10. INTERVAL ESTIMATION IN REGRESSION ANALYSIS

Engineers operate at the interface between science and society.

– Dean Gordon Brown

The regression equation $\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_i$ is one estimate of the linear relationship between *X* and *Y* based on a sample of *n* data points drawn from the population. The regression coefficients $\hat{a}_0$ and $\hat{a}_1$ are point estimates of the "true" population values a_0 and a_1 and the computed value $\hat{y}_i$ is a point estimate of the "true" mean value of *Y* at $X = x_i$. We can make inferences about the "true" mean value of *Y* by constructing a confidence interval about the sample mean value $\hat{y}_i$, just as we did earlier in the course.

If the residuals are normally distributed, then a 100(1–α) confidence interval on the population mean is given by

$$\text{C.I.} = \hat{y}_i \pm t_{\alpha/2,n-2}\, s_{y|x} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SSX}}$$

We can apply this formula at all values of *X* to obtain what is sometimes called a *confidence interval about the regression line*.

Example

Construct a 95% confidence interval about the regression line for the chemical process problem.

From our previous work:

$$n = 10 \quad t_{0.025,8} = 2.306 \quad s_{y|x} = 3.18 \quad SSX = 8250 \quad \bar{x} = 145$$

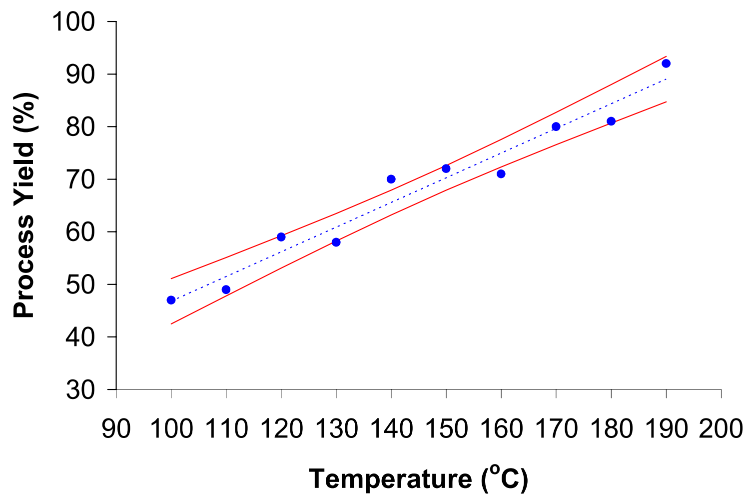
Substituting, we can write the confidence interval as

$$\text{C.I.} = \hat{y}_i \pm (2.306)(3.18)\sqrt{\frac{1}{10} + \frac{(x_i - 145)^2}{8250}}$$

The table below shows the predicted yields corresponding to the original *x* observations and the width of the confidence interval about those predicted yields.

Temperature (°C)	100	110	120	130	140	150	160	170	180	190
Yield (%)	46.8	51.5	56.2	60.9	65.6	70.2	74.9	79.6	84.3	89.0
C.I. (%)										

Note that the confidence interval is considerably narrower in the middle than it is at the ends. The minimum width is always at $X = \bar{x}$ and the grows symmetrically as $|x_i - \bar{x}|$ increases:



In our chemical process problem, the data is so very nearly linear that it's hard to tell what the confidence interval really looks like. Let's do a regression analysis on some data with a bit more scatter so we can see what's going on.

The data below represents final class averages for 20 randomly selected students taking a course in statistics and a course in numerical methods. It has been hypothesized that there is a linear relationship between statistics grades and numerical methods grades. Determine if this is true.

Statistics	Numerical Methods	Statistics	Numerical Methods
86	80	84	85
75	81	71	72
69	75	62	65
75	81	90	93
90	92	83	81
94	95	75	70
83	80	71	73
86	81	76	72
71	76	84	80
65	72	97	98

Using Excel's SLOPE and INTERCEPT functions, we determine the slope and intercept of the regression model to be

$\hat{a}_0 =$ $\hat{a}_1 =$

Having determined the model coefficients, we can predict a Numerical Methods grade for each of the students on the basis of their Statistics grade and compute the model residuals.

Using Excel's SUMSQ function, we can compute SST, SSR, and SSX as:

$$\mathrm{SST} =$$

$$\mathrm{SSR} =$$

$$\mathrm{SSX} =$$

Now we can compute the standard error of the estimate as

$$s_{y|x} = \sqrt{\frac{\mathrm{SSE}}{n-2}} = \sqrt{\frac{\mathrm{SST}-\mathrm{SSR}}{n-2}} =$$

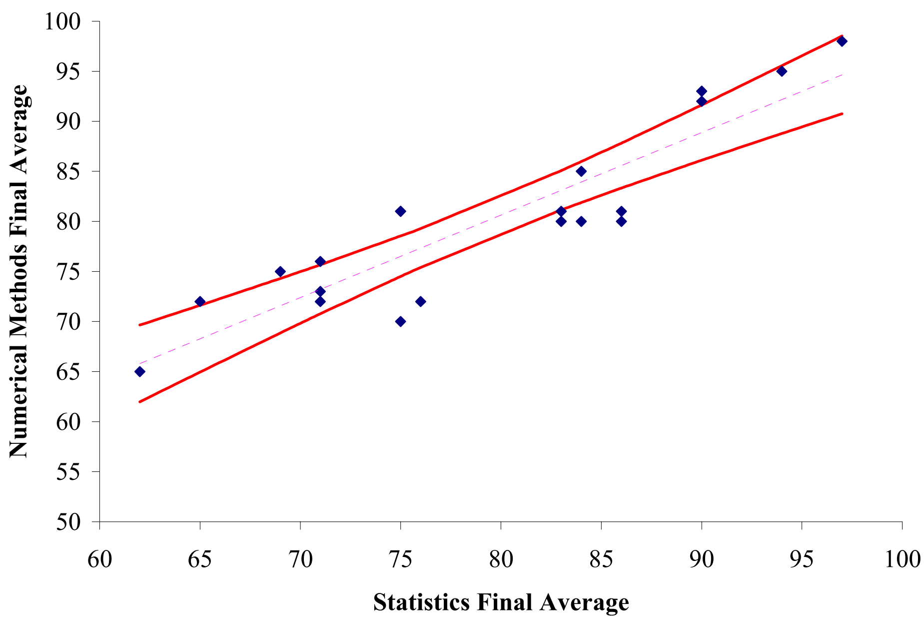
To compute a confidence interval, we need the appropriate *t*-test statistic. If we are interested in a 95% confidence interval, the *t*-test statistic is

$$t_{\alpha/2,n-2} = t_{0.025,18} =$$

Finally, the confidence interval is given by

$$\mathrm{C.I.} = \hat{y}_i \pm t_{\alpha/2,n-2}\, s_{y|x} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\mathrm{SSX}}} =$$

The results are plotted below:



What do these results mean? If we drew 100 samples from the population and did a regression analysis on every sample, then we would expect 95 of the regression lines to fall within the band shown above. Every one of those regression lines passes through the point $(\bar{x}, \bar{y})$ so the lines will move up and down slightly from one sample to the next. In addition, the slopes of the lines will change from one sample to the next. Taken together, this combination of translation and rotation forms the band of results shown in the plot.