2
SIMPLE REGRESSION ANALYSIS
There is a connection between the two, right?

That means...

There is a connection between the two, right?

Exactly!

Where did you learn so much about regression analysis, Miu?

Miu!

Blink

Blink

Earth to Miu! Are you there?
YOU WERE STARING AT THAT COUPLE.

ACK, YOU CAUGHT ME!

IT'S JUST... THEY'RE STUDYING TOGETHER.

I WISH I COULD STUDY WITH HIM LIKE THAT.

THAT'S WHY I AM TEACHING YOU! AND THERE'S NO CRYING IN STATISTICS!

PAT

PAT

THERE, THERE.

WE'RE FINALLY DOING REGRESSION ANALYSIS TODAY. DOESN'T THAT CHEER YOU UP?

YES. I WANT TO LEARN.

SIGH
All right then, let's go! This table shows the high temperature and the number of iced tea orders every day for two weeks.

<table>
<thead>
<tr>
<th>High temp. (°C)</th>
<th>Iced tea orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>22nd (Mon.)</td>
<td>29</td>
</tr>
<tr>
<td>23rd (Tues.)</td>
<td>28</td>
</tr>
<tr>
<td>24th (Wed.)</td>
<td>34</td>
</tr>
<tr>
<td>25th (Thurs.)</td>
<td>31</td>
</tr>
<tr>
<td>26th (Fri.)</td>
<td>25</td>
</tr>
<tr>
<td>27th (Sat.)</td>
<td>29</td>
</tr>
<tr>
<td>28th (Sun.)</td>
<td>32</td>
</tr>
<tr>
<td>29th (Mon.)</td>
<td>31</td>
</tr>
<tr>
<td>30th (Tues.)</td>
<td>24</td>
</tr>
<tr>
<td>31st (Wed.)</td>
<td>33</td>
</tr>
<tr>
<td>1st (Thurs.)</td>
<td>25</td>
</tr>
<tr>
<td>2nd (Fri.)</td>
<td>31</td>
</tr>
<tr>
<td>3rd (Sat.)</td>
<td>26</td>
</tr>
<tr>
<td>4th (Sun.)</td>
<td>30</td>
</tr>
</tbody>
</table>

Now... we'll first make this into a scatter plot...

See how the dots roughly line up? That suggests these variables are correlated. The correlation coefficient, called $R$, indicates how strong the correlation is. $R$ ranges from $+1$ to $-1$, and the further it is from zero, the stronger the correlation. I'll show you how to work out the correlation coefficient on page 78.

* A positive $R$ value indicates a positive relationship, meaning as $x$ increases, so does $y$. A negative $R$ value means as the $x$ value increases, the $y$ value decreases.
Here, R is large, indicating iced tea really does sell better on hotter days.

Yes, that makes sense!

Obviously more people order iced tea when it's hot out.

True, this information isn't very useful by itself.

You mean there's more?

Remember what I told you the other day? Using regression analysis...

You can predict the number of iced tea orders from the high temperature.

Sure! We haven't even begun the regression analysis.

Oh, yeah... but how?
Basically, the goal of regression analysis is...

...to obtain the regression equation...

...in the form of \( y = ax + b \).

If you input a high temperature for \( x \)...

...you can predict how many orders of iced tea there will be (\( y \)).
Regression analysis doesn’t seem too hard. Just wait...

As I said earlier, $y$ is the dependent (or outcome) variable and $x$ is the independent (or predictor) variable.

\[ y = ax + b \]

$\alpha$ is the regression coefficient, which tells us the slope of the line we make.

That leaves us with $b$, the intercept. This tells us where our line crosses the y-axis.

Okay, got it.

So how do I get the regression equation?

Hold on, Miu.

Finding the equation is only part of the story.

You also need to learn how to verify the accuracy of your equation by testing for certain circumstances. Let’s look at the process as a whole.
Here's an overview of regression analysis.

**General Regression Analysis Procedure**

**Step 1**
Draw a scatter plot of the independent variable versus the dependent variable. If the dots line up, the variables may be correlated.

**Step 2**
Calculate the regression equation.

**Step 3**
Calculate the correlation coefficient (R) and assess our population and assumptions.

**Step 4**
Conduct the analysis of variance.

**Step 5**
Calculate the confidence intervals.

**Step 6**
Make a prediction!
We have to do all these steps?

For a thorough analysis, yes.

What do steps 4 and 5 even mean?

It’s easier to explain with an example. Let’s use sales data from Norns.

Tea Room Norns

Independent variable

Dependent variable

All right!

We’ll go over that later.

For steps 4 and 5 even mean?

It’s easier to explain with an example. Let’s use sales data from Norns.

Tea Room Norns

Independent variable

Dependent variable

All right!

We’ll go over that later.

Step 1: Draw a scatter plot of the independent variable versus the dependent variable. If the dots line up, the variables may be correlated.

First, draw a scatter plot of the independent variable and the dependent variable.

We’ve done that already.

<table>
<thead>
<tr>
<th>High temp. (°C)</th>
<th>Iced tea orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>22nd (Mon.)</td>
<td>29</td>
</tr>
<tr>
<td>23rd (Tues.)</td>
<td>28</td>
</tr>
<tr>
<td>24th (Wed.)</td>
<td>34</td>
</tr>
<tr>
<td>25th (Thurs.)</td>
<td>31</td>
</tr>
<tr>
<td>26th (Fri.)</td>
<td>25</td>
</tr>
<tr>
<td>27th (Sat.)</td>
<td>29</td>
</tr>
<tr>
<td>28th (Sun.)</td>
<td>32</td>
</tr>
<tr>
<td>29th (Mon.)</td>
<td>31</td>
</tr>
<tr>
<td>30th (Tues.)</td>
<td>34</td>
</tr>
<tr>
<td>31st (Wed.)</td>
<td>33</td>
</tr>
<tr>
<td>1st (Thurs.)</td>
<td>30</td>
</tr>
<tr>
<td>2nd (Fri.)</td>
<td>26</td>
</tr>
<tr>
<td>3rd (Sat.)</td>
<td>31</td>
</tr>
<tr>
<td>4th (Sun.)</td>
<td></td>
</tr>
</tbody>
</table>
When we plot each day's high temperature against iced tea orders, they seem to line up.

And we know from earlier that the value of $R$ is 0.9069, which is pretty high.

It looks like these variables are correlated.

Do you really learn anything from all those dots? Why not just calculate $R$?

The shape of our data is important!

Look at this chart. Rather than flowing in a line, the dots are scattered randomly.

You can still find a regression equation, but it's meaningless. The low $R$ value confirms it, but the scatter plot lets you see it with your own eyes.

Always draw a plot first to get a sense of the data's shape.

Oh, I see. Plots...are...important!
Step 2: Calculate the regression equation.

Now, let's make a regression equation!

Let's find $a$ and $b$!

Finally, the time has come.

$y = ax + b$

Let's draw a straight line, following the pattern in the data as best we can.

The little arrows are the distances from the line, which represents the estimated values of each dot, which are the actual measured values. The distances are called residuals. The goal is to find the line that best minimizes all the residuals.

This is called linear least squares regression.

We square the residuals to find the sum of squares, which we use to find the regression equation.

Step 1: Calculate $S_{xx}$ (sum of squares of $x$), $S_{yy}$ (sum of squares of $y$), and $S_{xy}$ (sum of products of $x$ and $y$).

Step 2: Calculate $S_e$ (residual sum of squares).

Step 3: Differentiate $S_e$ with respect to $a$ and $b$, and set it equal to 0.

Step 4: Separate out $a$ and $b$.

Step 5: Isolate the $a$ component.

Step 6: Find the regression equation.

I'll add this to my notes.
Find

- The sum of squares of $x$, $S_{xx}$: $(x - \bar{x})^2$
- The sum of squares of $y$, $S_{yy}$: $(y - \bar{y})^2$
- The sum of products of $x$ and $y$, $S_{xy}$: $(x - \bar{x})(y - \bar{y})$

Note: The bar over a variable (like $\bar{y}$) is a notation that means average. We can call this variable $x$-bar.

<table>
<thead>
<tr>
<th>High temp. in °C</th>
<th>Iced tea orders</th>
<th>$x$</th>
<th>$y$</th>
<th>$x - \bar{x}$</th>
<th>$y - \bar{y}$</th>
<th>$(x - \bar{x})^2$</th>
<th>$(y - \bar{y})^2$</th>
<th>$(x - \bar{x})(y - \bar{y})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>22nd (Mon.) 29</td>
<td>77</td>
<td>-0.1</td>
<td>4.4</td>
<td>0.0</td>
<td>19.6</td>
<td>-0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23rd (Tues.) 28</td>
<td>62</td>
<td>-1.1</td>
<td>-10.6</td>
<td>1.3</td>
<td>111.8</td>
<td>12.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24th (Wed.) 34</td>
<td>93</td>
<td>4.9</td>
<td>20.4</td>
<td>23.6</td>
<td>417.3</td>
<td>99.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25th (Thurs.) 31</td>
<td>84</td>
<td>1.9</td>
<td>11.4</td>
<td>3.4</td>
<td>130.6</td>
<td>21.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26th (Fri.) 25</td>
<td>59</td>
<td>-4.1</td>
<td>-13.6</td>
<td>17.2</td>
<td>184.2</td>
<td>56.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27th (Sat.) 29</td>
<td>64</td>
<td>-0.1</td>
<td>-8.6</td>
<td>0.0</td>
<td>73.5</td>
<td>1.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28th (Sun.) 32</td>
<td>80</td>
<td>2.9</td>
<td>7.4</td>
<td>8.2</td>
<td>55.2</td>
<td>21.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29th (Mon.) 31</td>
<td>75</td>
<td>1.9</td>
<td>2.4</td>
<td>3.4</td>
<td>5.9</td>
<td>4.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30th (Tues.) 24</td>
<td>58</td>
<td>-5.1</td>
<td>-14.6</td>
<td>26.4</td>
<td>212.3</td>
<td>74.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31st (Wed.) 33</td>
<td>91</td>
<td>3.9</td>
<td>18.4</td>
<td>14.9</td>
<td>339.6</td>
<td>71.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st (Thurs.) 25</td>
<td>51</td>
<td>-4.1</td>
<td>-21.6</td>
<td>17.2</td>
<td>465.3</td>
<td>89.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd (Fri.) 31</td>
<td>73</td>
<td>1.9</td>
<td>0.4</td>
<td>3.4</td>
<td>0.2</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd (Sat.) 26</td>
<td>65</td>
<td>-3.1</td>
<td>-7.6</td>
<td>9.9</td>
<td>57.8</td>
<td>23.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th (Sun.) 30</td>
<td>84</td>
<td>0.9</td>
<td>11.4</td>
<td>0.7</td>
<td>130.6</td>
<td>9.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>408</td>
<td>1016</td>
<td>0</td>
<td>0</td>
<td>129.7</td>
<td>2203.4</td>
<td>484.9</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>29.1</td>
<td>72.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ x \quad y \quad x - \bar{x} \quad y - \bar{y} \quad (x - \bar{x})^2 \quad (y - \bar{y})^2 \quad (x - \bar{x})(y - \bar{y}) \]

\[ S_{xx} \quad S_{yy} \quad S_{xy} \]

*Some of the figures in this chapter are rounded for the sake of printing, but calculations are done using the full, unrounded values resulting from the raw data unless otherwise stated.
Find the residual sum of squares, $S_e$.

- $y$ is the observed value.
- $\hat{y}$ is the estimated value based on our regression equation.
- $y - \hat{y}$ is called the residual and is written as $e$.

Note: The caret in $\hat{y}$ is affectionately called a hat, so we call this parameter estimate $y$-hat.

<table>
<thead>
<tr>
<th>High temp. in °C</th>
<th>Actual iced tea orders</th>
<th>Predicted iced tea orders</th>
<th>Residuals ($e$)</th>
<th>Squared residuals $(y - \hat{y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>22nd (Mon.)</td>
<td>29</td>
<td>$\alpha \times 29 + b$</td>
<td>77 - $(\alpha \times 29 + b)$</td>
<td>$[77 - (\alpha \times 29 + b)]^2$</td>
</tr>
<tr>
<td>23rd (Tues.)</td>
<td>28</td>
<td>$\alpha \times 28 + b$</td>
<td>62 - $(\alpha \times 28 + b)$</td>
<td>$[62 - (\alpha \times 28 + b)]^2$</td>
</tr>
<tr>
<td>24th (Wed.)</td>
<td>34</td>
<td>$\alpha \times 34 + b$</td>
<td>93 - $(\alpha \times 34 + b)$</td>
<td>$[93 - (\alpha \times 34 + b)]^2$</td>
</tr>
<tr>
<td>25th (Thurs.)</td>
<td>31</td>
<td>$\alpha \times 31 + b$</td>
<td>84 - $(\alpha \times 31 + b)$</td>
<td>$[84 - (\alpha \times 31 + b)]^2$</td>
</tr>
<tr>
<td>26th (Fri.)</td>
<td>25</td>
<td>$\alpha \times 25 + b$</td>
<td>59 - $(\alpha \times 25 + b)$</td>
<td>$[59 - (\alpha \times 25 + b)]^2$</td>
</tr>
<tr>
<td>27th (Sat.)</td>
<td>29</td>
<td>$\alpha \times 29 + b$</td>
<td>64 - $(\alpha \times 29 + b)$</td>
<td>$[64 - (\alpha \times 29 + b)]^2$</td>
</tr>
<tr>
<td>28th (Sun.)</td>
<td>32</td>
<td>$\alpha \times 32 + b$</td>
<td>80 - $(\alpha \times 32 + b)$</td>
<td>$[80 - (\alpha \times 32 + b)]^2$</td>
</tr>
<tr>
<td>29th (Mon.)</td>
<td>31</td>
<td>$\alpha \times 31 + b$</td>
<td>75 - $(\alpha \times 31 + b)$</td>
<td>$[75 - (\alpha \times 31 + b)]^2$</td>
</tr>
<tr>
<td>30th (Tues.)</td>
<td>24</td>
<td>$\alpha \times 24 + b$</td>
<td>58 - $(\alpha \times 24 + b)$</td>
<td>$[58 - (\alpha \times 24 + b)]^2$</td>
</tr>
<tr>
<td>31st (Wed.)</td>
<td>33</td>
<td>$\alpha \times 33 + b$</td>
<td>91 - $(\alpha \times 33 + b)$</td>
<td>$[91 - (\alpha \times 33 + b)]^2$</td>
</tr>
<tr>
<td>1st (Thurs.)</td>
<td>25</td>
<td>$\alpha \times 25 + b$</td>
<td>51 - $(\alpha \times 25 + b)$</td>
<td>$[51 - (\alpha \times 25 + b)]^2$</td>
</tr>
<tr>
<td>2nd (Fri.)</td>
<td>31</td>
<td>$\alpha \times 31 + b$</td>
<td>73 - $(\alpha \times 31 + b)$</td>
<td>$[73 - (\alpha \times 31 + b)]^2$</td>
</tr>
<tr>
<td>3rd (Sat.)</td>
<td>26</td>
<td>$\alpha \times 26 + b$</td>
<td>65 - $(\alpha \times 26 + b)$</td>
<td>$[65 - (\alpha \times 26 + b)]^2$</td>
</tr>
<tr>
<td>4th (Sun.)</td>
<td>30</td>
<td>$\alpha \times 30 + b$</td>
<td>84 - $(\alpha \times 30 + b)$</td>
<td>$[84 - (\alpha \times 30 + b)]^2$</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td><strong>408</strong></td>
<td><strong>1016</strong></td>
<td><strong>408a + 14b</strong></td>
<td><strong>1016 - (408a + 14b)</strong></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>29.1</strong></td>
<td><strong>72.6</strong></td>
<td>$\bar{y} \alpha + b$</td>
<td>$\bar{y} - (\bar{y} + b)$</td>
</tr>
</tbody>
</table>

\[
S_e = \frac{\sum (y - \hat{y})^2}{14}
\]

\[
S_e = [77 - (\alpha \times 29 + b)]^2 + \ldots + [84 - (\alpha \times 30 + b)]^2
\]

**THE SUM OF THE RESIDUALS Squared is called the residual sum of squares. It is written as $S_e$ or RSS.**
Differentiate $S_e$ with respect to $a$ and $b$, and set it equal to 0. When differentiating $y = (ax + b)^n$ with respect to $x$, the result is
\[
\frac{dy}{dx} = n (ax + b)^{n-1} \times a.
\]

- Differentiate with respect to $a$.
\[
\frac{dS_e}{da} = 2 [77 - (29a + b)] \times (-29) + \cdots + 2 [84 - (30a + b)] \times (-30) = 0 \tag{1}
\]

- Differentiate with respect to $b$.
\[
\frac{dS_e}{db} = 2 [77 - (29a + b)] \times (-1) + \cdots + 2 [84 - (30a + b)] \times (-1) = 0 \tag{2}
\]

**Step 4**

Rearrange (1) and (2) from the previous step.

Rearrange (1).

\[
2 [77 - (29a + b)] \times (-29) + \cdots + 2 [84 - (30a + b)] \times (-30) = 0 \tag{3}
\]

- Divide both sides by 2.

Rearrange (2).

\[
2 [77 - (29a + b)] \times (-1) + \cdots + 2 [84 - (30a + b)] \times (-1) = 0 \tag{4}
\]

- Divide both sides by 2.

\[
[77 - (29a + b)] \times (-1) + \cdots + [84 - (30a + b)] \times (-1) = 0
\]

- Multiply by -1.

\[
[29a + 29b - 29 \times 77] + \cdots + [30a + 30b - 30 \times 84] = 0
\]

- Multiply.

\[
(29 \times 29a + 29 \times b - 29 \times 77) + \cdots + (30 \times 30a + 30 \times b - 30 \times 84) = 0
\]

- Separate out $a$ and $b$.

\[
29 \times 29a + 29b - 29 \times 77 + \cdots + 30b = b \times 77 + \cdots + 30 \times 84
\]

\[
29 \times 29a + 29b - 29 \times 77 + \cdots + 30b = b \times 77 + \cdots + 30 \times 84
\]

\[
29^2 + \cdots + 30^2 \times a + (29 + \cdots + 30) b = (29 \times 77 + \cdots + 30 \times 84)
\]

- Subtract $14b$ from both sides and multiply by -1.

\[
b = \frac{77 \times \cdots + 84 - 29 \times \cdots + 30}{14} a
\]

\[
b = \bar{y} - \bar{x} \tag{5}
\]

- Isolate $b$ on the left side of the equation.

The components in (5) are the averages of $y$ and $x$. 

74 CHAPTER 2 SIMPLE REGRESSION ANALYSIS
Plug the value of $b$ found in 4 into line 3 (3 and 4 are the results from Step 4).

Now $a$ is the only variable.

Combine the $a$ terms.

Transpose.

Rearrange the left side of the equation.

\[
(29^2 + \ldots + 30^2) - \left( \frac{29 + \ldots + 30}{14} \right)^2
= (29^2 + \ldots + 30^2) - 2 \times \left( \frac{29 + \ldots + 30}{14} \right)^2 + \left( \frac{29 + \ldots + 30}{14} \right)^2 \times 14
= (29^2 + \ldots + 30^2) - 2 \times (29 + \ldots + 30) \times \frac{29 + \ldots + 30}{14} + \left( \frac{29 + \ldots + 30}{14} \right)^2 \times 14
= (29^2 + \ldots + 30^2) - 2 \times (29 + \ldots + 30) \times (29 + \ldots + 30)\times 14 + \left( \frac{29 + \ldots + 30}{14} \right)^2 \times 14
= (29^2 + \ldots + 30^2) - 2 \times (29 + \ldots + 30) \times (x + (x)^2) \times 14
= (29^2 + \ldots + 30^2) - 2 \times (29 + \ldots + 30) \times (x + (x)^2)
= S_x
\]

The last term is multiplied by $\frac{14}{14}$.

We add and subtract $\frac{(29 + \ldots + 30)^2}{14}$.

Rearrange the right side of the equation.

\[
(29 \times 77 + \ldots + 30 \times 84) - \left( \frac{29 + \ldots + 30}{14} \right) \left( \frac{77 + \ldots + 84}{14} \right)
= (29 \times 77 + \ldots + 30 \times 84) - \frac{29 + \ldots + 30}{14} \times \frac{77 + \ldots + 84}{14} \times 14
= (29 \times 77 + \ldots + 30 \times 84) - \frac{29 + \ldots + 30}{14} \times (77 + \ldots + 84) + \frac{29 + \ldots + 30}{14} \times (77 + \ldots + 84)
= (29 \times 77 + \ldots + 30 \times 84) - (29 + \ldots + 30) \times (77 + \ldots + 84) + \frac{29 + \ldots + 30}{14} \times (77 + \ldots + 84)
= (29 - x)(77 - \bar{y}) + \ldots (30 - x)(84 - \bar{y})
= S_{xy}
\]

WE ADD AND SUBTRACT $\bar{x} \times \bar{y} \times 14$.

We add and subtract $\bar{x} \times \bar{y} \times 14$.

\[
S_xa = S_{xy}
\]

Isolate $a$ on the left side of the equation.

\[
a = \frac{S_{xy}}{S_{xx}}
\]
**Step 6** Calculate the regression equation.

From Step 5, \( a = \frac{S_{xy}}{S_{xx}} \). From Step 4, \( b = \bar{y} - \bar{x}a \).

If we plug in the values we calculated in Step 1,

\[
\begin{align*}
\alpha &= \frac{S_{xx}}{S_{xy}} = \frac{484.9}{129.7} = 3.7 \\
b &= \bar{y} - \bar{x}a = 72.6 - 29.1 \times 3.7 = -36.4
\end{align*}
\]

then the regression equation is

\[ y = 3.7x - 36.4. \]

It's that simple!

Note: The values shown are rounded for the sake of printing, but the result (36.4) was calculated using the full, unrounded values.

**THE RELATIONSHIP BETWEEN THE RESIDUALS AND THE SLOPE \( a \) AND INTERCEPT \( b \) IS ALWAYS**

\[
\begin{align*}
\alpha &= \frac{\text{sum of products of } x \text{ and } y}{\text{sum of squares of } x} = \frac{S_{xy}}{S_{xx}} \\
b &= \bar{y} - \bar{x}a
\end{align*}
\]

**THIS IS TRUE FOR ANY LINEAR REGRESSION.**
So, Miu, what are the average values for the high temperature and the iced tea orders?

Let me see...

29.1°C and 72.6 orders.

Remember, the average temperature is \( \bar{x} \) and the average number of orders is \( \bar{y} \). Now for a little magic.

The regression equation can be...

\[
\hat{y} = a\bar{x} + b
\]

\[
= a\bar{x} + (\bar{y} - \bar{x}a)
\]

\[
= a(x - \bar{x}) + \bar{y}
\]

...rearranged like this.

That's from step 4!

Without looking, I can tell you that the regression equation crosses the point (29.1, 72.6).

Now, if we set \( x \) to the average value (\( \bar{x} \)) we found before...

\[
= a(x - \bar{x}) + \bar{y}
\]

\[
= a(\bar{x} - \bar{x}) + \bar{y}
\]

\[
= a\times0 + \bar{y}
\]

\[
= \bar{y}
\]

See what happens?

When \( x \) is the average, so is \( y \)!
Step 3: Calculate the correlation coefficient (r) and assess our population and assumptions.

Next, we’ll determine the accuracy of the regression equation we have come up with.

Why? What will that tell us?

Our data and its regression equation

Example data and its regression equation

Well, the graph on the left has a steeper slope...

Miu, can you see a difference between these two graphs?

Hmm...

The dots are closer to the regression line in the left graph.

Right!
When a regression equation is accurate, the estimated values (the line) are closer to the observed values (dots).

Right. Accuracy is important, but determining it by looking at a graph is pretty subjective.

So accurate means realistic?

Right! We use $R$ to represent an index that measures the accuracy of a regression equation. The index compares our data to our predictions—in other words, the measured $x$ and $y$ to the estimated $\hat{x}$ and $\hat{y}$.

The correlation coefficient from earlier, right?

That's why we need $R$!

Ta-da!

The dots are close.

The dots are kind of far.

Yes, that's true.

Correlation coefficient.

R is also called the Pearson product moment correlation coefficient in honor of mathematician Karl Pearson.

I see!
Here's the equation. We calculate these like we did $S_{xx}$ and $S_{xy}$ before.

$$R = \frac{\text{sum of products } y \text{ and } \hat{y}}{\text{sum of squares of } y \times \text{sum of squares of } \hat{y}} = \frac{S_{yy}}{\sqrt{S_{yy} \times S_{\hat{y}\hat{y}}}}$$

$$= \frac{1812.3}{\sqrt{2203.4 \times 1812.3}} = 0.9069$$

That's not too bad!

Here is the regression function!

<table>
<thead>
<tr>
<th>Actual values</th>
<th>Estimated values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>$\hat{y} = 3.7x - 36.4$</td>
</tr>
<tr>
<td>$y - \bar{y}$</td>
<td>$\hat{y} - \bar{\hat{y}}$</td>
</tr>
<tr>
<td>$(y - \bar{y})^2$</td>
<td>$(\hat{y} - \bar{\hat{y}})^2$</td>
</tr>
<tr>
<td>$(y - \bar{y})(\hat{y} - \bar{\hat{y}})$</td>
<td>$(y - \hat{y})^2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>22nd (Mon.)</th>
<th>77</th>
<th>72.0</th>
<th>4.4</th>
<th>-0.5</th>
<th>19.6</th>
<th>0.3</th>
<th>-2.4</th>
<th>24.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>23rd (Tues.)</td>
<td>62</td>
<td>68.3</td>
<td>-10.6</td>
<td>-4.3</td>
<td>111.8</td>
<td>18.2</td>
<td>45.2</td>
<td>39.7</td>
</tr>
<tr>
<td>24th (Wed.)</td>
<td>93</td>
<td>90.7</td>
<td>20.4</td>
<td>18.2</td>
<td>417.3</td>
<td>329.6</td>
<td>370.9</td>
<td>5.2</td>
</tr>
<tr>
<td>25th (Thurs.)</td>
<td>84</td>
<td>79.5</td>
<td>11.4</td>
<td>6.9</td>
<td>130.6</td>
<td>48.2</td>
<td>79.3</td>
<td>20.1</td>
</tr>
<tr>
<td>26th (Fri.)</td>
<td>59</td>
<td>57.1</td>
<td>-13.6</td>
<td>-15.5</td>
<td>184.2</td>
<td>239.8</td>
<td>210.2</td>
<td>3.7</td>
</tr>
<tr>
<td>27th (Sat.)</td>
<td>64</td>
<td>72.0</td>
<td>-8.6</td>
<td>-0.5</td>
<td>73.5</td>
<td>0.3</td>
<td>4.6</td>
<td>64.6</td>
</tr>
<tr>
<td>28th (Sun.)</td>
<td>80</td>
<td>83.3</td>
<td>7.4</td>
<td>10.7</td>
<td>55.2</td>
<td>114.1</td>
<td>79.3</td>
<td>10.6</td>
</tr>
<tr>
<td>29th (Mon.)</td>
<td>75</td>
<td>79.5</td>
<td>2.4</td>
<td>6.9</td>
<td>5.9</td>
<td>48.2</td>
<td>16.9</td>
<td>20.4</td>
</tr>
<tr>
<td>30th (Tues.)</td>
<td>58</td>
<td>53.3</td>
<td>-14.6</td>
<td>-19.2</td>
<td>212.3</td>
<td>369.5</td>
<td>280.1</td>
<td>21.6</td>
</tr>
<tr>
<td>31st (Wed.)</td>
<td>91</td>
<td>87.0</td>
<td>18.4</td>
<td>14.4</td>
<td>339.6</td>
<td>207.9</td>
<td>265.7</td>
<td>16.1</td>
</tr>
<tr>
<td>1st (Thurs.)</td>
<td>51</td>
<td>57.1</td>
<td>-21.6</td>
<td>-15.5</td>
<td>465.3</td>
<td>239.8</td>
<td>334.0</td>
<td>37.0</td>
</tr>
<tr>
<td>2nd (Fri.)</td>
<td>73</td>
<td>79.5</td>
<td>0.4</td>
<td>6.9</td>
<td>0.2</td>
<td>48.2</td>
<td>3.0</td>
<td>42.4</td>
</tr>
<tr>
<td>3rd (Sat.)</td>
<td>65</td>
<td>60.8</td>
<td>-7.6</td>
<td>-11.7</td>
<td>57.3</td>
<td>138.0</td>
<td>88.9</td>
<td>17.4</td>
</tr>
<tr>
<td>4th (Sun.)</td>
<td>84</td>
<td>75.8</td>
<td>11.4</td>
<td>3.2</td>
<td>130.6</td>
<td>10.3</td>
<td>36.6</td>
<td>67.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sum</th>
<th>1016</th>
<th>1016</th>
<th>0</th>
<th>0</th>
<th>2203.4</th>
<th>1812.3</th>
<th>1812.3</th>
<th>391.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>72.6</td>
<td>72.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$S_e$ isn't necessary for calculating $R$, but I included it because we'll need it later.
If we square $R$, it's called the coefficient of determination and is written as $R^2$.

$R^2$ can be an indicator of...

...how much variance is explained by our regression equation.

$R^2$ can be an indicator of...

An $R^2$ of zero indicates that the outcome variable can't be reliably predicted from the predictor variable.

The higher the accuracy of the regression equation, the closer the $R^2$ value is to 1, and vice versa.

But generally we want a value of at least .5.

So how high does $R^2$ need to be for the regression equation to be considered accurate?

Unfortunately, there is no universal standard in statistics.

Now try finding the value of $R^2$.

$R^2 = (0.9069)^2 = 0.8225$

It's .8225.

Sure thing.

Lowest... .5...
The value of $R^2$ for our regression equation is well over .5, so our equation should be able to estimate iced tea orders relatively accurately.

Jot this equation down. $R^2$ can be calculated directly from these values. Using our norm data, \[ 1 - \frac{391.1}{2203.4} = 0.8225 \]

That's handy!

Now to assess the population and verify that our assumptions are met!

I meant to ask you about that. What population? Japan? Earth?

Actually, the population we're talking about isn't people—it's data.

\[
R^2 = \left( \frac{\text{correlation}}{\text{coefficient}} \right)^2 = \frac{a \times S_{xy}}{S_{yy}} = 1 - \frac{S_e}{S_{yy}}
\]
Here, look at the tea room data again.

How many days are there with a high temperature of 31°C?

The 25th, 29th, and 2nd... so three.

I can make a chart like this from your answer.

Now, consider that...

...these three days are not the only days in history with a high of 31°C, are they?

There must have been many others in the past, and there will be many more in the future, right?

Of course.

<table>
<thead>
<tr>
<th>Date</th>
<th>High temp. (°C)</th>
<th>Iced tea orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>22nd (Mon.)</td>
<td>29</td>
<td>77</td>
</tr>
<tr>
<td>23rd (Tues.)</td>
<td>28</td>
<td>62</td>
</tr>
<tr>
<td>24th (Wed.)</td>
<td>34</td>
<td>93</td>
</tr>
<tr>
<td>25th (Thurs.)</td>
<td>31</td>
<td>84</td>
</tr>
<tr>
<td>26th (Fri.)</td>
<td>25</td>
<td>59</td>
</tr>
<tr>
<td>27th (Sat.)</td>
<td>29</td>
<td>64</td>
</tr>
<tr>
<td>28th (Sun.)</td>
<td>32</td>
<td>80</td>
</tr>
<tr>
<td>29th (Mon.)</td>
<td>31</td>
<td>75</td>
</tr>
<tr>
<td>30th (Tues.)</td>
<td>24</td>
<td>58</td>
</tr>
<tr>
<td>31st (Wed.)</td>
<td>33</td>
<td>91</td>
</tr>
<tr>
<td>1st (Thurs.)</td>
<td>25</td>
<td>51</td>
</tr>
<tr>
<td>2nd (Fri.)</td>
<td>31</td>
<td>73</td>
</tr>
<tr>
<td>3rd (Sat.)</td>
<td>26</td>
<td>65</td>
</tr>
<tr>
<td>4th (Sun.)</td>
<td>30</td>
<td>84</td>
</tr>
</tbody>
</table>

So...
These three days are a sample...

...from the population of all days with a high temperature of 31°C. We use sample data when it's unlikely we'll be able to get the information we need from every single member of the population.

That makes sense.

Samples represent the population.

Ice tea orders.

Thanks, Risa. I get it now.

Good! On to diagnostics, then.
Assumptions of Normality

A regression equation is meaningful only if a certain hypothesis is viable.

Alternative Hypothesis

The number of orders of iced tea on days with temperature $x^\circ C$ follows a normal distribution with mean $Ax+B$ and standard deviation $\sigma$ (Sigma).

Let's take it slow. First look at the shapes on this graph.

These shapes represent the entire population of iced tea orders for each high temperature. Since we can't possibly know the exact distribution for each temperature, we have to assume that they must all be the same: a normal, bell-shaped curve.
“MUST ALL BE THE SAME”? WOULDN’T THE DISTRIBUTIONS BE SLIGHTLY DIFFERENT?

COULD THEY DIFFER ACCORDING TO TEMPERATURE?

THEY’RE NEVER EXACTLY THE SAME.

GOOD POINT.

BUT WE MUST ASSUME THAT THEY ARE! REGRESSION Depends ON THE ASSUMPTION OF NORMALITY!

YOU’RE A SHARP ONE...

AX + B IS CALLED THE POPULATION REGRESSION. THE EXPRESSION ax+b IS THE SAMPLE REGRESSION.

I CAN DO THAT.

BY THE WAY, AX+B I’LL PUT THAT IN MY NOTES.

Ax+B
(Population Regression)

86  CHAPTER 2  SIMPLE REGRESSION ANALYSIS
STEP 4: CONDUCT THE ANALYSIS OF VARIANCE.

NOW, LET'S RETURN TO THE STORY ABOUT A, B, AND σ.

A, LIKE A, IS A SLOPE. B, LIKE B, IS AN INTERCEPT. AND σ IS THE STANDARD DEVIATION.

A, B, AND σ ARE COEFFICIENTS OF THE ENTIRE POPULATION.

IF THE REGRESSION EQUATION IS \( y = ax+b \)

- \( a \) SHOULD BE CLOSE TO A
- \( b \) SHOULD BE CLOSE TO B
- \( \frac{s}{\sqrt{\text{number of individuals} - 2}} \) SHOULD BE CLOSE TO σ

DO YOU RECALL \( a \), \( b \), AND THE STANDARD DEVIATION FOR OUR NORMS DATA?

WELL, THE REGRESSION EQUATION WAS \( y = 3.7x - 36.4 \), SO...

- A IS ABOUT 3.7
- B IS ABOUT -36.4
- \( \sigma \) IS ABOUT \( \sqrt{\frac{391.1}{14-2}} = \sqrt{\frac{391.1}{12}} = 5.7 \)

IS THAT RIGHT?

PERFECT!
"CLOSE TO" SEEMS SO VAGUE. CAN'T WE FIND A, B, AND σ WITH MORE CERTAINTY?

However...

SINCE A, B, AND σ ARE COEFFICIENTS OF THE POPULATION, WE'D NEED TO USE ALL THE NORMS ICED TEA AND HIGH TEMPERATURE DATA THROUGHOUT HISTORY! WE COULD NEVER GET IT ALL.

...WE CAN DETERMINE ONCE AND FOR ALL WHETHER A = 0!

TAH-RUH!

...IMAGINE IF A WERE ZERO...

YOU SHOULD LOOK MORE EXCITED! THIS IS IMPORTANT!
**NULL HYPOTHESIS**

The number of orders of iced tea on days with high temperature $x$ °C follows a normal distribution with mean $B$ and standard deviation $\sigma$. (A is absent!)

**A ≠ 0**

If the slope $A = 0$, the line is horizontal. That means iced tea orders are the same, no matter what the high temperature is!

**A = 0**

The temperature doesn't matter!

How do we find out about $A$? **ANOVA**

Let's do the analysis and see what fate has in store for $A$.

This is getting exciting.

Assumptions of Normality 89
### The Steps of ANOVA

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Equation/Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td>Define the population. The population is “days with a high temperature of ( x ) degrees.”</td>
<td></td>
</tr>
<tr>
<td>Step 2</td>
<td>Set up a null hypothesis and an alternative hypothesis. Null hypothesis is ( A = 0 ). Alternative hypothesis is ( A \neq 0 ).</td>
<td></td>
</tr>
<tr>
<td>Step 3</td>
<td>Select which hypothesis test to conduct. We’ll use analysis of one-way variance.</td>
<td></td>
</tr>
<tr>
<td>Step 4</td>
<td>Choose the significance level. We’ll use a significance level of .05.</td>
<td></td>
</tr>
</tbody>
</table>
| Step 5 | Calculate the test statistic from the sample data. The test statistic is: \[
\frac{a^2}{\frac{1}{S_{xx}}} \frac{S_s}{\text{number of individuals} - 2}
\] Plug in the values from our sample regression equation: \[
\frac{\frac{3.7^2}{1}}{129.7} \frac{391.1}{14 - 2} = 55.6
\] The test statistic will follow an \( F \) distribution with first degree of freedom 1 and second degree of freedom 12 (number of individuals minus 2), if the null hypothesis is true. |                                                                                   |
| Step 6 | Determine whether the \( p \)-value for the test statistic obtained in Step 5 is smaller than the significance level. At significance level .05, with \( d_1 \) being 1 and \( d_2 \) being 12, the critical value is 4.7472. Our test statistic is 55.6. |                                                                                   |
| Step 7 | Decide whether you can reject the null hypothesis. Since our test statistic is greater than the critical value, we reject the null hypothesis. |                                                                                   |

The \( F \) statistic lets us test the slope of the line by looking at variance. If the variation around the line is much smaller than the total variance of \( Y \), that’s evidence that the line accounts for \( Y \)'s variation, and the statistic will be large. If the ratio is small, the line doesn’t account for much variation in \( Y \), and probably isn’t useful! So \( A \neq 0 \), what a relief!
STEP 5: CALCULATE THE CONFIDENCE INTERVALS.

NOW, LET’S TAKE A CLOSER LOOK AT HOW WELL OUR REGRESSION EQUATION REPRESENTS THE POPULATION.

IN THE POPULATION...

...LOTS OF DAYS HAVE A HIGH OF 31°C, AND THE NUMBER OF ICED TEA ORDERS ON THOSE DAYS VARIES. OUR REGRESSION EQUATION PREDICTS ONLY ONE VALUE FOR ICED TEA ORDERS AT THAT TEMPERATURE.

HOW DO WE KNOW THAT IT’S THE RIGHT VALUE?

WE CAN’T KNOW FOR SURE. WE CHOOSE THE MOST LIKELY VALUE: THE POPULATION MEAN.

OKAY, I’M READY!

IF THE POPULATION HAS A NORMAL DISTRIBUTION...

DAYS WITH A HIGH OF 31°C CAN EXPECT APPROXIMATELY THE MEAN NUMBER OF ICED TEA ORDERS. WE CAN’T KNOW THE EXACT MEAN, BUT WE CAN ESTIMATE A RANGE IN WHICH IT MIGHT FALL.

THE MEAN NUMBER OF ORDERS IS SOMEWHERE IN HERE.

HUH? THE RANGES DIFFER, DEPENDING ON THE VALUE OF x!
We calculate an interval for each temperature. As you noticed, the width varies. It's smaller near $\bar{x}$, which is the average high temperature value.

Even this interval isn't absolutely guaranteed to contain the true population mean. Our confidence is determined by the confidence coefficient.

Confidence interval

$23^\circ C$

More orders

Fewer orders

There is no equation to calculate it, no set rule.

You choose the confidence coefficient, and you can make it any percentage you want.

I will make it 42%.

When calculating a confidence interval, you choose the confidence coefficient first.

You would then say "A 42% confidence interval for iced tea orders when the temperature is $31^\circ C$ is 30 to 35 orders," for example!

Now, confidence...

...is no ordinary coefficient.

Sounds familiar!
Hey, if our confidence is based on the coefficient, isn't higher better?

Well, not necessarily.

Well, most people choose either 95% or 99%.

42% is too low to be very meaningful.

Hey, if our confidence is based on the coefficient, isn't higher better?

True, our confidence is higher when we choose 99%, but the interval will be much larger, too.

Hey, if our confidence is based on the coefficient, isn't higher better?

Now, shall we calculate the confidence interval for the population of days with a high temperature of 31°C?

Yes, let's!

Hm, I see.

Which should I choose?

95%

99%

Well, most people choose either 95% or 99%.

42% is too low to be very meaningful.

Hey, if our confidence is based on the coefficient, isn't higher better?

Yes, let's!

The number of orders of iced tea is probably between 40 and 80!

That's not surprising.

The number of orders of iced tea is almost certainly between 0 and 120!

Oh sure...

Sweet!

Now, shall we calculate the confidence interval for the population of days with a high temperature of 31°C?

Yes, let's!

However, if the confidence coefficient is too low, the result is not convincing.

Hm, I see.

True, our confidence is higher when we choose 99%, but the interval will be much larger, too.

Oh sure...

The number of orders of iced tea is probably between 40 and 80!

That's not surprising.

The number of orders of iced tea is almost certainly between 0 and 120!

Well, most people choose either 95% or 99%.

42% is too low to be very meaningful.

Hey, if our confidence is based on the coefficient, isn't higher better?

Well, not necessarily.

Hey, if our confidence is based on the coefficient, isn't higher better?

Yes, let's!

However, if the confidence coefficient is too low, the result is not convincing.

Hm, I see.
Here’s how to calculate a 95% confidence interval for iced tea orders on days with a high of 31°C.

This is the confidence interval.

\[ 79.5^* - 3.9 = 75.6 \quad 31 \times a + b = 31 \times 3.7 - 36.4 \quad 79.5 + 3.9 = 83.4 \]

Number of orders of iced tea

\[ 31 \times a + b \]

\[ 79.5 \]

\[ 3.9 \]

Distance from the estimated mean is

\[ \sqrt{F(1, n - 2; .05) \times \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \times \frac{S_e}{n - 2}} \]

\[ = \sqrt{F(1,14 - 2; .05) \times \left( \frac{1}{14} + \frac{(31 - 29.1)^2}{129.7} \right) \times \frac{391.1}{14 - 2}} \]

\[ = 3.9 \]

where \( n \) is the number of data points in our sample and \( F \) is a ratio of two chi-squared distributions, as described on page 57.

To calculate a 99% confidence interval, just change

\[ F(1,14 - 2; .05) = 4.7 \]

to

\[ F(1,14 - 2; .01) = 9.3 \]

(refer to page 58 for an explanation of \( F(1, n - 2; .05) = 4.7 \), and so on.)

* The value 79.5 was calculated using unrounded numbers.

So we are 95% sure that, if we look at the population of days with a high of 31°C, the mean number of iced tea orders is between 76 and 83.

Exactly!
STEP 6: MAKE A PREDICTION!

AT LAST, WE MAKE THE PREDICTION.

THE FINAL STEP!

IF TOMORROW'S HIGH TEMPERATURE IS 27°C...

...HOW MANY ICED TEA ORDERS WILL WE GET AT THE SHOP TOMORROW?

HMM, THE REGRESSION EQUATION IS

\[ y = 3.7x - 36.4 \]

\[ y = 3.7 \times 27 - 36.4 \]

\[ = 63.5 \]

\[ \approx 64 \]

BINGO!

* This calculation was performed using rounded figures. If you're doing the calculation with the full, unrounded figures, you should get 64.6.
But will there be exactly 64 orders?

How can we possibly know for sure?

That's a great question.

We should get close to 64 orders because the value of $R^2$ is 0.8225, but... how close?

We'll make a prediction interval!

We'll pick a coefficient and then calculate a range in which iced tea orders will most likely fall.

Didn't we just do that?

Not quite. Before, we were predicting the mean number of iced tea orders for the population of days with a certain high temperature, but now we're predicting the likely number of iced tea orders on a given day with a certain temperature.

I don't see the difference.
Confidence intervals help us assess the population.

The prediction interval looks like a confidence interval, but it’s not the same.

As with a confidence interval, we need to choose the confidence coefficient before we can do the calculation. Again, 95% and 99% are popular.

I just calculated an interval, so this should be a snap.

The calculation is very similar, with one important difference...

The prediction interval will be wider because it covers the range of all expected values, not just where the mean should be.

Now, try calculating the prediction interval for 27°C.

The future is always surprising.

No sweat!
Here’s how we calculate a 95% prediction interval for tomorrow’s iced tea sales.

This is the prediction interval.

\[
\begin{align*}
64.6 - 13.1 &= 51.5 \\
27 \times a + b &= 27 \times 3.7 - 36.4 \\
&= 64.6 \\
64.6 + 13.1 &= 77.7^* 
\end{align*}
\]

Distance from the estimated value is

\[
\sqrt{F(1, n - 2; .05) \times \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \times \frac{S_e}{n - 2}}
\]

\[
= \sqrt{F(1, 14 - 2; .05) \times \left(1 + \frac{1}{14} + \frac{(27 - 29.1)^2}{129.7}\right) \times \frac{391.1}{14 - 2}}
\]

\[
= 13.1
\]

The estimated number of tea orders we calculated earlier (on page 95) was rounded, but we’ve used the number of tea orders estimated using unrounded numbers, 64.6, here.

Here we used the F distribution to find the prediction interval and population regression. Typically, statisticians use the t distribution to get the same results.

\[
* \text{This calculation was performed using the rounded numbers shown here. The full, unrounded calculation results in 77.6.}
\]

So we’re 95% confident that the number of iced tea orders will be between 52 and 78 when the high temperature for that day is 27°C.
Asumptions of normality

What are you staring at?

Oh! I was daydreaming.

You made it through today's lesson.

How was it?

It was difficult at times...

...but I'm catching on. I think I can do this.

And predicting the future is really exciting!

We can make all kinds of predictions about the future.

Like, how many days until you finally talk to him.

Yeah, it rocks!

Heh heh!
WHICH STEPS ARE NECESSARY?

Remember the regression analysis procedure introduced on page 68?

1. Draw a scatter plot of the independent variable versus the dependent variable. If the dots line up, the variables may be correlated.

2. Calculate the regression equation.

3. Calculate the correlation coefficient ($R$) and assess our population and assumptions.

4. Conduct the analysis of variance.

5. Calculate the confidence intervals.

6. Make a prediction!

In this chapter, we walked through each of the six steps, but it isn’t always necessary to do every step. Recall the example of Miu’s age and height on page 25.

- Fact: There is only one Miu in this world.

- Fact: Miu’s height when she was 10 years old was 137.5 cm.

Given these two facts, it makes no sense to say that “Miu’s height when she was 10 years old follows a normal distribution with mean $Ax + B$ and standard deviation $\sigma$. In other words, it’s nonsense to analyze the population of Miu’s heights at 10 years old. She was just one height, and we know what her height was.

In regression analysis, we either analyze the entire population or, much more commonly, analyze a sample of the larger population. When you analyze a sample, you should perform all the steps. However, since Steps 4 and 5 assess how well the sample represents the population, you can skip them if you’re using data from an entire population instead of just a sample.

**NOTE** We use the term statistic to describe a measurement of a characteristic from a sample, like a sample mean, and parameter to describe a measurement that comes from a population, like a population mean or coefficient.

STANDARDIZED RESIDUAL

Remember that a residual is the difference between the measured value and the value estimated with the regression equation. The standardized residual is the residual divided by its estimated standard deviation. We use the standardized residual to assess whether a particular measurement deviates significantly from
the trend. For example, say a group of thirsty joggers stopped by Norns on the 4th, meaning that though iced tea orders were expected to be about 76 based on that day's high temperature, customers actually placed 84 orders for iced tea. Such an event would result in a large standardized residual.

Standardized residuals are calculated by dividing each residual by an estimate of its standard deviation, which is calculated using the residual sum of squares. The calculation is a little complicated, and most statistics software does it automatically, so we won't go into the details of the calculation here.

Table 2-1 shows the standardized residual for the Norns data used in this chapter.

**Table 2-1: Calculating the Standardized Residual**

<table>
<thead>
<tr>
<th>High temperature</th>
<th>Measured number of orders of iced tea</th>
<th>Estimated number of orders of iced tea $\hat{y} = 3.7x - 36.4$</th>
<th>Residual $y - \hat{y}$</th>
<th>Standardized residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>22nd (Mon.)</td>
<td>29</td>
<td>77</td>
<td>72.0</td>
<td>5.0</td>
</tr>
<tr>
<td>23rd (Tues.)</td>
<td>28</td>
<td>62</td>
<td>68.3</td>
<td>–6.3</td>
</tr>
<tr>
<td>24th (Wed.)</td>
<td>34</td>
<td>93</td>
<td>90.7</td>
<td>2.3</td>
</tr>
<tr>
<td>25th (Thurs.)</td>
<td>31</td>
<td>84</td>
<td>79.5</td>
<td>4.5</td>
</tr>
<tr>
<td>26th (Fri.)</td>
<td>25</td>
<td>59</td>
<td>57.1</td>
<td>1.9</td>
</tr>
<tr>
<td>27th (Sat.)</td>
<td>29</td>
<td>64</td>
<td>72.0</td>
<td>–8.0</td>
</tr>
<tr>
<td>28th (Sun.)</td>
<td>32</td>
<td>80</td>
<td>83.3</td>
<td>–3.3</td>
</tr>
<tr>
<td>29th (Mon.)</td>
<td>31</td>
<td>75</td>
<td>79.5</td>
<td>–4.5</td>
</tr>
<tr>
<td>30th (Tues.)</td>
<td>24</td>
<td>58</td>
<td>53.3</td>
<td>4.7</td>
</tr>
<tr>
<td>31st (Wed.)</td>
<td>33</td>
<td>91</td>
<td>87.0</td>
<td>4.0</td>
</tr>
<tr>
<td>1st (Thurs.)</td>
<td>25</td>
<td>51</td>
<td>57.1</td>
<td>–6.1</td>
</tr>
<tr>
<td>2nd (Fri.)</td>
<td>31</td>
<td>73</td>
<td>79.5</td>
<td>–6.5</td>
</tr>
<tr>
<td>3rd (Sat.)</td>
<td>26</td>
<td>65</td>
<td>60.8</td>
<td>4.2</td>
</tr>
<tr>
<td>4th (Sun.)</td>
<td>30</td>
<td>84</td>
<td>75.8</td>
<td>8.2</td>
</tr>
</tbody>
</table>

As you can see, the standardized residual on the 4th is 1.5. If iced tea orders had been 76, as expected, the standardized residual would have been 0.

Sometimes a measured value can deviate so much from the trend that it adversely affects the analysis. If the standardized residual is greater than 3 or less than –3, the measurement is considered an outlier. There are a number of ways to handle outliers, including removing them, changing them to a set value, or just keeping them in the analysis as is. To determine which approach is most appropriate, investigate the underlying cause of the outliers.
INTERPOLATION AND EXTRAPOLATION

If you look at the x values (high temperature) on page 64, you can see that the highest value is 34°C and the lowest value is 24°C. Using regression analysis, you can interpolate the number of iced tea orders on days with a high temperature between 24°C and 34°C and extrapolate the number of iced tea orders on days with a high below 24°C or above 34°C. In other words, extrapolation is the estimation of values that fall outside the range of your observed data.

Since we've only observed the trend between 24°C and 34°C, we don't know whether iced tea sales follow the same trend when the weather is extremely cold or extremely hot. Extrapolation is therefore less reliable than interpolation, and some statisticians avoid it entirely.

For everyday use, it's fine to extrapolate—as long as you're aware that your result isn't completely trustworthy. However, avoid using extrapolation in academic research or to estimate a value that's far beyond the scope of the measured data.

AUTOCORRELATION

The independent variable used in this chapter was high temperature; this is used to predict iced tea sales. In most places, it's unlikely that the high temperature will be 20°C one day and then shoot up to 30°C the next day. Normally, the temperature rises or drops gradually over a period of several days, so if the two variables are related, the number of iced tea orders should rise or drop gradually as well. Our assumption, however, has been that the deviation (error) values are random. Therefore, our predicted values do not change from day to day as smoothly as they might in real life.

When analyzing variables that may be affected by the passage of time, it's a good idea to check for autocorrelation. Autocorrelation occurs when the error is correlated over time, and it can indicate that you need to use a different type of regression model.

There's an index to describe autocorrelation—the Durbin-Watson statistic, which is calculated as follows:

\[ d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \]
The equation can be read as “the sum of the square of each residual minus the previous residual, divided by the sum of each residual squared.” You can calculate the value of the Durbin-Watson statistic for the example in this chapter:

$$\frac{(-6.3 - 5.0)^2 + (2.3 - (-6.3))^2 + \cdots + (8.2 - 4.2)^2}{5.0^2 + (-6.3)^2 + \cdots + 8.2^2} = 1.8$$

The exact critical value of the Durbin-Watson test differs for each analysis, and you can use a table to find it, but generally we use 1 as a cutoff: a result less than 1 may indicate the presence of autocorrelation. This result is close to 2, so we can conclude that there is no autocorrelation in our example.

**NONLINEAR REGRESSION**

On page 66, Risa said:

This equation is linear, but regression equations don’t have to be linear. For example, these equations may also be used as regression equations:

- $y = \frac{a}{x} + b$
- $y = a\sqrt{x} + b$
- $y = ax^2 + bx + c$
- $y = a \times \log x + b$

The regression equation for Miu’s age and height introduced on page 26 is actually in the form of $y = \frac{a}{x} + b$ rather than $y = ax + b$. 
Of course, this raises the question of which type of equation you should choose when performing regression analysis on your own data. Below are some steps that can help you decide.

1. Draw a scatter plot of the data points, with the dependent variable values on the x-axis and the independent variable values on the y-axis. Examine the relationship between the variables suggested by the spread of the dots: Are they in roughly a straight line? Do they fall along a curve? If the latter, what is the shape of the curve?

2. Try the regression equation suggested by the shape in the variables plotted in Step 1. Plot the residuals (or standardized residuals) on the y-axis and the independent variable on the x-axis. The residuals should appear to be random, so if there is an obvious pattern in the residuals, like a curved shape, this suggests that the regression equation doesn’t match the shape of the relationship.

3. If the residuals plot from Step 2 shows a pattern in the residuals, try a different regression equation and repeat Step 2. Try the shapes of several regression equations and pick one that appears to most closely match the data. It’s usually best to pick the simplest equation that fits the data well.

TRANSFORMING NONLINEAR EQUATIONS INTO LINEAR EQUATIONS

There’s another way to deal with nonlinear equations: simply turn them into linear equations. For an example, look at the equation for Miu’s age and height (from page 26):

\[ y = -\frac{326.6}{x} + 173.3 \]

You can turn this into a linear equation. Remember:

\[ \text{If } \frac{1}{x} = X, \text{ then } \frac{1}{X} = x. \]

So we’ll define a new variable \( X \), set it equal to \( \frac{1}{x} \), and use \( X \) in the normal \( y = aX + b \) regression equation. As shown on page 76, the value of \( a \) and \( b \) in the regression equation \( y = aX + b \) can be calculated as follows:

\[
\begin{align*}
\alpha &= \frac{S_{xy}}{S_{xx}} \\
\beta &= \bar{y} - \bar{X}\alpha
\end{align*}
\]
We continue with the analysis as usual. See Table 2-2.

<table>
<thead>
<tr>
<th>Age (x)</th>
<th>$\frac{1}{x}$</th>
<th>Height (y)</th>
<th>$(X - \bar{X})$</th>
<th>$y - \bar{y}$</th>
<th>$(X - \bar{X})^2$</th>
<th>$(y - \bar{y})^2$</th>
<th>$(X - \bar{X})(y - \bar{y})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.2500</td>
<td>100.1</td>
<td>0.1428</td>
<td>-38.1625</td>
<td>0.0204</td>
<td>1456.3764</td>
<td>-5.4515</td>
</tr>
<tr>
<td>5</td>
<td>0.2000</td>
<td>107.2</td>
<td>0.0928</td>
<td>-31.0625</td>
<td>0.0086</td>
<td>964.8789</td>
<td>-2.8841</td>
</tr>
<tr>
<td>6</td>
<td>0.1667</td>
<td>114.1</td>
<td>0.0595</td>
<td>-24.1625</td>
<td>0.0035</td>
<td>583.8264</td>
<td>-1.4381</td>
</tr>
<tr>
<td>7</td>
<td>0.1429</td>
<td>121.7</td>
<td>0.0357</td>
<td>-16.5625</td>
<td>0.0013</td>
<td>274.3164</td>
<td>-0.5914</td>
</tr>
<tr>
<td>8</td>
<td>0.1250</td>
<td>126.8</td>
<td>0.0178</td>
<td>-11.4625</td>
<td>0.0003</td>
<td>131.3889</td>
<td>-0.2046</td>
</tr>
<tr>
<td>9</td>
<td>0.1111</td>
<td>130.9</td>
<td>0.0040</td>
<td>-7.3625</td>
<td>0.0000</td>
<td>54.2064</td>
<td>-0.0292</td>
</tr>
<tr>
<td>10</td>
<td>0.1000</td>
<td>137.5</td>
<td>-0.0072</td>
<td>-0.7625</td>
<td>0.0001</td>
<td>0.5814</td>
<td>-0.0055</td>
</tr>
<tr>
<td>11</td>
<td>0.0909</td>
<td>143.2</td>
<td>-0.0162</td>
<td>4.9375</td>
<td>0.0003</td>
<td>24.3789</td>
<td>-0.0802</td>
</tr>
<tr>
<td>12</td>
<td>0.0833</td>
<td>149.4</td>
<td>-0.0238</td>
<td>11.1375</td>
<td>0.0006</td>
<td>124.0439</td>
<td>-0.2653</td>
</tr>
<tr>
<td>13</td>
<td>0.0769</td>
<td>151.6</td>
<td>-0.0302</td>
<td>13.3375</td>
<td>0.0009</td>
<td>177.889</td>
<td>-0.4032</td>
</tr>
<tr>
<td>14</td>
<td>0.0714</td>
<td>154.0</td>
<td>-0.0357</td>
<td>15.7375</td>
<td>0.0013</td>
<td>247.6689</td>
<td>-0.5622</td>
</tr>
<tr>
<td>15</td>
<td>0.0667</td>
<td>154.6</td>
<td>-0.0405</td>
<td>16.3375</td>
<td>0.0016</td>
<td>266.9139</td>
<td>-0.6614</td>
</tr>
<tr>
<td>16</td>
<td>0.0625</td>
<td>155.0</td>
<td>-0.0447</td>
<td>16.7375</td>
<td>0.0020</td>
<td>280.1439</td>
<td>-0.7473</td>
</tr>
<tr>
<td>17</td>
<td>0.0588</td>
<td>155.1</td>
<td>-0.0483</td>
<td>16.8375</td>
<td>0.0023</td>
<td>283.5014</td>
<td>-0.8137</td>
</tr>
<tr>
<td>18</td>
<td>0.0556</td>
<td>155.3</td>
<td>-0.0516</td>
<td>17.0375</td>
<td>0.0027</td>
<td>290.2764</td>
<td>-0.8790</td>
</tr>
<tr>
<td>19</td>
<td>0.0526</td>
<td>155.7</td>
<td>-0.0545</td>
<td>17.4375</td>
<td>0.0030</td>
<td>304.0664</td>
<td>-0.9507</td>
</tr>
</tbody>
</table>

Sum: 184  1.7144  2212.2  0.0000  0.0000  0.0489  5464.4575  -15.9563

Average: 11.5  0.1072  138.3

According to the table:

$$a = \frac{S_{xy}}{S_{xx}} = \frac{-15.9563}{0.0489} = -326.6$$
$$b = \bar{y} - \bar{X}a = 138.2625 - 0.1072 \times (-326.6) = 173.3$$

So the regression equation is this:

$$y = -326.6X + 173.3$$

* If your result is slightly different from 326.6, the difference might be due to rounding. If so, it should be very small.
which is the same as this:

\[ y = -\frac{326.6}{x} + 173.3 \]

\[ \uparrow \quad \uparrow \]

height    age

We’ve transformed our original, nonlinear equation into a linear one!