



Interpreting Residual Plots to Improve Your Regression

Feedback

WHAT'S ON THIS PAGE:

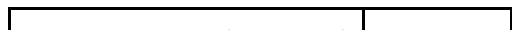


- Observations, Predictions, and Residuals
- Understanding Accuracy with Observed vs. Predicted
- Examining Predicted vs. Residual ("The Residual Plot")
- How much does it matter if my model isn't perfect?
- Example Residual Plots and Their Diagnoses
- Improving Your Model: Assessing the Impact of an Outlier
- Improving Your Model: Transforming Variables
- Improving Your Model: Missing Variables
- Improving Your Model: Fixing Nonlinearity
- FAQs

When you run a regression, Stats iQ automatically calculates and plots residuals to help you understand and improve your regression model. Read below to learn everything you need to know about interpreting residuals (including definitions and examples).

Observations, Predictions, and Residuals

To demonstrate how to interpret residuals, we'll use a lemonade stand data set, where each row was a day of "Temperature" and "Revenue."



By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

32.9	\$43
24.0	\$30
etc.	etc.

The regression equation describing the relationship between “Temperature” and “Revenue” is:

$$\text{Revenue} = 2.7 * \text{Temperature} - 35$$

Let's say one day at the lemonade stand it was 30.7 degrees and “Revenue” was \$50. That 50 is our observed or actual output, the value that actually happened.

So if we insert 30.7 at our value for “Temperature”...

$$\text{Revenue} = 2.7 * 30.7 - 35$$

$$\text{Revenue} = 48$$

...we get \$48. That's the predicted value for that day, also known as the value for “Revenue” the regression equation would have predicted based on the “Temperature.”

Your model isn't always perfectly right, of course. In this case, the prediction is off by 2; that difference, the 2, is called the residual. The residual is the bit that's left when you subtract the predicted value from the observed value.

$$\text{Residual} = \text{Observed} - \text{Predicted}$$


You can imagine that every row of data now has, in addition, a predicted value and a residual.

Temperature (Celsius)	Revenue (Observed)	Revenue (Predicted)	Residual (Observed – Predicted)
28.2	\$44	\$41	\$3
21.4	\$23	\$23	\$0
32.9	\$43	\$54	-\$11
24.0	\$30	\$29	\$1
etc.	etc.	etc.	etc.

We're going to use the observed, predicted, and residual values to assess and improve the model.

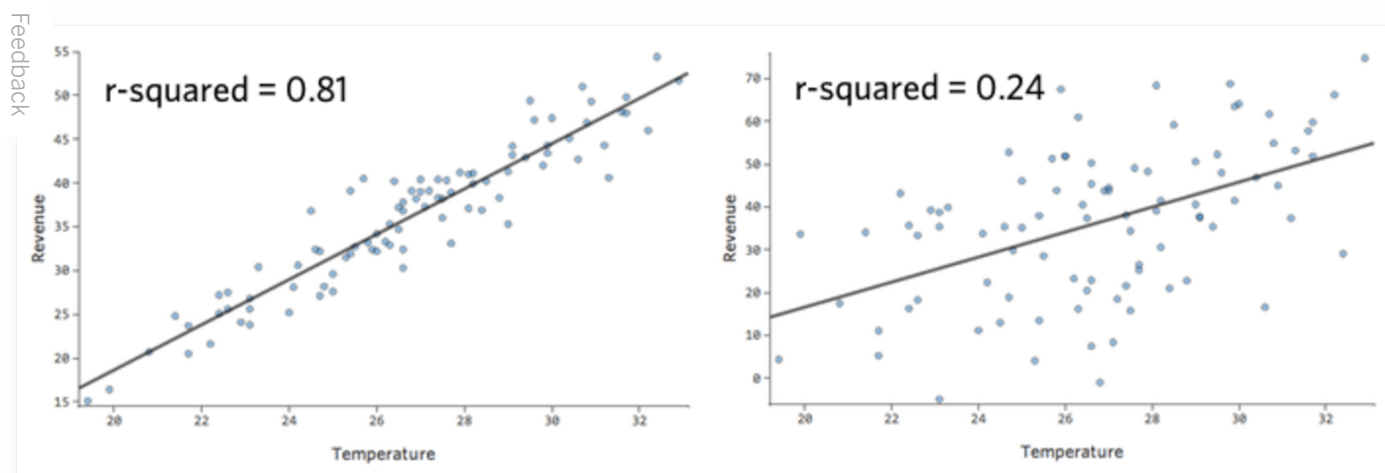
By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES

 Need Help?

Understanding Accuracy with Observed vs. Predicted

In a simple model like this, with only two variables, you can get a sense of how accurate the model is just by relating “Temperature” to “Revenue.” Here’s the same regression run on two different lemonade stands, one where the model is very accurate, one where the model is not:



It’s clear that for both lemonade stands, a higher “Temperature” is associated with higher “Revenue.” But at a given “Temperature,” you could forecast the “Revenue” of the left lemonade stand much more accurately than the right lemonade stand, which means the model is much more accurate.

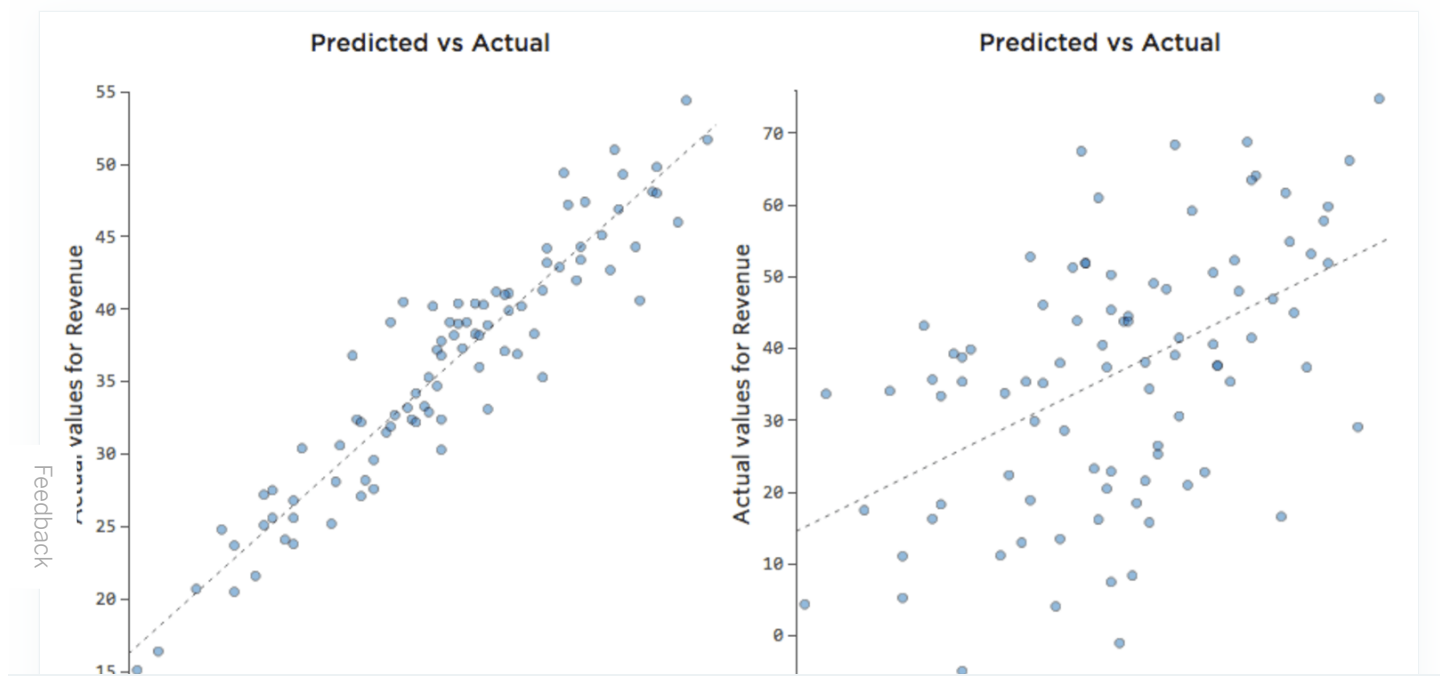
But most models have more than one explanatory variable, and it’s not practical to represent more variables in a chart like that. So instead, let’s plot the predicted values versus the observed values for

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES

 Need Help?

these same data sets.



Again, the model for the chart on the left is very accurate; there's a strong correlation between the model's predictions and its actual results. The model for the chart on the far right is the opposite; the model's predictions aren't very good at all.

Note that these charts look just like the "Temperature" vs. "Revenue" charts above them, but the x-axis is predicted "Revenue" instead of "Temperature." That's common when your regression equation only has one explanatory variable. More often, though, you'll have multiple explanatory variables, and these charts will look quite different from a plot of any one explanatory variable vs. "Revenue."

Was this helpful?

Examining Predicted vs. Residual ("The Residual Plot")

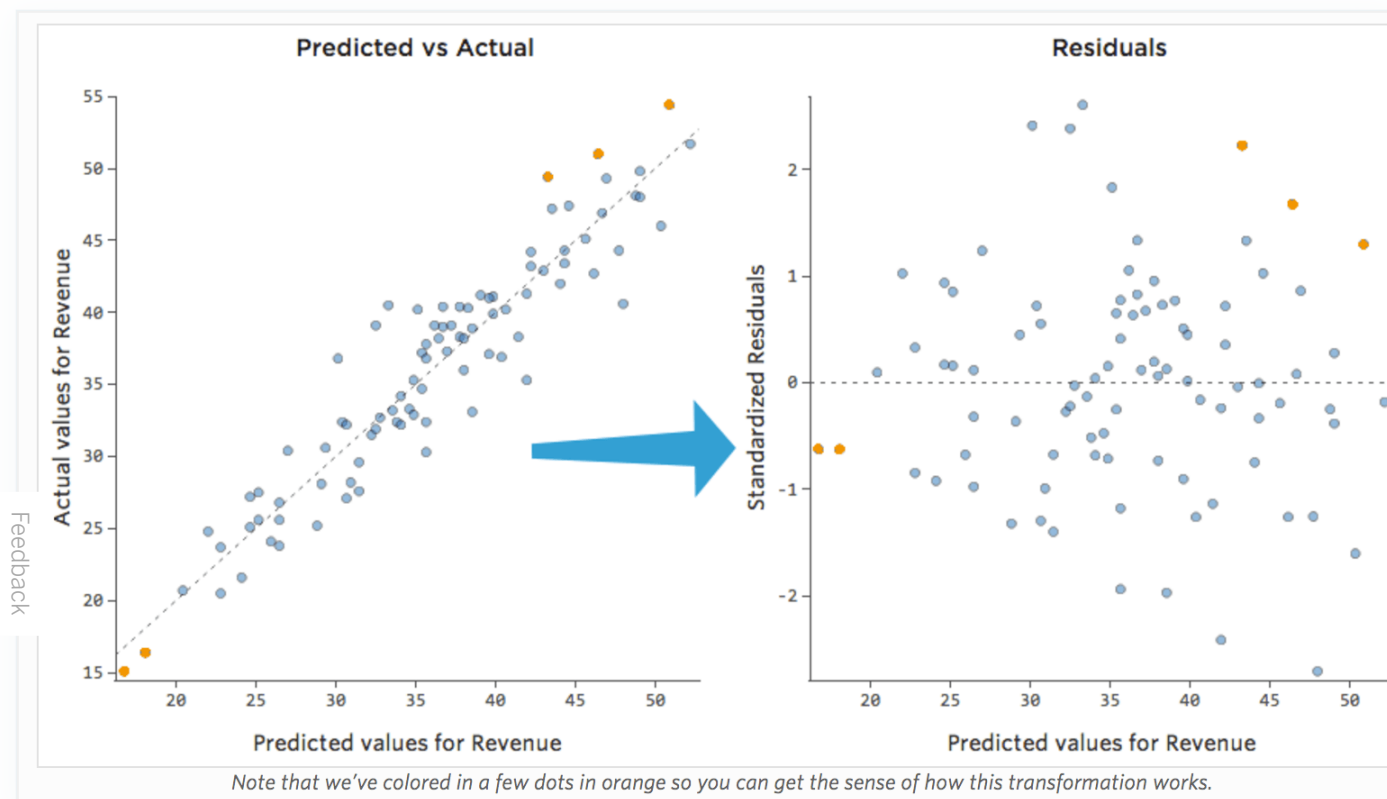
The most useful way to plot the residuals, though, is with your predicted values on the x-axis and your residuals on the y-axis.

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?



In the plot on the right, each point is one day, where the prediction made by the model is on the x-axis and the accuracy of the prediction is on the y-axis. The distance from the line at 0 is how bad the prediction was for that value.

Since...

Residual = Observed – Predicted

...positive values for the residual (on the y-axis) mean the prediction was too low, and negative values mean the prediction was too high; 0 means the guess was exactly correct.

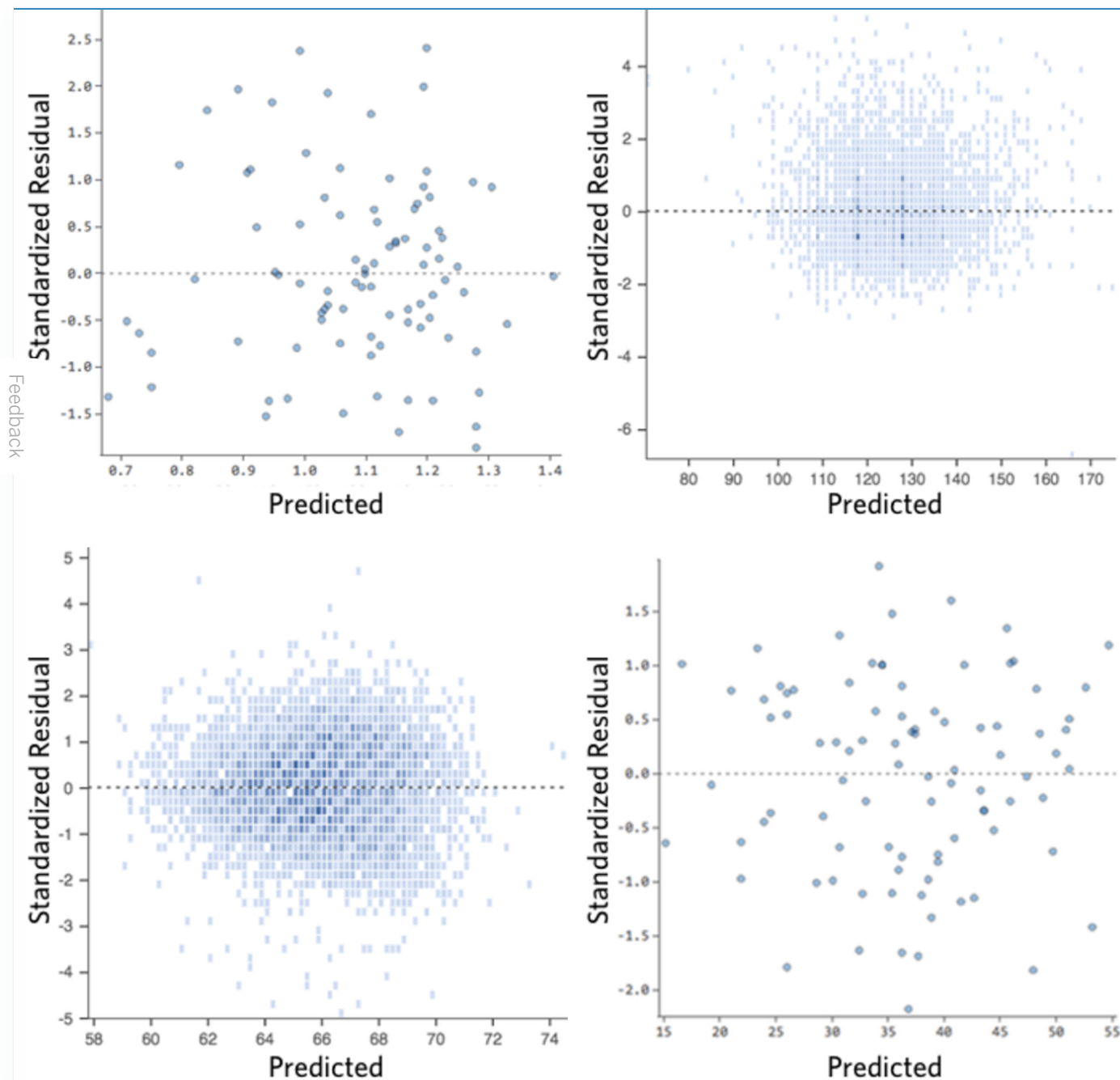
By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

Ideally your plot of the residuals looks like one of these:



That is,

- (1) they're pretty symmetrically distributed, tending to cluster towards the middle of the plot.
- (2) they're clustered around the lower single digits of the y-axis (e.g., 0.5 or 1.5, not 30 or 150).
- (3) in general, there aren't any clear patterns.

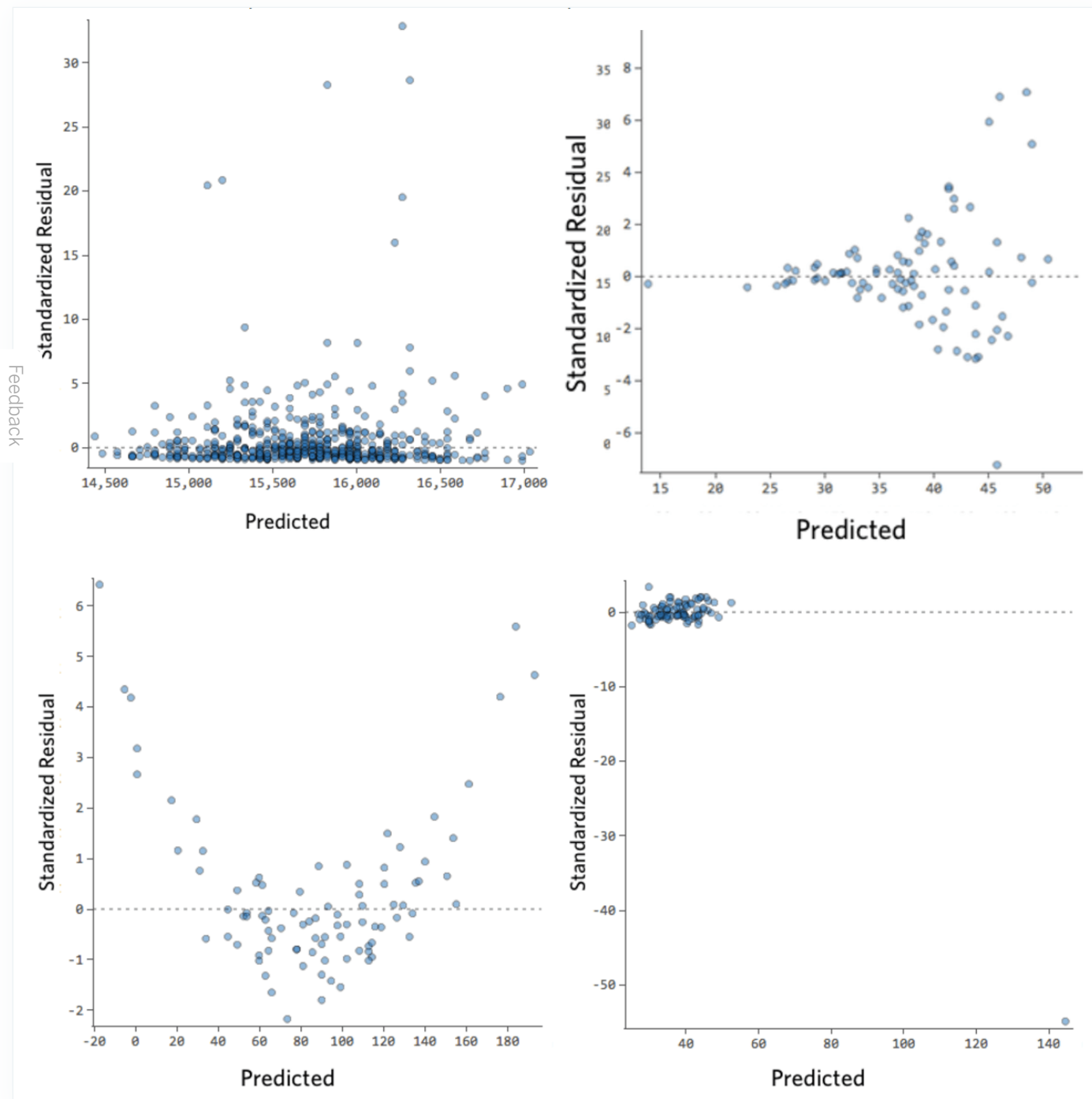
By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

Here's some residual plots that don't meet those requirements:




These plots aren't evenly distributed vertically, or they have an outlier, or they have a clear shape to them.

If you can detect a clear pattern or trend in your residuals, then your model has room for improvement.

In a second we'll break down why and what to do about it.

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES

 Need Help?

How much does it matter if my model isn't perfect?

How concerned should you be if your model isn't perfect, if your residuals look a bit unhealthy? It's up to you.

Feedback If you're publishing your thesis in particle physics, you probably want to make sure your model is as accurate as humanly possible. If you're trying to run a quick and dirty analysis of your nephew's lemonade stand, a less-than-perfect model might be good enough to answer whatever questions you have (e.g., whether "Temperature" appears to affect "Revenue").

Most of the time a decent model is better than none at all. So take your model, try to improve it, and then decide whether the accuracy is good enough to be useful for your purposes.

Was this helpful?

Example Residual Plots and Their Diagnoses

If you're not sure what a residual is, take five minutes to read the above, then come back here.

Below is a gallery of unhealthy residual plots. Your residual may look like one specific type from below, or some combination.

If yours looks like one of the below, click that residual to understand what's happening and learn how to fix it.

(Throughout we'll use a lemonade stand's "Revenue" vs. that day's "Temperature" as an example data set.)

Y-AXIS UNBALANCED

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

Feedback

Show details about this plot, and how to fix it.



HETEROSCEDASTICITY

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

Show details about this plot, and how to fix it.



NONLINEAR



Feedback

Show details about this plot, and how to fix it.



OUTLIERS

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

Feedback

Show details about this plot, and how to fix it.



LARGE Y-AXIS DATAPOINTS

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

X-AXIS UNBALANCED



Was this helpful?

Yes

No

Improving Your Model: Assessing the Impact of an Outlier

Let's assume that you have an outlying datapoint that is legitimate, not a measurement or data error. To decide how to move forward, you should assess the impact of the datapoint on the regression.

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES

 Need Help?

If that changes the model significantly, examine the model (particularly actual vs. predicted), and decide which one feels better to you. It's okay to ultimately discard the outlier as long as you can theoretically defend that, saying, "In this case we're not interested in outliers, they're just not of interest," or "That was the day Uncle Jerry came buy and tipped me \$100; that's not predictable, and it's not worth including in the model."

Was this helpful?

Yes

No

Feedback

Improving Your Model: Transforming Variables

OVERVIEW

The most common way to improve a model is to transform one or more variables, usually using a "log" transformation.

Transforming a variable changes the shape of its distribution. Typically the best place to start is a variable that has an asymmetrical distribution, as opposed to a more symmetrical or bell-shaped distribution. So find a variable like this to transform:

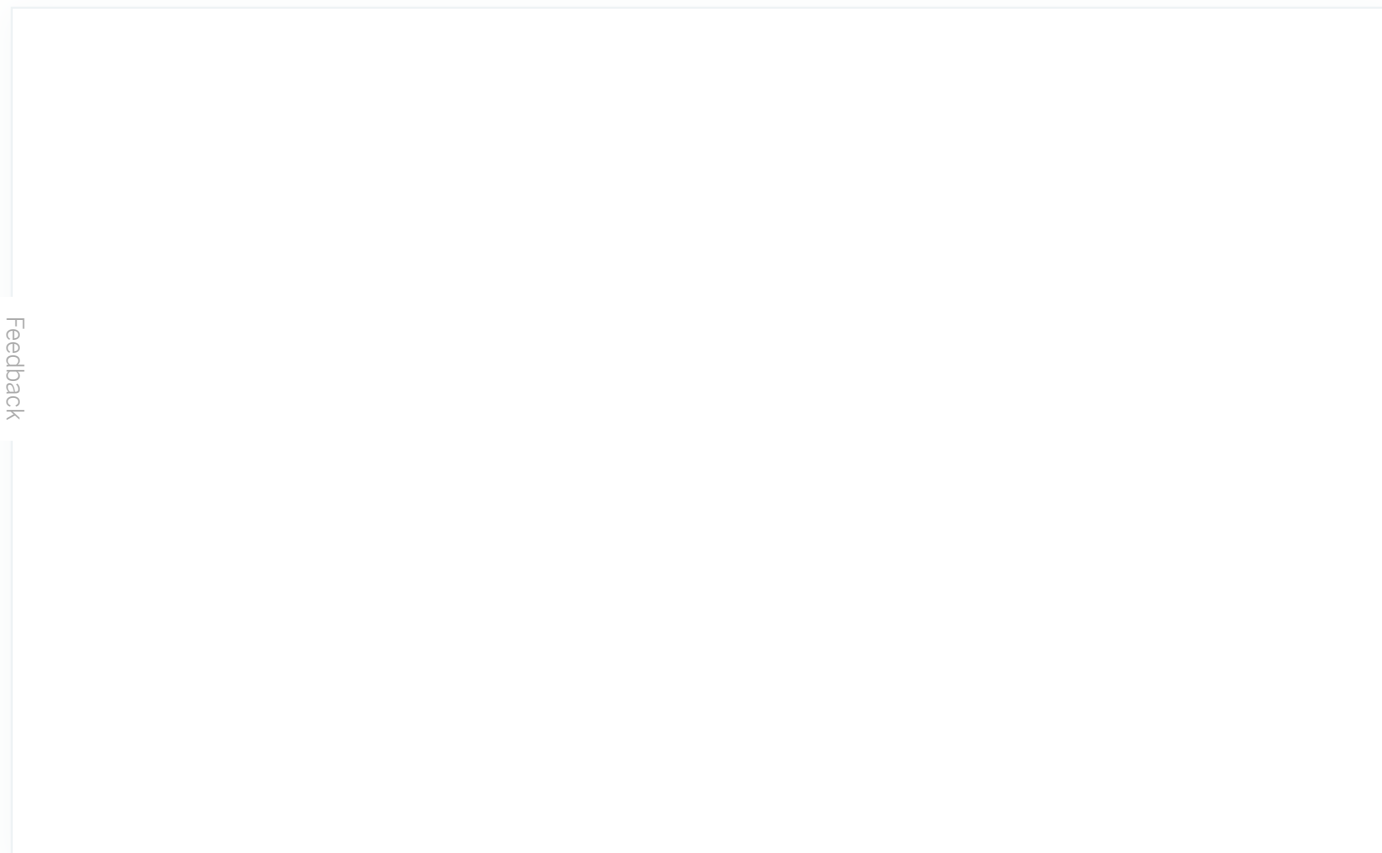
By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

In general, regression models work better with more symmetrical, bell-shaped curves. Try different kinds of transformations until you hit upon the one closest to that shape. It's often not possible to get close to that, but that's the goal. So let's say you take the square root of "Revenue" as an attempt to get to a more symmetrical shape, and your distribution looks like this:



That's good, but it's still a bit asymmetrical. Let's try taking the log of "Revenue" instead, which yields this shape:

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

Feedback

That's nice and symmetrical. You're probably going to get a better regression model with $\log(\text{"Revenue"})$ instead of "Revenue." Indeed, here's how your equation, your residuals, and your r -squared might change:

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

Stats iQ shows a small version of the variable's distribution inline with the regression equation:

Select the transformation f_x button to the left of the variable...

Feedback

...then select a transformation, most often $\log(x)$...

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

Feedback

...then examine the histogram to see if it's more centered, as this one is after transformation:

After transforming a variable, note how its distribution, the r-squared of the regression, and the patterns of the residual plot change. If those improve (particularly the r-squared and the residuals), it's probably best to keep the transformation.

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES

 Need Help?

to transform includes zeros or negative values, though. To learn why taking a log is so useful, or if you have non-positive numbers you want to transform, or if you just want to get a better understanding of what’s happening when you transform data, read on through the details below.

DETAILS

If you take the $\log_{10}()$ of a number, you’re saying “10 to what power gives me that number.” For example, here’s a simple table of four datapoints, including both “Revenue” and Log(“Revenue”):

Feedback

Temperature	Revenue	Log(Revenue)
20	100	2
30	1,000	3
40	10,000	4
45	31,623	4.5

Note that if we plot “Temperature” vs. “Revenue,” and “Temperature” vs. Log(“Revenue”), the latter model fits much better.



The interesting thing about this transformation is that your regression is no longer linear. When “Temperature” went from 20 to 30, “Revenue” went from 10 to 100, a 90-unit gap. Then when “Temperature” went from 30 to 40, “Revenue” went from 100 to 1000, a much larger gap.

If you’ve taken a log of your response variable, it’s no longer the case that a one-unit increase in

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES

 Need Help?

Also note that you can't take the log of 0 or of a negative number (there is no X where $10^X = 0$ or $10^X = -5$), so if you do a log transformation, you'll lose those datapoints from the regression. There's 4 common ways of handling the situation:

- 1 Take a square root, or a cube root. Those won't change the shape of the curve as dramatically as taking a log, but they allow zeros to remain in the regression.
- 2 If it's not too many rows of data that have a zero, and those rows aren't theoretically important, you can decide to go ahead with the log and lose a few rows from your regression.
- 3 Instead of taking $\log(y)$, take $\log(y+1)$, such that zeros become ones and can then be kept in the regression. This biases your model a bit and is somewhat frowned upon, but in practice, its negative side effects are typically pretty minor.

Feedback

Was this helpful?

Improving Your Model: Missing Variables

Probably the most common reason that a model fails to fit is that not all the right variables are included. This particular issue has a lot of possible solutions.

ADDING A NEW VARIABLE

Sometimes the fix is as easy as adding another variable to the model. For example, if lemonade stand "Revenue" traffic was much larger on weekends than weekdays, your predicted vs. actual plot might look like the below (r-squared of 0.053) since the model is just taking the average of weekend days

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

and weekdays:

Feedback



If the model includes a variable called “Weekend,” then the predicted vs. actual plot might look like this (r-squared of 0.974):

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

Feedback

The model makes far more accurate predictions because it's able to take into account whether a day of the week is a weekday or not.

Note that sometimes you'll need to create variables in Stats iQ to improve your model in this fashion. For example, you might have had a "Date" variable (with values like "10/26/2014") and you might need to create a new variable called "Day of Week" (i.e., Sunday) or Weekend (i.e., Weekend).

UNAVAILABLE OMITTED VARIABLE

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

something like “Number of Competitors in the Area” that you failed to collect at the time.

If the variable you need is unavailable, or you don’t even know what it would be, then your model can’t really be improved and you have to assess it and decide how happy you are with it (whether it’s useful or not, even though it’s flawed).

INTERACTIONS BETWEEN VARIABLES

Perhaps on weekends the lemonade stand is always selling at 100% of capacity, so regardless of the “Temperature,” “Revenue” is high. But on weekdays, the lemonade stand is much less busy, so “Temperature” is an important driver of “Revenue.” If you ran a regression that included “Weekend” and “Temperature,” you might see a predicted vs. actual plot like this, where the row along

Feedback

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

the top are the weekend days.

Feedback

We would say that there's an **interaction** between "Weekend" and "Temperature"; the effect of one of them on "Revenue" is different based on the value of the other. If we create an interaction variable, we

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

get a much better model, where predicted vs. actual looks like this:



Feedback

Was this helpful?

Improving Your Model: Fixing Nonlinearity

Let's say you have a relationship that looks like this:

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES

 Need Help?

Feedback

You might notice that the shape is that of a parabola, which you might recall is typically associated with formulas that look like this:

$$y = x^2 + x + 1$$

By default, regression uses a linear model that looks like this:

$$y = x + 1$$

In fact, the line in the plot above has this formula:

$$y = 1.7x + 51$$

But it's a terrible fit. So if we add an x^2 term, our model has a better chance of fitting the curve. In fact, it creates this:

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

Feedback

The formula for that curve is:

$$y = -2x^2 + 111x - 1408$$

That means our diagnostic plots change from this...

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

Feedback

Note that these are healthy diagnostic plots, even though the data appears to be unbalanced to the right side of it.

The above approach can be extended to other kinds of shapes, particularly an S-shaped curve, by adding an x^3 term. That's relatively uncommon, though.

A few cautions:

- Generally speaking, if you have an x^2 term because of a nonlinear pattern in your data, you want to have a plain-old- x -not- x^2 term. You may find that your model is perfectly good without it, but you should definitely try both to start.
- The regression equation may be difficult to understand. For the linear equation at the beginning of this section, for each additional unit of "Temperature," "Revenue" went up 1.7 units. When you have both x^2 and x in the equation, it's not easy to say "When *Temperature* goes up one degree, here's what happens." Sometimes for that reason it's easier to just use a linear equation, assuming that equation fits well enough.

Was this helpful?

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

How do I create a new Stats iQ analysis?	+
What are the options for analyzing my data in Stats iQ?	+
I don't know what this statistical term means. Can you tell me?	+
How do I filter the data that appears in Stats iQ?	+
How do I get my new responses to show up in Stats iQ?	+
How are analysis cards ordered in my Stats iQ Workspace?	+
What's Stats iQ? / Where's Statwing?	+

USER-FRIENDLY GUIDE TO LOGISTIC REGRESSION

THE CONFUSION MATRIX & PRECISION-RECALL TRADEOFF

Related Articles

USER-FRIENDLY GUIDE TO LINEA...

Regression estimates a mathematical formula that relates one or more input variables to one output variable.

USER-FRIENDLY GUIDE TO LOGIS...

Logistic regression estimates a mathematical formula that relates one or more input variables to one output v

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES

 Need Help?

together impact an output variable. For example, if both the

POPULAR USE CASES

Customer Experience Management (CXM)
NPS Software
Employee Engagement Software
Online Survey Software
Market Research Software
360 Development
Customer Survey Software
Website & App Feedback
Voice of Customer Software
Employee Pulse Surveys
Onboarding & New Hire Surveys
Reputation Management Software

SUPPORT

Submit a Ticket
Online Help
Qualtrics Community
Professional Services
Product Roadmap
Status

COMPANY

About Us
Investors
X4 Summit
Careers
Partnerships
Contact Us

By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?

CUSTOMERS

Integrations

Blog

Events

Training & Certification

Resource Library

XM Basecamp

© 2021 Qualtrics®

Terms of Service

Privacy Statement

Security Statement

Cookie Preferences



By using this site, you agree with our use of cookies. [Want to know more?](#)

I CONSENT TO COOKIES



Need Help?