Chemometrics

MA4605

Week 6. Lecture 12. Confidence Intervals for Regression

October 11, 2011

The regression line

We shall assume that there is a linear relationship between the signal y read from an instrument and the concentration x.

The regression line

We shall assume that there is a linear relationship between the signal y read from an instrument and the concentration x. Many procedures assume all the errors are in y and the x values are known fixed quantities free of error.

The regression line

We shall assume that there is a linear relationship between the signal y read from an instrument and the concentration x.
Many procedures assume all the errors are in y and the x values are known fixed quantities free of error.
Other assumptions are :

- the y-values have a Normal error distribution

The regression line

We shall assume that there is a linear relationship between the signal y read from an instrument and the concentration x.
Many procedures assume all the errors are in y and the x values are known fixed quantities free of error.
Other assumptions are :

- the y-values have a Normal error distribution
- the magnitude of the errors in the y-values is independent of the analyte concentration.

The regression line

We shall assume that there is a linear relationship between the signal y read from an instrument and the concentration x.
Many procedures assume all the errors are in y and the x values are known fixed quantities free of error.
Other assumptions are :

- the y-values have a Normal error distribution
- the magnitude of the errors in the y-values is independent of the analyte concentration.

Since we assume the errors are in the y-values, we are seeking the line that minimizes the deviations in the y direction between the experimental points and the calculated line.

The regression line

We shall assume that there is a linear relationship between the signal y read from an instrument and the concentration x.
Many procedures assume all the errors are in y and the x values are known fixed quantities free of error.
Other assumptions are :

- the y-values have a Normal error distribution
- the magnitude of the errors in the y-values is independent of the analyte concentration.

Since we assume the errors are in the y-values, we are seeking the line that minimizes the deviations in the y direction between the experimental points and the calculated line.

Some of the deviations(residuals) are positive and some are negative, hence we minimize the **sum of squares of the residuals**.

The straight line is $y = \alpha + \beta \cdot x$
The estimates of α and β that minimize the sum of squared residuals are given by:
Slope of least squares line: b= $\frac{\sum_i [(x_i - \overline{x}) \cdot (y_i - \overline{y})]}{\sum_i (x_i - \overline{x})^2}$
Intercept of least squares line: a= $\overline{y} - b\overline{x}$

- These equations can be used only when the the visual inspection of the observation points and the values of r indicate that a straight line relationship is realistic for the experiment in question.

Errors in the slope and the intercept of the regression line

To estimate the regression parameters α and β we don't need to assume any distribution form for the residuals.

Errors in the slope and the intercept of the regression line

To estimate the regression parameters α and β we don't need to assume any distribution form for the residuals.
However, if we want to calculate any confidence intervals or perform hypothesis testing, we will need to do this.

Errors in the slope and the intercept of the regression line

To estimate the regression parameters α and β we don't need to assume any distribution form for the residuals.
However, if we want to calculate any confidence intervals or perform hypothesis testing, we will need to do this.
The usual assumption is that the residuals are normally distributed:

$$\epsilon \sim N(0, \sigma^2)$$

- The regression line will be used to estimate the concentrations of test materials by interpolation. The random errors of the slope and the intercept are thus of importance and they depend on the assumption of normality of the residuals.

The random errors in the y-direction is defined as

$$s_{y/x} = \sqrt{\frac{\sum_i (y_i - \widehat{y}_i)^2}{n-2}}$$

The number of degrees of freedom df= $n-2$, which reflects the obvious consideration that only one straight line can be drawn through two points.
Knowing the value of the estimate of random errors in the y-direction $s_{y/x}$, we can calculate the standard errors for the slope (b) and intercept(a).

- Standard error of the slope: $s_b = \frac{s_{y/x}}{\sqrt{\sum_i (x_i - \overline{x})^2}}$
- Standard error of the intercept: $s_a = s_{y/x} \sqrt{\frac{\sum_i x_i^2}{n \sum_i (x_i - \overline{x})^2}}$

Confidence Intervals for the slope and intercept

The values of s_b and s_a can be used the usual way to calculate confidence intervals for the slope and the intercept.

- The CI for the slope is given by $b \pm t_{n-2} s_b$
- The CI for the intercept is given by $a \pm t_{n-2} s_a$

where the t values is taken at $n - 2$ degrees of freedom and the desired confidence level.

Example 5.3.1

Standard aqueous solutions of fluorescein are examined in a fluorescence spectrometer, and yield the following fluorescence intensities:

Fluorescence intensities(Y)	2.1	5.0	9.0	12.6	17.3	21.0	24.7
Concentration(X)	0	2	4	6	8	10	12

The regression line

$$\begin{aligned} y &= \alpha + \beta x \\ \textit{Intensities} &= 1.5179 + 1.9304\,\textit{Concentration} \end{aligned}$$

Calculate the confidence limits for the slope and the intercept of the regression line.

Confidence Intervals

The random errors estimate in the y-direction is

$$s_{y/x} = \sqrt{\frac{\sum_i (y_i - \widehat{y}_i)^2}{n-2}} = \sqrt{\frac{0.9368}{7-2}} = \sqrt{0.18736} = 0.4329$$

The t-value for $n-2$=5 degrees of freedom and 95% confidence level is 2.57.

- **95%CI for the slope**
  $s_b = \frac{s_{y/x}}{\sqrt{\sum_i (x_i - \overline{x})^2}} = \frac{0.4329}{\sqrt{112}} = 0.0409$
  The 95%CI for the slope = $b \pm t_{n-2} s_b = 1.9304 \pm 2.57(0.0409)$ =[1.825,2.035]
- **95%CI for the intercept**
  $s_a = s_{y/x} \sqrt{\frac{\sum_i x_i^2}{n \sum_i (x_i - \overline{x})^2}} = 0.4329 \sqrt{\frac{364}{7 \cdot 112}} = 0.295$
  The 95%CI for the intercept = $a \pm t_{n-2} s_a = 1.5179 \pm 2.57(0.295)$ =[0.759, 2.276]

Regression Coefficients in R

Intensities $<-$ c(2.1,5.0,9.0,12.6,17.3,21.0,24.7)
Concentration$<-$c(0,2,4,6,8,10,12)
The regression estimates can be obtained in R using lm(y $\sim$ x).
Store the output from the *lm* function in a object called *model*.
$>$ *model* $<-$ *lm*(*Intensities* $\sim$ *Concentration*)
Extract from this object only the values of the regression coefficients using the *coef* command.
$>$ *coef*(*model*)

```
(Intercept)           x
   1.517857    1.930357
```

CI for the regression coefficients in R

The confidence intervals for the regression estimates can be obtained in R using *confint* command.
The confidence limits are obtained by default in the *lm* model.
Extract from the *lm* output only the values of the confidence limits:

$> confint(model)$

	2.5 %	97.5%
(Intercept)	0.759700	2.276014
x	1.825220	2.035495