Math 218, Mathematical Statistics

D Joyce, Spring 2016

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

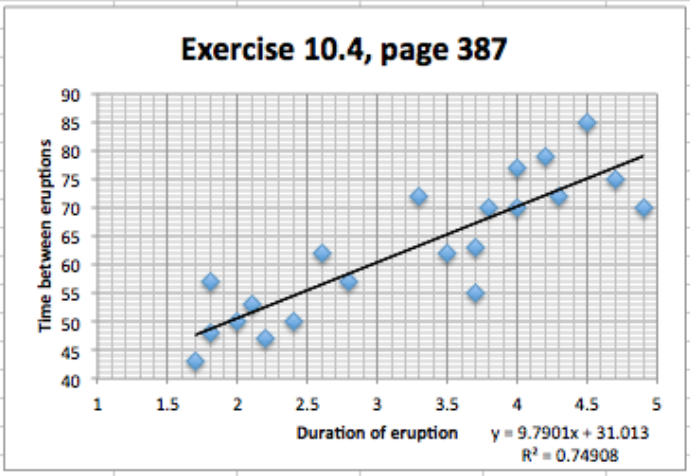
4. The time between eruptions of Old Faithful geyser in Yellowstone National Park is random but is related to the duration of the last eruption. The table in the exercise shows these times for 21 consecutive eruptions.

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	y
2	50
1.8	57
3.7	55
2.2	47
2.1	53
2.4	50
2.6	62
2.8	57
3.3	72
3.5	62
3.7	63
3.8	70
4.5	85
4.7	75
4	77
4	70
1.7	43
1.8	48
4.9	70
4.2	79
4.3	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 $\overline{x}$ of the x values at the bottom of the first column and $\overline{y}$ at the bottom of the second column. The next two columns computed $S_{xx} = \sum(x - \overline{x})^2 = 22.23$, the two after that $\text{SST} = S_{yy} = 2844.28$, 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 $\text{SSE} = 716.16$.

x	y	x-x bar	(x-x bar)^2	y-y bar	(y-y bar)^2	(x-x bar)(y-y ba	^y=9.79x+31.0	e=y-^y	(y-^y)^2
2	50	-1.23	1.5129	-12.71	161.5441	15.6333	50.39	-0.39	0.1521
1.8	57	-1.43	2.0449	-5.71	32.6041	8.1653	48.452	8.548	73.068304
3.7	55	0.47	0.2209	-7.71	59.4441	-3.6237	66.863	-11.863	140.730769
2.2	47	-1.03	1.0609	-15.71	246.8041	16.1813	52.328	-5.328	28.387584
2.1	53	-1.13	1.2769	-9.71	94.2841	10.9723	51.359	1.641	2.692881
2.4	50	-0.83	0.6889	-12.71	161.5441	10.5493	54.266	-4.266	18.198756
2.6	62	-0.63	0.3969	-0.71	0.5041	0.4473	56.204	5.796	33.593616
2.8	57	-0.43	0.1849	-5.71	32.6041	2.4553	58.142	-1.142	1.304164
3.3	72	0.07	0.0049	9.29	86.3041	0.6503	62.987	9.013	81.234169
3.5	62	0.27	0.0729	-0.71	0.5041	-0.1917	64.925	-2.925	8.555625
3.7	63	0.47	0.2209	0.29	0.0841	0.1363	66.863	-3.863	14.922769
3.8	70	0.57	0.3249	7.29	53.1441	4.1553	67.832	2.168	4.700224
4.5	85	1.27	1.6129	22.29	496.8441	28.3083	74.615	10.385	107.848225
4.7	75	1.47	2.1609	12.29	151.0441	18.0663	76.553	-1.553	2.411809
4	77	0.77	0.5929	14.29	204.2041	11.0033	69.77	7.23	52.2729
4	70	0.77	0.5929	7.29	53.1441	5.6133	69.77	0.23	0.0529
1.7	43	-1.53	2.3409	-19.71	388.4841	30.1563	47.483	-4.483	20.097289
1.8	48	-1.43	2.0449	-14.71	216.3841	21.0353	48.452	-0.452	0.204304
4.9	70	1.67	2.7889	7.29	53.1441	12.1743	78.491	-8.491	72.097081
4.2	79	0.97	0.9409	16.29	265.3641	15.8013	71.708	7.292	53.173264
4.3	72	1.07	1.1449	9.29	86.3041	9.9403	72.677	-0.677	0.458329
x bar	y bar		Sxx		Syy	Sxy			SSE
3.23809524	62.7142857		22.2309		2844.2861	217.6293			716.157062

The regression sum of squares is $\text{SSR} = \text{SST} - \text{SSE} = 2844.28 - 716.16 = 2128.12$. The coefficient of determination $r^2 = \dfrac{\text{SSR}}{\text{SST}} = 0.749$ 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 σ.

From page 356, this is $s^2 = \dfrac{\text{SSE}}{n-2} = \dfrac{716.16}{19} = 24.7$. Therefore the estimate s for σ 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

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

where as usual $\alpha = 1 - 0.95 = 0.05$. We have the values $n = 21$, so $t_{19,0.25} = 2.093$. 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 60.38 ± 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

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

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 at `http://math.clarku.edu/~djoyce/ma218/`