Residuals

This document will introduce you to the concept of residuals in statistics. You'll learn how to calculate, plot, and interpret residuals to enhance your analysis of data.

Furry Friends Animal Sanctuary (FFAS) needs our help! In just a few weeks, ten giant pandas will call the preserve home. To prepare for their arrival, FFAS needs to stock up on bamboo. But how much do they need? Let's use our data science knowledge to help them estimate how much bamboo the pandas will eat each week.

A panda's weight might give us a pretty good idea of how much food it needs. Using data about the weight and bamboo consumption of other pandas, we can create a linear regression model to fit this data. Here, the independent variable will be a panda's weight, and the dependent variable will be how much bamboo it needs to consume.

Once we have our model, though, how do we know whether or not we can trust the predictions it makes? We want to make sure that we advise FFAS wisely, for the sake of their pandas' happiness and well-being. One way to evaluate our model is through calculating and plotting **residuals**.

We can think of residuals as a sort of estimate of error for our model. A residual is the difference between a value observed for the dependent variable and the corresponding value predicted by the model. In other words,

residual = value observed - value predicted

In the FFAS case, our residuals will compare the amount of bamboo our model predicts that a panda of a certain weight will eat to the amount that a panda of that weight actually ate. In this way, we can use residuals to figure out whether or not the model we're using is appropriate for the data at hand. This is a good complement to finding the R^2 value for the model.

> **Question:** Perhaps we created a linear regression model for our panda data and found the following for one particular panda:

Panda's Weight (kg)	Predicted Weekly Bamboo Consumption (kg)	Actual Weekly Bamboo Consumption (kg)
100	151	148

What would the residual for this data point be?

Answer: To find the residual, we subtract the value predicted for the weekly bamboo consumption from the value observed:

151 - 148 = 3

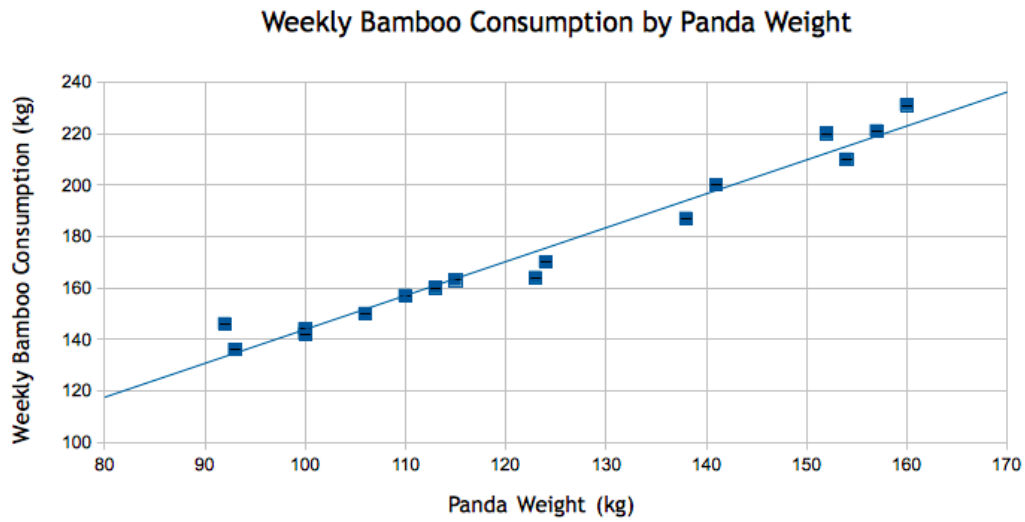
The residual for this data point is thus 3 kg.

This is all well and good, but the number 3 doesn't mean a whole lot on its own. In order to get an overall idea of how well our model is predicting bamboo consumption, we should find out some other residual values for this model.

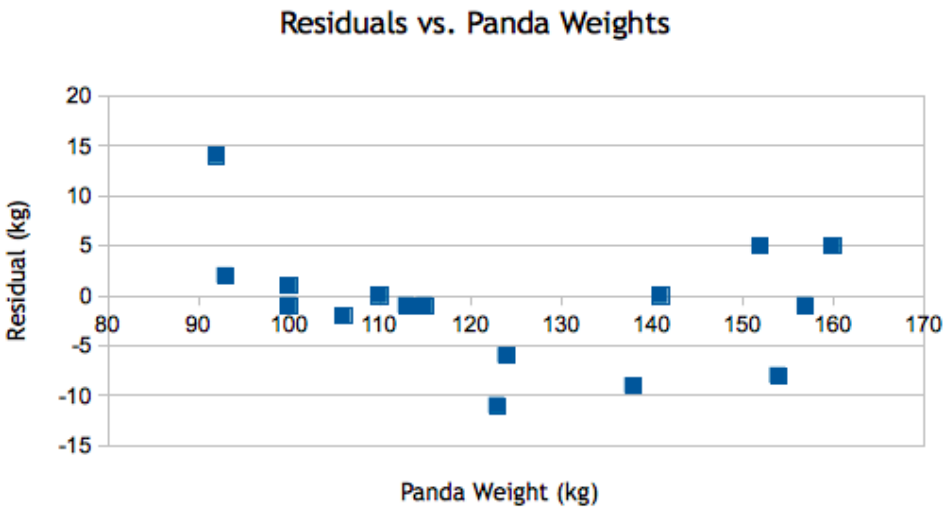
Perhaps after creating a model with data gathered from more pandas, we end up with the information below. As before, the residual is just the actual weekly bamboo consumption minus the predicted weekly bamboo consumption.

Panda's Weight (kg)	Actual Weekly Bamboo Consumption (kg)	Predicted Weekly Bamboo Consumption (kg)	Residual (kg)
100	137	143	-6
160	231	226	5
115	163	164	-1
124	170	176	-6
157	221	222	-1
113	149	161	-12
138	187	196	-9
145	228	205	23
106	150	152	-2
92	146	132	14
100	144	143	1
152	220	215	5
110	157	157	0
123	164	175	-11
141	200	200	0
154	210	218	-8
93	144	134	10

It's hard to see any trends when the data is formatted like this, so let's give ourselves a visual to help. On the graph below, the points in the scatter plot represent the actual data collected, and the line represents the model we're using to predict future bamboo consumption based on weight.



It looks like our data is relatively linear, but let's examine this further using the residuals. If we plot the residuals vs. the independent variable (panda weight), we end up with a graph like this:



Question: The plot of residuals vs. panda weights seems to have a relatively random pattern, with positive and negative residuals mixed around together. What do you think this says about how our model fits the data?

Answer: This randomness means that our linear model is a pretty good fit! If the residuals seemed to depend on panda weight in some way, that would indicate that the relationship between weight and bamboo consumption was not approximately linear. We can now tell FFAS with confidence that this linear model will help them figure out how much bamboo to buy.

Now that you've had an introduction to residuals and helped out FFAS, pause and think for a moment about what a histogram of residuals from a well-fitted model would look like. What information would this give us that we couldn't get from an R^2 value?

Although an R^2 value can help us evaluate the overall fit of a model, it can't tell us about any non-random, systematic errors. Seeing a distribution of the residuals can reveal any clustering of errors, cluing us in about problems with the model's fit. If a histogram of the residuals reveals a normal distribution, however, this further tells us that our model is working well. The more residuals we have close to 0, the smaller the overall difference between our model and the actual data.