:

$$\beta_1^* = \frac{c}{c^2} \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} x_i^2 - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2} \Rightarrow \beta_1^* = \frac{1}{c} \hat{\beta}_1.$$  

For $\hat{\beta}_0^*$ we have:

$$\hat{\beta}_0^* = \bar{y} - \beta_1^* \bar{x} = \bar{y} - \frac{1}{c} \hat{\beta}_1 \bar{x} \Rightarrow \hat{\beta}_0^* = \hat{\beta}_0.$$  

b. The value of $R^2$ remains the same:

$$(R^2)^2 = (\beta_1^*)^2 S_{xy}^2 = \frac{1}{c^2} \hat{\beta}_1^* c^2 S_{xy}^2 = \hat{\beta}_1^* S_{xy}^2 = R^2$$

EXERCISE 5
We can write the centered model as

$$y_i = \gamma_0 + \beta_1 z_i + \epsilon_i$$

where $z_i = x_i - \bar{x}$. The estimates of $\beta_1$ and $\gamma_0$ are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} z_i y_i - \frac{1}{n} (\sum_{i=1}^{n} z_i)(\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} z_i^2 - \left(\frac{\sum_{i=1}^{n} z_i}{n}\right)^2}$$

We note however that $\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$. Therefore the estimate of $\beta_1$ is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i y_i - \bar{x} y_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$  

We observe that this estimate is the same as the estimate of the uncentered model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

And since $\bar{z} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n} = 0$ the estimate of $\gamma_0$ is:

$$\gamma_0 = \bar{y} - \hat{\beta}_1 \bar{z} \Rightarrow \gamma_0 = \bar{y}.$$  

EXERCISE 6
You are given $s_y = 10$, $\sum_{i=1}^{19} (y_i - \hat{y}_i)^2 = 180$.

a. The proportion of the variation in $y$ that can be explained by $x$ is the $R^2$. We know that $R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$.  

$\text{SST} = (n - 1) S_y^2 = (19 - 1)10^2 = 1800$, \text{SSE} = $\sum_{i=1}^{19} (y_i - \hat{y}_i)^2 = 180$. Therefore $R^2 = 1 - \frac{180}{1800} = 0.90$. So, 90% of the variation in $y$ can be explained by $x$.

b. The standard error of the estimate is $s_x = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{180}{18-2}} = 3.25$.  

EXERCISE 7
You are given the following: \( \bar{x} = 76, \bar{y} = 880, \sum_{i=1}^{n}(x_i - \bar{x})^2 = 6800, \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 14200, r_{xy} = 0.72, s_e = 20.13.\)

a. \( \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{14200}{6800} = 2.088.\)

b. \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 880 - 2.088(76) = 721.312.\)

c. \( r = \hat{\beta}_1 \frac{s_x}{s_y} \Rightarrow s_y = \frac{\hat{\beta}_1^2 s_x^2}{r^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n - 1} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{r^2(n - 1)} \Rightarrow \sum_{i=1}^{n}(y_i - \bar{y})^2 = 2.088^2(6800) = 2288 \Rightarrow \sum_{i=1}^{n}(y_i - \bar{y})^2 = 57188.\)

d. \( R^2 = 1 - \frac{SSE}{SST} \Rightarrow r^2 = \frac{SST - SSE}{SST} \Rightarrow SST - SSE = r^2(SST) \Rightarrow SSE = SST - r^2(SST) \Rightarrow SSE = SST(1 - r^2) = 57188(1 - 0.72^2) \Rightarrow SSE = 27541.74.\)

And finally:
\( S_e^2 = \frac{SSE}{n - k - 1} = \frac{SSE}{n - 2} \Rightarrow n - 2 = \frac{SSE}{S_e^2} = \frac{27541.74}{20.13^2} = 67.97 \Rightarrow n = 70.\)

EXERCISE 8
Here are the R commands:

```r
#Read the "asthma.txt" data:
a1 <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/asthma.txt", sep="", header=TRUE)

#Initialize the vector b and r:
b <- rep(0,1000)
r <- rep(0,1000)

#A for loop that will run 1000 regressions: x is fixed, the y values are permuted.
for(i in 1:1000){
y <- sample(a1$resistance)
qqq <- lm(y ~ a1$height)
b[i] <- qqq$coef[2]
r[i] <- cor(y, a1$height)
}

#Construct a histogram of using the 1000 values of b:
hist(b)

#Compute beta_hat from the actual data (original data):
q1 <- lm(a1$resistance ~ a1$height)
beta_1 <- q1$coef[2]

#Place the actual beta_1 on the histogram to see how plausible it value is under H0:
segments(beta_1,0,beta_1,200, col="green")

#Count how many of the 1000 simulated beta values are larger than the actual beta_1:
sum(b < beta_1)

#You can construct the histogram using the correlations r and find the same results.
```
#Read the "cystfibr.txt" data:
a2 <- read.table("http://www.stat.ucla.edu/~nchristo/statistics13/cystfibr.txt", sep="", header=TRUE)

#Initialize the vector b and r:
b <- rep(0,1000)
r <- rep(0,1000)

#A for loop that will run 1000 regressions: x is fixed, the y values are permuted.
for(i in 1:1000){
y <- sample(a2$resistance)
qqq <- lm(y ~ a2$height)
b[i] <- qqq$coef[2]
r[i] <- cor(y, a2$height)
}

#Contruct a histogram of using the 1000 values of b:
hist(b)

#Compute beta_hat from the actual data (original data):
q1 <- lm(a2$resistance ~ a2$height)
beta_1 <- q1$coef[2]

#Place the actual beta_1 on the histogram to see how plausible it value is under H0:
segments(beta_1,0,beta_1,200, col="green")

#Count how many of the 1000 simulated beta values are larger than the actual beta_1:
sum(b < beta_1)

#You can construct the histogram using the correlations r and find the same results.