In a single-point external standardization we determine the value of $k_A$ by measuring the signal for a single standard containing a known concentration of analyte. Using this value of $k_A$ and the signal for our sample, we then calculate the concentration of analyte in our sample (Example 5.1). With only a single determination of $k_A$, a quantitative analysis using a single-point external standardization is straightforward.

A multiple-point standardization presents a more difficult problem. Consider the data in Table 5.1 for a multiple-point external standardization. What is our best estimate of the relationship between $S_{std}$ and $C_{std}$? It is tempting to treat this data as five separate single-point standardizations, determining $k_A$ for each standard, and reporting the mean value. Despite its simplicity, this is not an appropriate way to treat a multiple-point standardization.

**Table 5.1: Data for a Hypothetical Multiple-Point External Standardization**

<table>
<thead>
<tr>
<th>$C_{std}$ (arbitrary units)</th>
<th>$S_{std}$ (arbitrary units)</th>
<th>$k_A = S_{std} / C_{std}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>—</td>
</tr>
<tr>
<td>0.100</td>
<td>12.36</td>
<td>123.6</td>
</tr>
<tr>
<td>0.200</td>
<td>24.83</td>
<td>124.2</td>
</tr>
<tr>
<td>0.300</td>
<td>35.91</td>
<td>119.7</td>
</tr>
<tr>
<td>0.400</td>
<td>48.79</td>
<td>122.0</td>
</tr>
<tr>
<td>0.500</td>
<td>60.42</td>
<td>122.8</td>
</tr>
<tr>
<td>mean value for $k_A =$</td>
<td></td>
<td>122.5</td>
</tr>
</tbody>
</table>

So why is it inappropriate to calculate an average value for $k_A$ as done in Table 5.1? In a single-point standardization we assume that our reagent blank (the first row in Table 5.1) corrects for all constant sources of determinate error. If this is not the case, then the value of $k_A$ from a single-point standardization has a determinate error. Table 5.2 demonstrates how an uncorrected constant error affects our determination of $k_A$. The first three columns show the concentration of analyte in the standards, $C_{std}$, the signal without any source of constant error, $S_{std}$, and the actual value of $k_A$ for five standards. As we expect, the value of $k_A$ is the same for each standard. In the fourth column we add a constant determinate error of +0.50 to the signals, $(S_{std})e$. The last column contains the corresponding apparent values of $k_A$. Note that we obtain a different value of $k_A$ for each standard and that all of the apparent $k_A$ values are greater than the true value.

**Table 5.2: Effect of a Constant Determinate Error on the Value of**

<table>
<thead>
<tr>
<th>$C_{std}$ (without constant error)</th>
<th>$S_{std}$ (actual)</th>
<th>$k_A = S_{std} / C_{std}$ (actual)</th>
<th>$(S_{std})e$ (with constant error)</th>
<th>$k_A = (S_{std})e / C_{std}$ (apparent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.50</td>
<td>1.50</td>
</tr>
<tr>
<td>2.00</td>
<td>2.00</td>
<td>1.00</td>
<td>2.50</td>
<td>1.25</td>
</tr>
<tr>
<td>3.00</td>
<td>3.00</td>
<td>1.00</td>
<td>3.50</td>
<td>1.17</td>
</tr>
</tbody>
</table>
How do we find the best estimate for the relationship between the signal and the concentration of analyte in a multiple-point standardization? Figure 5.8 shows the data in Table 5.1 plotted as a normal calibration curve. Although the data certainly appear to fall along a straight line, the actual calibration curve is not intuitively obvious. The process of mathematically determining the best equation for the calibration curve is called linear regression.

<table>
<thead>
<tr>
<th>$C_{\text{std}}$</th>
<th>$S_{\text{std}}$ (without constant error)</th>
<th>$k_A = S_{\text{std}}/C_{\text{std}}$ (actual)</th>
<th>$(S_{\text{std}})_e$ (with constant error)</th>
<th>$k_A = (S_{\text{std}})<em>e/ C</em>{\text{std}}$ (apparent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.00</td>
<td>4.00</td>
<td>1.00</td>
<td>4.50</td>
<td>1.13</td>
</tr>
<tr>
<td>5.00</td>
<td>5.00</td>
<td>1.00</td>
<td>5.50</td>
<td>1.10</td>
</tr>
<tr>
<td>mean $k_A$ (true) = 1.00</td>
<td>mean $k_A$ (apparent) = 1.00</td>
<td>mean $k_A$ (apparent) = 1.23</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.8: Normal calibration curve for the hypothetical multiple-point external standardization in Table 5.1.

### 5.4.1 Linear Regression of Straight Line Calibration Curves

When a calibration curve is a straight-line, we represent it using the following mathematical equation

$$\hat{y} = \beta_0 + \beta_1 x \quad \text{(Equation 5.14)}$$

where $y$ is the signal, $S_{\text{std}}$, and $x$ is the analyte’s concentration, $C_{\text{std}}$. The constants $\beta_0$ and $\beta_1$ are, respectively, the calibration curve’s expected $y$-intercept and its expected slope. Because of uncertainty in our measurements, the best we
can do is to estimate values for $\beta_0$ and $\beta_1$, which we represent as $b_0$ and $b_1$. The goal of a linear regression analysis is to determine the best estimates for $b_0$ and $b_1$. How we do this depends on the uncertainty in our measurements.

### 5.4.2 Unweighted Linear Regression with Errors in $y$

The most common approach to completing a linear regression for Equation \( \hat{y} = b_0 + b_1x \) makes three assumptions:

1. that any difference between our experimental data and the calculated regression line is the result of indeterminate errors affecting $y$,
2. that indeterminate errors affecting $y$ are normally distributed, and
3. that the indeterminate errors in $y$ are independent of the value of $x$.

Because we assume that the indeterminate errors are the same for all standards, each standard contributes equally in estimating the slope and the $y$-intercept. For this reason the result is considered an unweighted linear regression.

The second assumption is generally true because of the central limit theorem, which we considered in Chapter 4. The validity of the two remaining assumptions is less obvious and you should evaluate them before accepting the results of a linear regression. In particular the first assumption is always suspect since there will certainly be some indeterminate errors affecting the values of $x$. When preparing a calibration curve, however, it is not unusual for the uncertainty in the signal, $S_{\text{std}}$, to be significantly larger than that for the concentration of analyte in the standards $C_{\text{std}}$. In such circumstances the first assumption is usually reasonable.

### How a Linear Regression Works

To understand the logic of an linear regression consider the example shown in Figure 5.9, which shows three data points and two possible straight-lines that might reasonably explain the data. How do we decide how well these straight-lines fits the data, and how do we determine the best straight-line?

![Figure 5.9: Illustration showing three data points and two possible straight-lines that might explain the data. The goal of a linear regression is to find the mathematical model, in this case a straight-line, that best explains the data.](image)

Let’s focus on the solid line in Figure 5.9. The equation for this line is

$$\hat{y} = b_0 + b_1x \label{5.15}$$
where \( b_0 \) and \( b_1 \) are our estimates for the \( y \)-intercept and the slope, and \( \hat{y} \) is our prediction for the experimental value of \( y \) for any value of \( x \). (If you are reading this aloud, you pronounce \( \hat{y} \) as \( y \)-hat.) Because we assume that all uncertainty is the result of indeterminate errors affecting \( y \), the difference between \( y \) and \( \hat{y} \) for each data point is the residual error, \( r \), in the our mathematical model for a particular value of \( x \).

\[ r_i = (y_i - \hat{y}_i) \]

Figure 5.10 shows the residual errors for the three data points. The smaller the total residual error, \( R \), which we define as

\[ R = \sum_i (y_i - \hat{y}_i)^2 \]

the better the fit between the straight-line and the data. In a linear regression analysis, we seek values of \( b_0 \) and \( b_1 \) that give the smallest total residual error.

The reason for squaring the individual residual errors is to prevent positive residual error from canceling out negative residual errors. You have seen this before in the equations for the sample and population standard deviations. You also can see from this equation why a linear regression is sometimes called the method of least squares.

---

**Finding the Slope and \( y \)-Intercept**

Although we will not formally develop the mathematical equations for a linear regression analysis, you can find the derivations in many standard statistical texts. The resulting equation for the slope, \( b_1 \), is

\[ b_1 = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2} \]

and the equation for the \( y \)-intercept, \( b_0 \), is

\[ b_0 = \frac{\sum_i y_i - b_1 \sum_i x_i}{n} \]

Although Equations 5.17 and 5.18 appear formidable, it is only necessary to evaluate the following four summations
Many calculators, spreadsheets, and other statistical software packages are capable of performing a linear regression analysis based on this model. To save time and to avoid tedious calculations, learn how to use one of these tools. (See Section 5.6 in this chapter for details on completing a linear regression analysis using Excel and R.) For illustrative purposes the necessary calculations are shown in detail in the following example.

Example 5.9

Using the data from Table 5.1, determine the relationship between $S_{\text{std}}$ and $C_{\text{std}}$ using an unweighted linear regression.

**Hint:**

Equations $\text{(5.17)}$ and $\text{(5.18)}$ are written in terms of the general variables $x$ and $y$. As you work through this example, remember that $x$ corresponds to $(C_{\text{std}})$, and that $y$ corresponds to $(S_{\text{std}})$.

**Solution**

We begin by setting up a table to help us organize the calculation.

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$y_i$</th>
<th>$x_i y_i$</th>
<th>$x_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.10</td>
<td>12.36</td>
<td>1.236</td>
<td>0.010</td>
</tr>
<tr>
<td>0.20</td>
<td>24.83</td>
<td>4.966</td>
<td>0.040</td>
</tr>
<tr>
<td>0.30</td>
<td>35.91</td>
<td>10.773</td>
<td>0.090</td>
</tr>
<tr>
<td>0.40</td>
<td>48.79</td>
<td>19.516</td>
<td>0.160</td>
</tr>
<tr>
<td>0.50</td>
<td>60.42</td>
<td>30.210</td>
<td>0.250</td>
</tr>
</tbody>
</table>

Adding the values in each column gives

\[
\sum x_i = 1.500 \quad \sum y_i = 182.31 \quad \sum x_i y_i = 66.701 \quad \sum x_i^2 = 0.550
\]

Substituting these values into Equation $\text{(5.17)}$ and Equation $\text{(5.18)}$, we find that the slope and the y-intercept are

\[
b_1 = \frac{(6 \times 66.701) - (1.500 \times 182.31)}{(6 \times 0.550) - (1.500)^2} = 120.706 \approx 120.71
\]
The relationship between the signal and the analyte, therefore, is

\[ S_{\text{std}} = 120.71 \times C_{\text{std}} + 0.21 \]

For now we keep two decimal places to match the number of decimal places in the signal. The resulting calibration curve is shown in Figure 5.11.

### Uncertainty in the Regression Analysis

As shown in Figure 5.11, because of indeterminate error affecting our signal, the regression line may not pass through the exact center of each data point. The cumulative deviation of our data from the regression line—that is, the total residual error—is proportional to the uncertainty in the regression. We call this uncertainty the **standard deviation about the regression**, \( s_r \), which is equal to

\[ s_r = \sqrt{\frac{\sum_i(y_i - \hat{y}_i)^2}{n-2}} \] \( \text{(Equation 5.19)} \)

where \( y_i \) is the \( i \)th experimental value, and \( \hat{y}_i \) is the corresponding value predicted by the regression line in Equation \( \text{(5.15)} \). Note that the denominator of equation 5.19 indicates that our regression analysis has \( n-2 \) degrees of freedom—we lose two degree of freedom because we use two parameters, the slope and the \( y \)-intercept, to calculate \( \hat{y}_i \).

Did you notice the similarity between the standard deviation about the regression (Equation \( \text{(5.19)} \)) and the standard deviation for a sample (Equation 4.1)?

![Figure 5.11: Calibration curve for the data in Table 5.1 and Example 5.9.](image)

A more useful representation of the uncertainty in our regression is to consider the effect of indeterminate errors on the slope, \( b_1 \), and the \( y \)-intercept, \( b_0 \), which we express as standard deviations.
\[ s_{b_1} = \sqrt{\frac{ns_r^2}{n\sum_i x_i^2 - \sum_i(x_i - \bar{x})^2}} \text{ label(5.20)} \]

\[ s_{b_0} = \sqrt{\frac{s_r^2\sum_i x_i^2}{n\sum_i x_i^2 - \sum_i(x_i - \bar{x})^2}} \text{ label(5.21)} \]

We use these standard deviations to establish confidence intervals for the expected slope, \( \beta_1 \), and the expected y-intercept, \( \beta_0 \)

\[ \beta_1 = b_1 \pm ts_{b_1} \text{ label(5.22)} \]

\[ \beta_0 = b_0 \pm ts_{b_0} \text{ label(5.23)} \]

where we select \( t \) for a significance level of \( \alpha \) and for \( n-2 \) degrees of freedom. Note that Equation \( \text{ ref(5.22)} \) and Equation \( \text{ ref(5.23)} \) do not contain a factor of \( (\sqrt{n})^{-1} \) because the confidence interval is based on a single regression line. Again, many calculators, spreadsheets, and computer software packages provide the standard deviations and confidence intervals for the slope and y-intercept. Example 5.10 illustrates the calculations.

You might contrast this with Equation 4.12 for the confidence interval around a sample’s mean value.

Example 5.10

Calculate the 95% confidence intervals for the slope and y-intercept from Example 5.9.

(As you work through this example, remember that \( x \) corresponds to \( C_{\text{std}} \), and that \( y \) corresponds to \( S_{\text{std}} \).)

**Solution**

We begin by calculating the standard deviation about the regression. To do this we must calculate the predicted signals, \( \hat{y}_i \), using the slope and y-intercept from Example 5.9, and the squares of the residual error, \( (y_i - \hat{y}_i)^2 \). Using the last standard as an example, we find that the predicted signal is

\[ \hat{y}_6 = b_0 + b_1 x_6 = 0.209 + (120.706 \times 0.500) = 60.562 \]

and that the square of the residual error is

\[ (y_i - \hat{y}_i)^2 = (60.42 - 60.562)^2 = 0.2016 \approx 0.202 \]

The following table displays the results for all six solutions.
\( x_i \quad y_i \quad \hat{y}_i \quad (y_i - \hat{y}_i)^2 \\
0.000 \quad 0.00 \quad 0.209 \quad 0.0437 \\
0.100 \quad 12.36 \quad 12.280 \quad 0.0064 \\
0.200 \quad 24.83 \quad 24.350 \quad 0.2304 \\
0.300 \quad 35.91 \quad 36.421 \quad 0.2611 \\
0.400 \quad 48.79 \quad 48.491 \quad 0.0894 \\
0.500 \quad 60.42 \quad 60.562 \quad 0.0202 \\

Adding together the data in the last column gives the numerator of Equation \( \text{(5.19)} \) as 0.6512. The standard deviation about the regression, therefore, is

\[
 s_r = \sqrt{\frac{0.6512}{6-2}} = 0.4035
\]

Next we calculate the standard deviations for the slope and the \( y \)-intercept using Equation \( \text{(5.20)} \) and Equation \( \text{(5.21)} \). The values for the summation terms are from in Example 5.9.

\[
 \text{s}_{(b_1)} = \sqrt{\frac{ns_r^2}{n\sum x_i^2 - \sum x_i^2}} = \sqrt{\frac{(0.4035)^2 \times 0.550}{(6 \times 0.550) - (1.550)^2}} = 0.965
\]

\[
 \text{s}_{(b_0)} = \sqrt{\frac{s_r^2 \sum x_i^2}{n\sum x_i^2 - \sum x_i^2}} = \sqrt{\frac{(0.4035)^2 \times 0.550}{(6 \times 0.550) - (1.550)^2}}
\]

Finally, the 95\% confidence intervals (\( \alpha = 0.05 \), 4 degrees of freedom) for the slope and \( y \)-intercept are

\[
 \beta_1 = b_1 \pm ts_{b_1} = 120.706 \pm (2.78 \times 0.965) = 120.7 \pm 2.7
\]

\[
 \beta_0 = b_0 \pm ts_{b_0} = 0.209 \pm (2.78 \times 0.292) = 0.2 \pm 0.8
\]

(You can find values for \( t \) in Appendix 4.)

The standard deviation about the regression, \( s_r \), suggests that the signal, \( S_{\text{std}} \), is precise to one decimal place. For this reason we report the slope and the \( y \)-intercept to a single decimal place.
Minimizing Uncertainty in Calibration Curves

To minimize the uncertainty in a calibration curve’s slope and y-intercept, you should evenly space your standards over a wide range of analyte concentrations. A close examination of Equation (5.20) and Equation (5.21) will help you appreciate why this is true. The denominators of both equations include the term \( \Sigma(x_i - \bar{x})^2 \). The larger the value of this term—which you accomplish by increasing the range of \( x \) around its mean value—the smaller the standard deviations in the slope and the y-intercept. Furthermore, to minimize the uncertainty in the y-intercept, it also helps to decrease the value of the term \( \Sigma x_i \) in Equation (5.21), which you accomplish by including standards for lower concentrations of the analyte.

Obtaining the Analyte’s Concentration From a Regression Equation

Once we have our regression equation, it is easy to determine the concentration of analyte in a sample. When using a normal calibration curve, for example, we measure the signal for our sample, \( S_{\text{Samp}} \), and calculate the analyte’s concentration, \( C_A \), using the regression equation.

\[
[C_\text{A}]=\frac{S_{\text{Samp}}-b_0}{b_1}\label{5.24}
\]

What is less obvious is how to report a confidence interval for \( C_A \) that expresses the uncertainty in our analysis. To calculate a confidence interval we need to know the standard deviation in the analyte’s concentration, \( s_{C_A} \), which is given by the following equation

\[
s_{C_\text{A}}=\frac{s_r}{b_1}\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{(\bar{S}_\text{Samp}-\bar{S}_\text{std})^2}{(b_1)^2\sum_i(C_{\text{std}_i}-\bar{C}_\text{std})^2}}\label{5.25}
\]

where \( m \) is the number of replicate used to establish the sample’s average signal (\( \bar{S}_\text{Samp} \)), \( n \) is the number of calibration standards, \( \bar{S}_\text{std} \) is the average signal for the calibration standards, and \( C_{\text{std}_i} \) and \( \bar{C}_\text{std} \) are the individual and mean concentrations for the calibration standards.\(^7\) Knowing the value of \( s_{C_A} \), the confidence interval for the analyte’s concentration is

\[
[\mu_{C_\text{A}}=C_\text{A} \pm ts_{C_\text{A}}]
\]

where \( \mu_{C_A} \) is the expected value of \( C_A \) in the absence of determinate errors, and with the value of \( t \) based on the desired level of confidence and \( n–2 \) degrees of freedom.

Note

Equation \( (5.25) \) is written in terms of a calibration experiment. A more general form of the equation, written in terms of \( x \) and \( y \), is given here.

\[
s_x=\frac{s_r}{b_1}\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{(\bar{Y}-\bar{y})^2}{(b_1)^2\sum_i(x_i-\bar{x})^2}}\]

A close examination of Equation \( (5.25) \) should convince you that the uncertainty in \( C_A \) is smallest when the sample’s average signal, \( \bar{S}_\text{Samp} \), is equal to the average signal for the standards, \( \bar{S}_\text{std} \). When practical, you should plan your calibration curve so that \( S_{\text{Samp}} \) falls in the middle of the calibration curve.
Example 5.11

Three replicate analyses for a sample containing an unknown concentration of analyte, yield values for $S_{\text{samp}}$ of 29.32, 29.16 and 29.51. Using the results from Example 5.9 and Example 5.10, determine the analyte's concentration, $C_A$, and its 95% confidence interval.

**Solution**

The average signal, $\bar{S}_{\text{samp}}$, is 29.33, which, using Equation (5.24) and the slope and the $y$-intercept from Example 5.9, gives the analyte’s concentration as

$$C_\text{ice(A)} = \frac{\bar{S}_{\text{samp}} - b_0}{b_1} = \frac{29.33 - 0.209}{120.706} = 0.241$$

To calculate the standard deviation for the analyte’s concentration we must determine the values for $S_{\text{std}}$ and $\Sigma(C_{\text{std}_i} - \bar{C}_{\text{std}})^2$. The former is just the average signal for the calibration standards, which, using the data in Table 5.1, is 30.385. Calculating $\Sigma(C_{\text{std}_i} - \bar{C}_{\text{std}})^2$ looks formidable, but we can simplify its calculation by recognizing that this sum of squares term is the numerator in a standard deviation equation; thus,

$$\Sigma(C_{\text{std}_i} - \bar{C}_{\text{std}})^2 = (s_{C_{\text{std}}})^2 \times (n-1)$$

where $s_{C_{\text{std}}}$ is the standard deviation for the concentration of analyte in the calibration standards. Using the data in Table 5.1 we find that $s_{C_{\text{std}}}$ is 0.1871 and

$$\Sigma(C_{\text{std}_i} - \bar{C}_{\text{std}})^2 = (0.1871)^2 \times (6-1) = 0.175$$

Substituting known values into Equation (5.25) gives

$$s_{C_A} = \frac{0.4035}{120.706} \sqrt{\frac{1}{3} + \frac{1}{6} + \frac{(29.33 - 30.385)^2}{(120.706)^2 \times 0.175}} = 0.0024$$

Finally, the 95% confidence interval for 4 degrees of freedom is

$$\mu_{C_A} = C_A \pm ts_{C_A} = 0.241 \pm (2.78 \times 0.0024) = 0.241 \pm 0.007$$

You can find values for $t$ in Appendix 4.

Figure 5.12 shows the calibration curve with curves showing the 95% confidence interval for $C_A$. 
**Figure 5.12**: Example of a normal calibration curve with a superimposed confidence interval for the analyte’s concentration. The points in blue are the original data from Table 5.1. The black line is the normal calibration curve as determined in Example 5.9. The red lines show the 95% confidence interval for $C_A$ assuming a single determination of $S_{samp}$.

**Exercise 5.4**

**Figure 5.3** shows a normal calibration curve for the quantitative analysis of Cu$^{2+}$. The data for the calibration curve are shown here:

<table>
<thead>
<tr>
<th>$[\text{Cu}^{2+}]$ (M)</th>
<th>Absorbance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$1.55 \times 10^{-3}$</td>
<td>0.050</td>
</tr>
<tr>
<td>$3.16 \times 10^{-3}$</td>
<td>0.093</td>
</tr>
<tr>
<td>$4.74 \times 10^{-3}$</td>
<td>0.143</td>
</tr>
<tr>
<td>$6.34 \times 10^{-3}$</td>
<td>0.188</td>
</tr>
<tr>
<td>$7.92 \times 10^{-3}$</td>
<td>0.236</td>
</tr>
</tbody>
</table>
Complete a linear regression analysis for this calibration data, reporting the calibration equation and the 95% confidence interval for the slope and the y-intercept. If three replicate samples give an $S_{\text{Samp}}$ of 0.114, what is the concentration of analyte in the sample and its 95% confidence interval?

Click here to review your answer to this exercise.

In a standard addition we determine the analyte’s concentration by extrapolating the calibration curve to the x-intercept. In this case the value of $C_A$ is

$$\[C_{\text{A}} = x\text{-intercept} = \frac{-b_0}{b_1}\]$$

and the standard deviation in $C_A$ is

$$\[s_{C_{\text{A}}} = \frac{s_r}{b_1} \sqrt{\frac{1}{n} + \frac{(\overline{S}_{\text{std}})^2}{(b_1)^2 \sum_{i} (C_{\text{std}_i} - \overline{C}_{\text{std}})^2}}\]$$

where $n$ is the number of standard additions (including the sample with no added standard), and $\overline{S}_{\text{std}}$ is the average signal for the $n$ standards. Because we determine the analyte’s concentration by extrapolation, rather than by interpolation, $s_{\text{CA}}$ for the method of standard additions generally is larger than for a normal calibration curve.

Evaluating a Linear Regression Model

You should never accept the result of a linear regression analysis without evaluating the validity of the model. Perhaps the simplest way to evaluate a regression analysis is to examine the residual errors. As we saw earlier, the residual error for a single calibration standard, $r_i$, is

$$\[r_i = (y_i - \hat{y}_i)\]$$

If your regression model is valid, then the residual errors should be randomly distributed about an average residual error of zero, with no apparent trend toward either smaller or larger residual errors (Figure 5.13a). Trends such as those shown in Figure 5.13b and Figure 5.13c provide evidence that at least one of the model’s assumptions is incorrect. For example, a trend toward larger residual errors at higher concentrations, as shown in Figure 5.13b, suggests that the indeterminate errors affecting the signal are not independent of the analyte’s concentration. In Figure 5.13c, the residual errors are not random, suggesting that the data can not be modeled with a straight-line relationship. Regression methods for these two cases are discussed in the following sections.

Figure 5.13: Plot of the residual error in the signal, $S_{\text{std}}$, as a function of the concentration of analyte, $C_{\text{std}}$ for an unweighted straight-line regression model. The red line shows a residual error of zero. The distribution of the residual error in (a) indicates that the unweighted linear regression model is appropriate. The increase in the residual errors in (b) for higher concentrations of analyte, suggest that a weighted straight-line regression is more appropriate. For (c), the
curved pattern to the residuals suggests that a straight-line model is inappropriate; linear regression using a quadratic model might produce a better fit.

Exercise 5.5

Using your results from Practice Exercise 5.4, construct a residual plot and explain its significance.

Click here to review your answer to this exercise.

5.4.3 Weighted Linear Regression with Errors in y

Our treatment of linear regression to this point assumes that indeterminate errors affecting $y$ are independent of the value of $x$. If this assumption is false, as is the case for the data in Figure 5.13b, then we must include the variance for each value of $y$ into our determination of the $y$-intercept, $b_0$, and the slope, $b_1$; thus

$$b_0 = \frac{\sum_i w_i y_i - b_1 \sum_i w_i x_i}{n} \label{5.26}$$

$$b_1 = \frac{n \sum_i w_i x_i y_i - \sum_i w_i x_i \sum_i w_i y_i}{n \sum_i w_i x_i^2 - (\sum_i w_i x_i)^2} \label{5.27}$$

where $w_i$ is a weighting factor that accounts for the variance in $y_i$

$$w_i = \frac{n(s_{y_i})^{-2}}{\sum_{i}(s_{y_i})^{-2}} \label{5.28}$$

and $s_{y_i}$ is the standard deviation for $y_i$. In a weighted linear regression, each $x y$-pair’s contribution to the regression line is inversely proportional to the precision of $y_i$—that is, the more precise the value of $y$, the greater its contribution to the regression.

Example 5.12

Shown here are data for an external standardization in which $s_{\text{std}}$ is the standard deviation for three replicate determination of the signal.

<table>
<thead>
<tr>
<th>$C_{\text{std}}$ (arbitrary units)</th>
<th>$S_{\text{std}}$ (arbitrary units)</th>
<th>$S_{\text{std}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>0.100</td>
<td>12.36</td>
<td>0.02</td>
</tr>
<tr>
<td>0.200</td>
<td>24.83</td>
<td>0.07</td>
</tr>
<tr>
<td>0.300</td>
<td>35.91</td>
<td>0.13</td>
</tr>
</tbody>
</table>
Determine the calibration curve’s equation using a weighted linear regression.

**Hint:**

This is the same data used in Example 5.9 with additional information about the standard deviations in the signal. As you work through this example, remember that $x$ corresponds to $C_{\text{std}}$, and that $y$ corresponds to $S_{\text{std}}$.

**Solution**

We begin by setting up a table to aid in calculating the weighting factors.

<table>
<thead>
<tr>
<th>$C_{\text{std}}$ (arbitrary units)</th>
<th>$S_{\text{std}}$ (arbitrary units)</th>
<th>$S_{\text{std}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.400</td>
<td>48.79</td>
<td>0.22</td>
</tr>
<tr>
<td>0.500</td>
<td>60.42</td>
<td>0.33</td>
</tr>
</tbody>
</table>

As a check on your calculations, the sum of the individual weights must equal the number of calibration standards, $n$. The sum of the entries in the last column is 6.0000, so all is well.

Adding together the values in the forth column gives

\[
\sum_{i}(s_{yi})^2 = 5293.09
\]

which we use to calculate the individual weights in the last column. After calculating the individual weights, we use a second table to aid in calculating the four summation terms in Equation \(\text{(5.26)}\) and Equation \(\text{(5.27)}\).
Adding the values in the last four columns gives

\[
\sum_{i}w_ix_i = 0.3644 \quad \sum_{i}w_iy_i = 44.9499
\]

\[
\sum_{i}w_ix_i^2 = 0.0499 \quad \sum_{i}w_ix_iy_i = 6.1451
\]

Substituting these values into the Equation \( \ref{5.26} \) and Equation \( \ref{5.27} \) gives the estimated slope and estimated \( y \)-intercept as

\[
b_1 = \dfrac{(6\times6.1451)-(0.3644\times44.9499)}{(6\times0.0499)-(0.3644)^2} = 122.985
\]

\[
b_0 = \dfrac{44.9499-(122.985\times0.3644)}{6} = 0.0224
\]

The calibration equation is

\[
S_{\text{std}} = 122.98 \times C_{\text{std}} + 0.02
\]

Figure 5.14 shows the calibration curve for the weighted regression and the calibration curve for the unweighted regression in Example 5.9. Although the two calibration curves are very similar, there are slight differences in the slope and in the \( y \)-intercept. Most notably, the \( y \)-intercept for the weighted linear regression is closer to the expected value of zero. Because the standard deviation for the signal, \( S_{\text{std}} \), is smaller for smaller concentrations of analyte, \( C_{\text{std}} \), a weighted linear regression gives more emphasis to these standards, allowing for a better estimate of the \( y \)-intercept.

Equations for calculating confidence intervals for the slope, the \( y \)-intercept, and the concentration of analyte when using a weighted linear regression are not as easy to define as for an unweighted linear regression.\textsuperscript{8} The confidence interval
for the analyte’s concentration, however, is at its optimum value when the analyte’s signal is near the weighted centroid, \( y_c \), of the calibration curve.

\[
\begin{align*}
\hat{y}_c &= \frac{1}{n \sum_i w_i x_i}
\end{align*}
\]

**Figure 5.14:** A comparison of unweighted and weighted normal calibration curves. See Example 5.9 for details of the unweighted linear regression and Example 5.12 for details of the weighted linear regression.

### 5.4.4 Weighted Linear Regression with Errors in Both x and y

If we remove our assumption that the indeterminate errors affecting a calibration curve exist only in the signal \( y \), then we also must factor into the regression model the indeterminate errors affecting the analyte’s concentration in the calibration standards \( x \). The solution for the resulting regression line is computationally more involved than that for either the unweighted or weighted regression lines.\(^9\) Although we will not consider the details in this textbook, you should be aware that neglecting the presence of indeterminate errors in \( x \) can bias the results of a linear regression.

See Figure 5.2 for an example of a calibration curve that deviates from a straight-line for higher concentrations of analyte.

### 5.4.5 Curvilinear and Multivariate Regression

A straight-line regression model, despite its apparent complexity, is the simplest functional relationship between two variables. What do we do if our calibration curve is curvilinear—that is, if it is a curved-line instead of a straight-line? One approach is to try transforming the data into a straight-line. Logarithms, exponentials, reciprocals, square roots, and trigonometric functions have been used in this way. A plot of \( \log(y) \) versus \( x \) is a typical example. Such transformations are not without complications. Perhaps the most obvious complication is that data with a uniform variance in \( y \) will not maintain that uniform variance after the transformation.
Another approach to developing a linear regression model is to fit a polynomial equation to the data, such as \( y = a + bx + cx^2 \). You can use linear regression to calculate the parameters \( a, b, \) and \( c \), although the equations are different than those for the linear regression of a straight line.\(^{10}\) If you cannot fit your data using a single polynomial equation, it may be possible to fit separate polynomial equations to short segments of the calibration curve. The result is a single continuous calibration curve known as a spline function.

It is worth noting that in mathematics, the term “linear” does not mean a straight-line. A linear function may contain many additive terms, but each term can have one and only one adjustable parameter. The function

\[
y = ax + bx^2
\]

is linear, but the function

\[
y = ax^b
\]

is nonlinear. This is why you can use linear regression to fit a polynomial equation to your data. Sometimes it is possible to transform a nonlinear function. For example, taking the log of both sides of the nonlinear function shown above gives a linear function.

\[
\log(y) = \log(a) + b\log(x)
\]

The regression models in this chapter apply only to functions containing a single independent variable, such as a signal that depends upon the analyte’s concentration. In the presence of an interferent, however, the signal may depend on the concentrations of both the analyte and the interferent

\[
S = k_A C_A + k_I C_I + S_{reag}
\]

where \( k_I \) is the interferent’s sensitivity and \( C_I \) is the interferent’s concentration. Multivariate calibration curves can be prepared using standards that contain known amounts of both the analyte and the interferent, and modeled using multivariate regression.\(^{11}\)

Check out the [Additional Resources](#) at the end of the textbook for more information about linear regression with errors in both variables, curvilinear regression, and multivariate regression.

---

**Contributors**

David Harvey (DePauw University)