The simple linear regression model relates two population variables, $Y$ and $x$, in a linear fashion. To be more specific, it sets

$$ Y = \beta_0 + \beta_1 x + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2). $$

Another way to write this is that

$$ Y|\ x \sim N(\beta_0 + \beta_1 x, \sigma^2). $$

In this simple linear regression model, there are three parameters: $\beta_0$, $\beta_1$, and $\sigma^2$. The interpretation of these model parameters are given below.

- $\beta_1$: $\beta_1$ is the change in the expected value of $Y$ for a unit increase in $x$. The reason for this is as follows:
  $$ E(Y) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x \quad \text{and} \quad dE(Y)/dx = \beta_1. $$

- $\beta_0$: $\beta_0 = \beta_0 + \beta_1 \times 0$ is the expected response of $Y$ when $x = 0$.

- $\sigma^2$: $\sigma^2$ is the variability in $Y$ at a particular value of $x$.

To estimate the parameters $\beta_0$, $\beta_1$, and $\sigma^2$, we first collect $n$ paired observations, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The estimates of $\beta_0$ and $\beta_1$ are those values of $\beta_0$ and $\beta_1$ which minimize the sum of the squared distances between the predicted responses (the line) and the observed responses. These are called the least squares estimators and are the values of $\beta_0$ and $\beta_1$ which minimize

$$ SS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2. $$

The estimates of $\beta_0$ and $\beta_1$ are found by setting

$$ \frac{\partial SS(\beta_0, \beta_1)}{\partial \beta_0} = \frac{\partial SS(\beta_0, \beta_1)}{\partial \beta_1} = 0, $$

and with some algebra are shown to be

$$ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 $$

A simple linear regression model can be fit in R using the following commands (in this case I will be regressing price on age).

```r
> age = c(1,1,3,4,4,5,6,7,7,8,8,10,10,13)
> price = c(13990,13495,12999,9500,10495,8995,9495,6999,6950,7850,6999,5995,4950,4495,2850)
> lm(price ~ age)
```

**Homework:** These problems are due at the beginning of class on 5/7/10.

1. Devore, page 384, #12.

**Homework Solutions:**
1. Proof:

\[
\sum_{i=1}^{I} \sum_{j=1}^{J} (x_{i,j} - \bar{x}_i)^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{i,j} - 2\bar{x}_i + \bar{x}_i)^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j}^2 - 2\bar{x}_i \sum_{i=1}^{I} x_{i,j} + \sum_{i=1}^{I} \sum_{j=1}^{J} \bar{x}_i^2
\]

\[
= \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j}^2 - 2\bar{x}_i \sum_{i=1}^{I} x_{i,j} + IJ\bar{x}_i^2
\]

\[
= \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j}^2 - IJ \left( \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j} \right)^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j}^2 - \frac{1}{IJ} \left( \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j} \right)^2
\]

2. \(MSE = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} (x_{i,j} - \bar{x}_i)^2}{I(J-1)}\) where \(\sigma^2 = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} (x_{i,j} - \bar{x}_i)^2}{I(J-1)}\).

3. In this case, \(SSTr = 8(4.39 - 5.19)^2 + 8(4.52 - 5.19)^2 + 8(5.49 - 5.19)^2 + 8(6.36 - 5.19)^2 = 20.38\), and using the formula given in (1)

\[
SST = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j}^2 - \frac{1}{IJ} \left( \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j} \right)^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j}^2 - \frac{1}{IJ} \left( IJ \bar{x}_i \right)^2 = 911.91 - \frac{1}{4 \times 8} \left( 4 \times 8 \times 5.19 \right)^2 = 49.9548.
\]

So the ANOVA table looks like:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment/Group</td>
<td>3</td>
<td>20.38</td>
<td>6.79</td>
<td>6.429</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>29.5724</td>
<td>1.056</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>31</td>
<td>49.9548</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The \(p\)-value in this case is \(P(F \geq 6.429)\) where \(F \sim F_{3,28}\), and this is calculated in R as \(1 - pf(6.429, 3, 28) = .0018\). Since this is less than .05, we reject \(H_0\) in favor of \(H_a\).

4. The complete ANOVA table is given below:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment/Group</td>
<td>2</td>
<td>152.18</td>
<td>76.09</td>
<td>5.56</td>
</tr>
<tr>
<td>Error</td>
<td>71</td>
<td>970.96</td>
<td>13.675</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
<td>1123.14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The \(p\)-value is calculated as \(P(F > 5.56)\) where \(F \sim F_{2,71}\). This is calculated in R as \(1 - pf(5.56, 2, 71) = .00571\), so \(H_0\) would be rejected in favor of \(H_a\) at \(\alpha = .05\).

5. (a) For this problem, \(\bar{x}_i = 30.82\), and \(SSE = 8 \times 4.99^2 + 8 \times 5.33^2 + 8 \times 3.33^2 + 8 \times 2.94^2 + 8 \times 2.74^2 = 724.94\), and \(SSTr = 9(29.3 - 30.82)^2 + 9(28 - 30.82)^2 + 9(30.2 - 30.82)^2 + 9(32.4 - 30.82)^2 + 9(34.2 - 30.82)^2 = 221.2\), and the ANOVA table is:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment/Group</td>
<td>4</td>
<td>221.1</td>
<td>55.278</td>
<td>3.43</td>
</tr>
<tr>
<td>Error</td>
<td>40</td>
<td>644.4</td>
<td>16.11</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>865.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The \(p\)-value is \(P(F > 3.43)\) where \(F \sim F_{4,40}\). This is calculated in R as \(1 - pf(3.43, 4, 40) = .0167\), so we would reject \(H_0\) in favor of \(H_a\).

(b) For this problem, \(\bar{x}_i = 29.3\), \(SSE = 804.35\), \(SSTr = 132.48\), and the ANOVA table is

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment/Group</td>
<td>4</td>
<td>132.48</td>
<td>33.12</td>
<td>1.65</td>
</tr>
<tr>
<td>Error</td>
<td>40</td>
<td>804.35</td>
<td>20.108</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>936.83</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This \(p\)-value is calculated as \(P(F > 1.65)\) where \(F \sim F_{4,40}\), and this is calculated in R as \(1 - pf(1.65, 4, 40) = .18\)