

Solutions to Pollard Sheet 9

The data:

Row	year	tilt	year18	tilt18
1	75	642	18	71
2	76	644	75	642
3	77	656	76	644
4	78	667	77	656
5	79	673	78	667
6	80	688	79	673
7	81	696	80	688
8	82	698	81	696
9	83	713	82	698
10	84	717	83	713
11	85	725	84	717
12	86	742	85	725
13	87	757	86	742
14			87	757

First fit the least squares line for 1975-1987. Predict the value for 1918 at the same time:

```
MTB > Regress 'tilt' 1 'year';
SUBC> Constant;
SUBC> Predict 18; <<<< get the prediction for 1918
SUBC> Brief 1.
```

The regression equation is $\text{tilt} = -61.1 + 9.32 \text{ year}$

Predictor	Coef	StDev	T	P
Constant	-61.12	25.13	-2.43	0.033
year	9.3187	0.3099	30.07	0.000

$\begin{array}{c} | \\ b \end{array}$
 $\begin{array}{c} | \\ \text{estimated standard error for } b \end{array}$

S = 4.181 R-Sq = 98.8% R-Sq(adj) = 98.7%

Predicted Values

Fit	StDev Fit	95.0% CI	95.0% PI
106.62	19.56	(63.57, 149.66)	(62.59, 150.64) XX

X denotes a row with X values away from the center
XX denotes a row with very extreme X values

Note the extreme width of the 95% prediction interval (62.59, 150.64) for the 1918 tilt. Minitab gave a warning. It is dangerous to extrapolate so far beyond the range of the data, placing so much faith in the validity of the model.

Calculate the value k_1 for which $P\{|t_{11}| \leq k_1\} = 0.95$:

```
MTB > invcdf 0.975 k1;
SUBC> t 11.
MTB > print k1
K1 2.20099 <<<< should be the same as the value from the t11 table
```

The 95% confidence interval for the slope is
 $b \pm (\text{constant from } t_{11} \text{ table}) \times (\text{estimated standard deviation for } b)$

```
MTB > let k2 = 9.3187 - k1 * 0.3099
MTB > let k3 = 9.3187 + k1 * 0.3099
MTB > print k2-k3
K2 8.63661
K3 10.0008
```

That is, the 95% confidence interval for the slope is (8.6, 10.0) tenths of a millimeter per year.

Repeat the fit, but asking for the prediction interval for 1997:

```
Fit StDev Fit    95.0% CI      95.0% PI
842.79    5.09 ( 831.58, 854.00) ( 828.29, 857.29) XX
X denotes a row with X values away from the center
XX denotes a row with very extreme X values
```

The first interval gives a range that contains the mean value for 1997 with probability 0.95 (under the model). The second interval takes into account the variability of the tilt about its mean value. The larger interval would be more useful if the engineer were worried about the consequences of the actual tilt.

Now include 1918 values in the fit. Get the prediction interval for 1997. Save the influence values:

```
MTB > Name c9 = 'HI1'
MTB > Regress 'tilt18' 1 'year18';
SUBC> Hi 'HI1';
SUBC> Constant;
SUBC> Predict 97;
SUBC> Brief 1.
```

The regression equation is
tilt18 = - 105 + 9.86 year18 <<< note the huge changes in the coefficients

Predictor	Coef	StDev	T	P
Constant	-104.877	5.719	-18.34	0.000
year18	9.85740	0.07306	134.93	0.000

S = 4.543 R-Sq = 99.9% R-Sq(adj) = 99.9%

Predicted Values

Fit	StDev Fit	95.0% CI	95.0% PI
851.29	1.93	(847.09, 855.49)	(840.54, 862.04)

```
MTB > let k2 = 9.85740 - 0.07306*k1
MTB > let k3 = 9.85740 + 0.07306*k1
MTB > print k2-k3
K2 9.69660
K3 10.0182
```

The 95% confidence interval for the slope based on the data augmented by the 1918 value is (9.7, 10.0) tenths of a millimeter per year. The interval is shorter than before, but we have purchased the "improvement" by means of a most dubious assumption: that the linear trend extends over the whole period from 1918 to 1997. The prediction interval for 1997 suffers the same defect.

Row	year	leverage18
1	75	0.956303
2	76	0.072010
3	77	0.071493
4	78	0.071493
5	79	0.072010
6	80	0.073045
7	81	0.074596
8	82	0.076665
9	83	0.079250
10	84	0.082353
11	85	0.085973
12	86	0.090110
13	87	0.094764
14		0.099935

The residual plot gives some hints of trouble. The fitted value (for 1918), sitting over at the left, suggests leverage problems—but we could see that from the raw data. The leverage values tabulated at left show that the 1918 value is having an overwhelming effect on the fit. If you look very carefully at the residual plot you might notice a downward trend in the residuals for 1975-87. I couldn't see it myself, without plotting those residuals separately from the 1918 residual. The compressed horizontal scale made it hard for me to judge slopes. The strongest warning signal came from the major change in the fit when the 1918 value was added in. If one data point has a huge effect on a fit, you should wonder why.

