From chapter 10, page 387, exercises 4, 11.

4. The time between eruptions of Old Faithful geyser in Yellowstone National Park is random but is related to the duration of the last eruption. The table in the exercise shows these times for 21 consecutive eruptions.

a. Make a scatter plot of the data. Does it appear to be approximately linear?

You can do this by hand. You’ll see it looks approximately linear. But you can also do it with R, Matlab, Excel, Maple, Mathematica, or several other software packages. I used Excel. It’s not as good as the rest since they’ll do all the work for you after you enter the data. Excel only does some of it.

I entered data in the first two columns. Then selected it and asked Excel to make a scatterplot. Then I adjusted the scales and added legends for the axes.

\[
x \quad | \quad y \\
2 \quad | \quad 50 \\
1.8 \quad | \quad 57 \\
3.7 \quad | \quad 55 \\
2.2 \quad | \quad 47 \\
2.1 \quad | \quad 53 \\
2.4 \quad | \quad 50 \\
2.6 \quad | \quad 62 \\
2.8 \quad | \quad 57 \\
3.3 \quad | \quad 72 \\
3.5 \quad | \quad 62 \\
3.7 \quad | \quad 63 \\
3.8 \quad | \quad 70 \\
4.5 \quad | \quad 85 \\
4.7 \quad | \quad 75 \\
4 \quad | \quad 77 \\
4 \quad | \quad 70 \\
1.7 \quad | \quad 43 \\
1.8 \quad | \quad 48 \\
4.9 \quad | \quad 70 \\
4.2 \quad | \quad 79 \\
4.3 \quad | \quad 72 \\
\]

It looks like there’s a linear trend.

b. Fit a least squares regression line. Use it to predict the time to the next eruption if the last eruption lasted 3 minutes.

In Excel all you have to do to get that line is select the linear trendline option. I also asked to display the equation which was \( \hat{y} = \beta_1 x + \beta_0 = 9.7901x + 31.013 \). That’s enough to predict that if \( x = 3 \), then the corresponding value of \( y \) will be \( \hat{y} = 60.38 \). You could also read it off from the graph as about \( \hat{y} = 60.5 \).

c. What proportion of variability in the time between eruptions \( y \) is accounted for by the duration of eruptions \( x \)? Does it suggest that \( x \) is a good predictor for \( y \)?

\( r^2 \) indicates the fraction of the variation in \( y \) accounted for by \( x \). That’s 86.5% of the variation, which is quite a bit.
You can also compute these things without Excel’s built-in trendline. I computed the average \( \bar{x} \) of the \( x \) values at the bottom of the first column and \( \bar{y} \) at the bottom of the second column. The next two columns computed \( S_{xx} = \sum (x - \bar{x})^2 = 22.23 \), the two after that \( \text{sst} = S_{xy} = 2844.28 \), and the one after that \( S_{xy} = 217.63 \). The next column computes the predicted \( \hat{y} = 9.79x + 31.0 \) values. The last two columns compute the the error sum of squares \( \text{sse} = 716.16 \).

<table>
<thead>
<tr>
<th>x</th>
<th>( \bar{y} )</th>
<th>( x \times \bar{y} ) bar</th>
<th>( x \times \bar{y} ) bar(^2)</th>
<th>( x \times \bar{y} ) bar</th>
<th>( x \times \bar{y} ) bar(^2)</th>
<th>( x \times \bar{y} ) bar</th>
<th>( \text{cov} \times \bar{y} ) bar</th>
<th>( x \times \bar{y} ) bar</th>
<th>( x \times \bar{y} ) bar(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>50</td>
<td>1.23</td>
<td>1.5123</td>
<td>1.1217</td>
<td>161.6441</td>
<td>15.6333</td>
<td>50.39</td>
<td>-0.39</td>
<td>0.1521</td>
</tr>
<tr>
<td>1.8</td>
<td>57</td>
<td>-1.43</td>
<td>2.0449</td>
<td>-5.71</td>
<td>32.6541</td>
<td>8.1653</td>
<td>48.452</td>
<td>8.548</td>
<td>79.068304</td>
</tr>
<tr>
<td>3.75</td>
<td>55</td>
<td>0.47</td>
<td>0.2209</td>
<td>-7.71</td>
<td>59.4441</td>
<td>-3.6237</td>
<td>66.863</td>
<td>-11.863</td>
<td>100.730769</td>
</tr>
<tr>
<td>2.2</td>
<td>47</td>
<td>-1.03</td>
<td>1.0605</td>
<td>-15.71</td>
<td>246.8041</td>
<td>16.1813</td>
<td>52.328</td>
<td>-5.326</td>
<td>28.387854</td>
</tr>
<tr>
<td>2.1</td>
<td>53</td>
<td>-1.13</td>
<td>1.2769</td>
<td>-9.71</td>
<td>94.2841</td>
<td>10.9733</td>
<td>51.359</td>
<td>1.641</td>
<td>2.662881</td>
</tr>
<tr>
<td>2.4</td>
<td>50</td>
<td>-0.88</td>
<td>0.8888</td>
<td>-12.71</td>
<td>161.5441</td>
<td>10.5493</td>
<td>54.266</td>
<td>-4.265</td>
<td>18.198756</td>
</tr>
<tr>
<td>2.6</td>
<td>62</td>
<td>0.67</td>
<td>0.3667</td>
<td>-0.71</td>
<td>0.5041</td>
<td>0.4773</td>
<td>56.204</td>
<td>5.796</td>
<td>33.593626</td>
</tr>
<tr>
<td>2.8</td>
<td>57</td>
<td>-0.43</td>
<td>0.3847</td>
<td>-6.71</td>
<td>32.041</td>
<td>2.6553</td>
<td>58.142</td>
<td>-1.142</td>
<td>1.304164</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
<td>0.07</td>
<td>0.0049</td>
<td>9.29</td>
<td>85.3041</td>
<td>0.6593</td>
<td>62.987</td>
<td>9.013</td>
<td>82.124159</td>
</tr>
<tr>
<td>3.5</td>
<td>62</td>
<td>0.07</td>
<td>0.0049</td>
<td>-0.71</td>
<td>0.5041</td>
<td>-0.1971</td>
<td>64.925</td>
<td>-2.925</td>
<td>85.567265</td>
</tr>
<tr>
<td>3.7</td>
<td>63</td>
<td>0.42</td>
<td>0.2209</td>
<td>-0.71</td>
<td>0.5041</td>
<td>0.1963</td>
<td>65.883</td>
<td>-3.663</td>
<td>94.292769</td>
</tr>
<tr>
<td>3.8</td>
<td>70</td>
<td>0.57</td>
<td>0.3247</td>
<td>-7.29</td>
<td>53.1441</td>
<td>4.1533</td>
<td>67.832</td>
<td>5.186</td>
<td>70.400224</td>
</tr>
<tr>
<td>4.5</td>
<td>85</td>
<td>1.27</td>
<td>1.6129</td>
<td>-22.29</td>
<td>496.6441</td>
<td>28.3003</td>
<td>74.615</td>
<td>10.385</td>
<td>107.648255</td>
</tr>
<tr>
<td>4.7</td>
<td>75</td>
<td>1.47</td>
<td>2.1609</td>
<td>-12.29</td>
<td>151.0441</td>
<td>18.0663</td>
<td>76.553</td>
<td>-1.553</td>
<td>241.1839</td>
</tr>
<tr>
<td>4</td>
<td>77</td>
<td>0.77</td>
<td>0.5929</td>
<td>-14.29</td>
<td>204.2441</td>
<td>11.0333</td>
<td>69.777</td>
<td>7.23</td>
<td>52.2729</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>0.77</td>
<td>0.5929</td>
<td>-7.29</td>
<td>53.1441</td>
<td>4.1533</td>
<td>67.832</td>
<td>5.186</td>
<td>70.400224</td>
</tr>
<tr>
<td>1.7</td>
<td>43</td>
<td>-1.53</td>
<td>2.3409</td>
<td>-19.71</td>
<td>398.4041</td>
<td>30.1563</td>
<td>47.483</td>
<td>-4.483</td>
<td>20.097299</td>
</tr>
<tr>
<td>1.8</td>
<td>48</td>
<td>-1.43</td>
<td>2.0449</td>
<td>-16.71</td>
<td>216.3841</td>
<td>21.0353</td>
<td>48.452</td>
<td>-0.263</td>
<td>0.020430</td>
</tr>
<tr>
<td>4.9</td>
<td>70</td>
<td>1.57</td>
<td>2.7889</td>
<td>-7.29</td>
<td>53.1441</td>
<td>12.1743</td>
<td>78.491</td>
<td>-8.951</td>
<td>72.097051</td>
</tr>
<tr>
<td>4.2</td>
<td>79</td>
<td>0.97</td>
<td>0.9405</td>
<td>-16.29</td>
<td>265.3341</td>
<td>15.8013</td>
<td>73.708</td>
<td>7.292</td>
<td>53.173204</td>
</tr>
<tr>
<td>4.3</td>
<td>72</td>
<td>1.07</td>
<td>1.1449</td>
<td>9.29</td>
<td>85.3041</td>
<td>9.4030</td>
<td>72.677</td>
<td>-0.677</td>
<td>0.458329</td>
</tr>
</tbody>
</table>

The regression sum of squares is \( \text{ssr} = \text{sst} - \text{sse} = 2844.28 - 716.16 = 2128.12 \). The coefficient of determination \( r^2 = \frac{\text{ssr}}{\text{sst}} = 0.749 \) which agrees with Excel’s computation. Its positive square root (positive because the slope of the line is positive) is the sample correlation coefficient \( r = 0.865 \).

d. Calculate the mean square estimate of \( \sigma \).

From page 356, this is \( s^2 = \frac{\text{sse}}{n - 2} = \frac{716.16}{19} = 24.7 \). Therefore the estimate \( s \) for \( \sigma \) is \( \sqrt{24.7} = 4.97 \).

11. This exercise continues exercise 4.

a. Calculate a 95% prediction interval for the time to the next eruption if the last eruption lasted 3 minutes.

A prediction interval applies to a specific value of \( x \) denoted \( x^* \), in this case \( x^* = 3 \). We saw in 4b that a point estimator for \( y \) was \( \hat{y} = 60.38 \). Now we want an interval estimate. The formula for this interval is given on the top of page 362 where \( Y^* \) denotes the point estimator \( \hat{Y}^* = 60.38 \). Its endpoints are

\[
\hat{Y}^* \pm t_{n-2,\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \]

where as usual \( \alpha = 1 - 0.95 = 0.05 \). We have the values \( n = 21 \), so \( t_{19,0.025} = 2.093 \). Also, \( s = 4.97 \), and

\[
\sqrt{1 + \frac{1}{21} + \frac{(3 - 3.238)^2}{22.23}} = \sqrt{1 + 0.0476 + 0.0025} = 1.05
\]
So the interval has endpoints $0.38 \pm 13.03$. It’s the interval $[47.24, 73.53]$.

**b.** Calculate a 95% confidence interval for the mean time to the next eruption for a last eruption lasting 3 minutes. Compare this CI ot the PI in part a.

The point estimator for the mean $\hat{\mu}^*$ has the same value as $Y^*$, namely 60.5, but the interval estimate is much narrower. See the figures on page 363. Its endpoints are

$$
\hat{\mu}^* \pm t_{n-2,\alpha/2} \frac{s}{\sqrt{n}} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}
$$

which only differs from the formula for $\hat{Y}^*$ in that a 1 is missing under the radical sine. The interval turns out to be $[57.51, 63.26]$. It’s only about $\frac{1}{4}$ as wide.

**c.** Repeat part a if the last eruption lasted only 1 minute. Do you think this prediction is reliable? Why or Why not?

The computations give the interval $[26.33, 55.28]$. Our data only goes from $x = 1.7$ to $x = 4.9$. If the linear model were accurate for all values of $x$, then the interval would be reliable. But $x = 1$ lies outside the range of our data. It could well be that other physical actions affect that outcome at $x = 1$ that don’t happen in the data range. Best not to depend on this prediction interval.

[Math 218 Home Page](http://math.clarku.edu/~djoyce/ma218/)