

Fitting Models By Maximum Likelihood and Least Squares

Class 3 in MATH 615: Approaches to Quantitative Analysis in the Life Sciences

Fitting models — simple linear regression by maximum likelihood — a Poisson regression model by maximum likelihood — simple linear regression by algebraic ordinary least squares — checking residuals — variance explained — populations and samples — biased and unbiased estimators

Setup

Simple linear regression

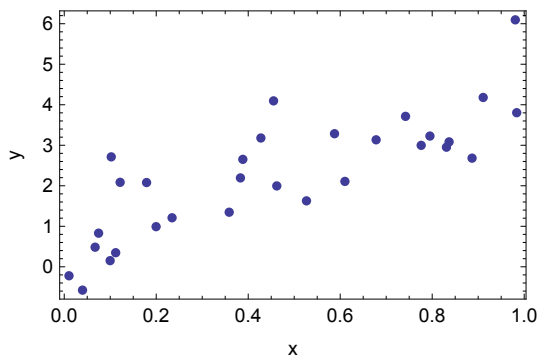
- Some data

x

```
{0.0746841, 0.742037, 0.983763, 0.836746, 0.830982, 0.0397565, 0.111839, 0.610239, 0.427565, 0.121527, 0.910857, 0.795068, 0.102191, 0.0997766, 0.980675, 0.678066, 0.0671251, 0.587646, 0.388313, 0.77606, 0.526689, 0.0103662, 0.199705, 0.358644, 0.462104, 0.455095, 0.234453, 0.179004, 0.383091, 0.886679}
```

y

```
{0.829985, 3.71352, 3.80383, 3.08382, 2.95229, -0.576366, 0.347498, 2.10775, 3.1794, 2.08534, 4.17968, 3.22734, 2.71354, 0.151644, 6.09583, 3.13462, 0.483795, 3.28492, 2.65234, 2.99816, 1.62799, -0.221725, 0.990624, 1.34724, 1.99625, 4.09572, 1.20976, 2.07922, 2.19408, 2.68162}
```



■ Maximum likelihood the hard way

Let's go back to the first model we looked at; a simple linear regression model with one response variable (y) and one predictor variable (x).

$$y_i = c + m x_i + \epsilon_i$$

We'll assume in this case that ϵ_i is distributed as $N(0, \sigma)$. We saw before that finding the mean of a normal distribution means that we find the value of μ that *maximizes the likelihood* of getting the observed data values. One could do this by calculating the likelihood of the data for a given value of μ (and σ), and then trying different values until we find the maximum likelihood. We also saw that one shortcut was simply to find the value that minimizes the sum of squared deviations around μ . We also saw that an even better shortcut was to simply find the mean of the data. All three methods give the same estimate for μ .

The method of maximizing likelihood is the most general, so let's try that first. Now that we have a regression model, we are not finding the MLE of a single parameter μ , but of two parameters, c and m . These two specify a line, and for any given value of x , the value of the line is the mean of the Normal distribution for data values at that point. To perhaps make this clearer, we can write the model for a simple Normal distribution as $y_i = N(\mu, \sigma)$ or, equivalently like this:

$$y_i = \mu + N(0, \sigma)$$

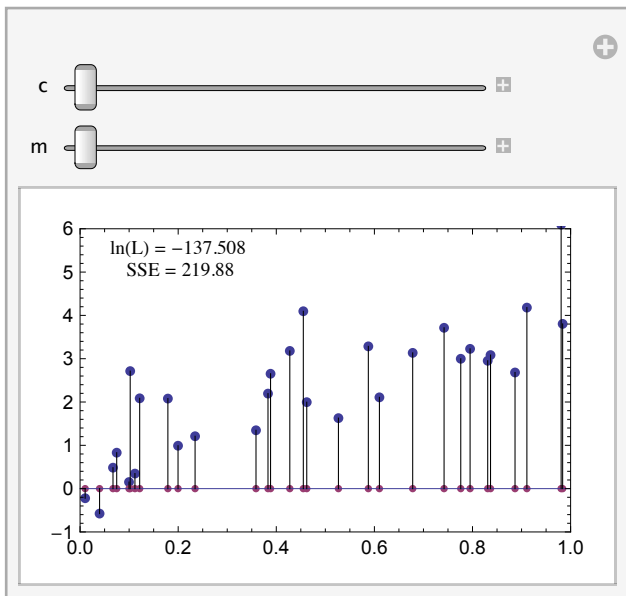
and the model for a regression like this:

$$y_i = c + m x_i + N(0, \sigma)$$

or this:

$$y_i = N(c + m x_i, \sigma)$$

So to calculate the likelihood of the data for a pair of values of c and m , we first calculate the value of \hat{y} ("y-hat"), which is $c + m x_i$, for all x_i , then get the value of the Normal PDF with that value as the mean and some specified value as the standard deviation σ . In the interactive plot below, the purple dots are the \hat{y} values.



Let's let the computer find the maximum likelihood

$$\{-41.5339, \{c \rightarrow 1.1165, m \rightarrow 2.76991\}\}$$

or, equivalently, the minimum sum of squared errors

$\{27.9316, \{c \rightarrow 1.1165, m \rightarrow 2.76991\}\}$

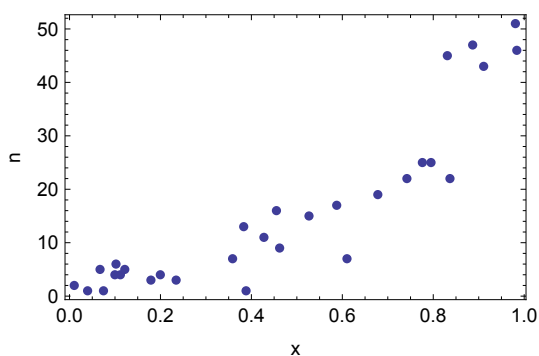
This method maximizing likelihood is, in fact, very general, and we will next use it again with a different kind of regression model. But in the particular case of normally-distributed errors, the method is known as **ordinary least squares**, and we will return to it to demonstrate how the best fit model can be obtained as if by magic!

Poisson regression

■ Some data

Suppose our response data are actually counts.

$\{1, 22, 46, 22, 45, 1, 4, 7, 11, 5, 43, 25, 6, 4, 51, 19, 5, 17, 1, 25, 15, 2, 4, 7, 9, 16, 3, 3, 13, 47\}$

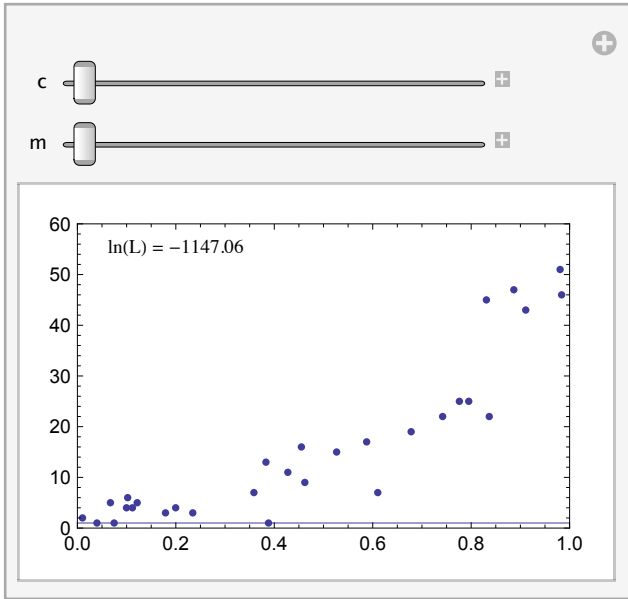


In this case, we know that the data are non-negative integers, and so the Normal distribution is not appropriate — the Poisson is a better candidate. Also the shape of the function cannot, strictly speaking, be a straight line, because any straight line (except one with zero slope) will at some point cross the x -axis and become negative. If that occurred a long way outside our data, we might be ok, but looking at the plot above, we see that our data come close to the x -axis, and so any fitted straight line would be negative in our range of interest. A better function for count data is an exponential function.

$$n_i = \text{Poisson}(e^{c+m x_i})$$

■ Maximum likelihood again

Even though our function and distribution model are different, we can apply an identical, maximum likelihood approach to that described above.



Let's let the computer find the maximum likelihood

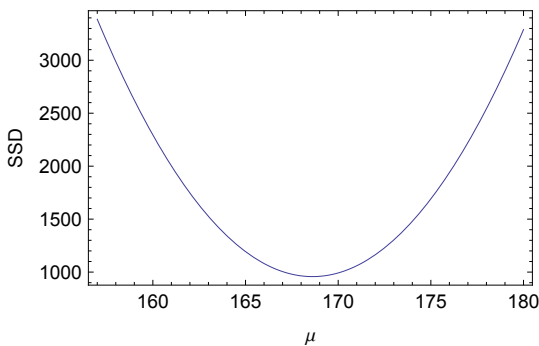
$$\{-73.5282, \{c \rightarrow 0.773595, m \rightarrow 3.31998\}\}$$

(There is no sum-of-squares equivalent here.)

Ordinary least squares by magic (actually, linear algebra)

Remember that we didn't need to actually calculate likelihoods in order to fit a Normal distribution to a set of data — we just needed to calculate the mean, because that *was* the MLE of μ . A similar shortcut applies to fitting a linear regression via ordinary least squares (i.e., assuming Normally-distributed errors). Remember the sum-of-squares plot for our class height data?

$$\{163.84, 179.2, 158.72, 161.28, 168.96, 166.4, 184.32, 168.96, 161.28, 163.84, 166.4, 168.96, 163.84, 176.64, 180.48, 171, 159, 172\}$$



Because the sum of squares involves, well, squares, the function is *quadratic*, meaning that it has a single, global minimum. This can therefore be identified by finding the derivative (slope) of the function, setting it to zero and solving for μ . To show the answer, however, we have to introduce some new notation that we will see repeatedly.

For the simple linear regression, we have been writing it like this:

$$y_i = c + m x_i + \epsilon_i$$

Statisticians, however, like to gather up all their parameters into a vector (or list) and call it β (or sometimes θ). In our case, there are two parameters, c and m , and so β is the vector $\{c, m\}$. But the question then arises: how to we write an equation with β , when only the m term applies to the x_i , and the c term is a constant that stands alone? The answer is to redefine x as a matrix X , in which the first column is a list of 1's and the second column is the x_i . The column of 1's is a placeholder for the constant to apply to (multiplying something by 1 doesn't change its value) The end result is that we can express our model as

$$y = X \beta + \epsilon$$

Don't worry if you don't see how the matrix version works — let's just demonstrate that it does. Here is a small example of an X matrix

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix}$$

and the β vector

$$\begin{pmatrix} c \\ m \end{pmatrix}$$

If we *matrix multiply* X by β , we get

$$\begin{pmatrix} c + m x_1 \\ c + m x_2 \\ c + m x_3 \end{pmatrix}$$

which is what the values of y_i should be.

Ok, so that's the notation. In that notation, the expected values of the y_i are given by $X \beta$, so the deviations are $y - X \beta$, so the sum of squared deviations is $(y - X \beta)'(y - X \beta)$. If you differentiate this with respect to β , set the result to zero and solve for β , you get

$$\hat{\beta} = (X' X)^{-1} X' y$$

You are essentially solving for the location of minimum of the sum-of-squares function, which is the location where the slope of the function is zero in all dimensions. Thus the least-squares estimated values of c and m ($\hat{\beta}$, pronounced " β -hat") can be obtained from a single matrix calculation, avoiding the need for an iterative search. Let's see if it works! Here is the computer search method with the data from the first section above:

$$\{27.9316, \{c \rightarrow 1.1165, m \rightarrow 2.76991\}\}$$

And here is our new method. First we create the X matrix:

```

( 1  0.139617
  1  0.298572
  1  0.405814
  1  0.593891
  1  0.694011
  1  0.857173
  1  0.191217
  1  0.489539
  1  0.885339
  1  0.241131
  1  0.791622
  1  0.0166526
  1  0.543152
  1  0.358873
  1  0.332927
  1  0.126654
  1  0.63567
  1  0.436975
  1  0.300465
  1  0.0440153
  1  0.71205
  1  0.246368
  1  0.646902
  1  0.953103
  1  0.96134
  1  0.65965
  1  0.622353
  1  0.296653
  1  0.167212
  1  0.268925 )

```

And then the calculation:

```

βHat = Inverse [Transpose [xMatrix] . xMatrix] . Transpose [xMatrix] . y
{1.1165, 2.76991}

```

The exact same answer!

This fitting method may not seem *that* magical, because our computer was able to calculate the estimates by numerical optimization very quickly indeed. But that's partly because the model was so simple. With multi-variable models, fitting by trial-and-error optimization quickly becomes slow and unreliable, even with fast computers, whereas the algebraic solution is always fast and exact.

Checking residuals

Our model fitting process assumed Normally-distributed errors, so we should check. Once we have the fitted parameters, we can get the expected y_i and thus the deviations of the data from that: the *residuals*

```
yHat = xMatrix.βHat
```

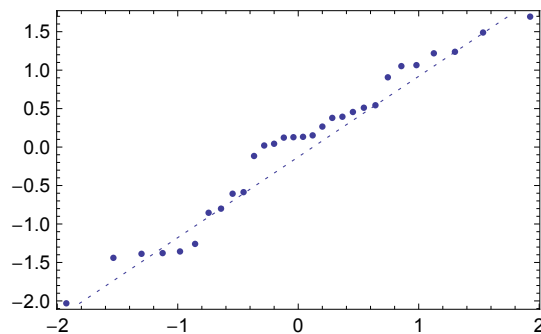
```
{1.50322, 1.94351, 2.24056, 2.76152, 3.03884, 3.49079, 1.64615,
 2.47248, 3.56881, 1.78441, 3.30922, 1.16262, 2.62098, 2.11054,
 2.03868, 1.46732, 2.87725, 2.32688, 1.94876, 1.23841, 3.08881, 1.79891,
 2.90836, 3.75651, 3.77932, 2.94367, 2.84036, 1.9382, 1.57966, 1.86139}
```

```
residuals = y - yHat
```

```
{0.456024, -1.44003, 0.266246, 0.0430066, 0.0204247, -0.115948, -1.25888,
 -0.586128, -0.606466, -1.3876, -2.03112, 0.152179, 0.12815, 0.12278, 1.065,
 0.133016, -1.35713, 0.543602, -1.37851, 0.394242, -0.85406, -0.800517,
 1.23867, 0.379304, 1.21885, 1.4889, 0.906874, 0.511311, 1.05236, 1.69545}
```

A Q-Q plot will give us a good visualization

```
QuantilePlot[residuals]
```



■ Residuals are only interesting for Normally-distributed data

The idea that residuals — the difference between an observed data point and the expected value predicted by the model — have a distribution makes sense if the distribution is Normal, because of the following equivalency

$$y_i = N(\mu, \sigma) = \mu + N(0, \sigma)$$

In the third version, the mean has been extracted from the Normal description, which now describes just the residuals. But this isn't possible for, say, the Poisson distribution. If data points are Poisson-distributed, with parameter (and therefore mean) λ , the deviations from λ cannot be Poisson. For one thing, some of them would be negative. Also, while the data are integers, λ need not be, and so the deviations are not necessarily integers either.

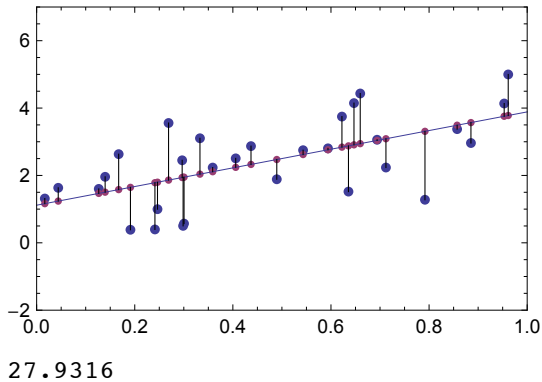
The bottom line is that concept of the examining the distribution of deviations only applies if the distribution you are assuming is Normal.

Sums of squares and the variance explained by a model

We have seen that fitting a model with Normal errors is accomplished by minimizing the sum of squared deviations from the model, also called the residual sum of squares (RSS), or sometimes the sum of squared errors (SSE). Here, we will explore some other sums of squares. These concepts will be used next week when we consider hypothesis testing.

Residual sum of squares (RSS)

We already know that the residual sum of squares is the sum of the squared deviations from the fitted model (the lengths of vertical lines in the figure below).

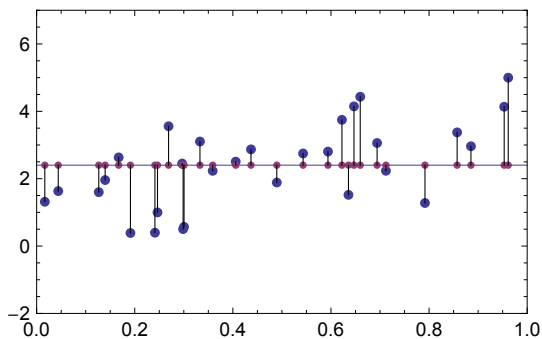


Total sum of squares (TSS)

If there was no relationship between x and y in a dataset, then the simplest possible description of y is as a distribution on its own. Again, assuming a Normal distribution, the y could be described by a single mean and standard deviation.

$$y_i = \mu + N(0, \sigma)$$

Taking the regression data from the sections above, we can visualize this by still spreading the data out along the x axis, but having a single predicted y , the mean.



This is equivalent to writing the regression version of the model with m (the slope) set to zero

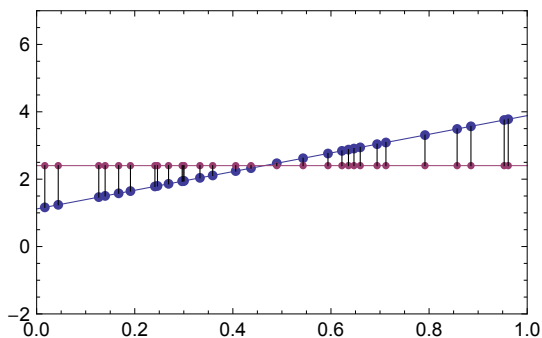
$$y_i = c + 0 x_i + N(0, \sigma) = c + N(0, \sigma)$$

in which case c (the intercept) is just the mean of the y_i . For this simpler model, we can also calculate the sum of squares, and because it is the simplest possible model for the y_i , and therefore simply a measure of total variation in the y_i , we will call it the total sum of squares (TSS).

44.7963

Model sum of squares

Finally, we can calculate the variability in the y_i predicted by the regression model (ignoring the error), compared to the simpler model. This is the regression, or model sum of squares (MSS).



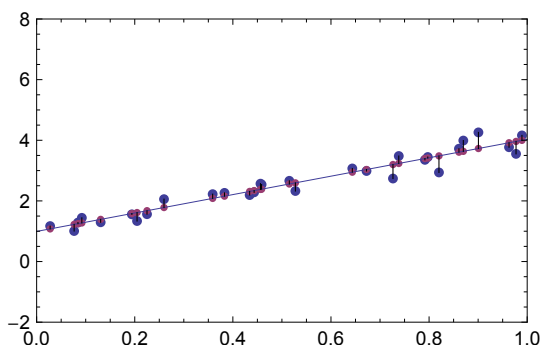
16.8647

■ Variance explained by the model

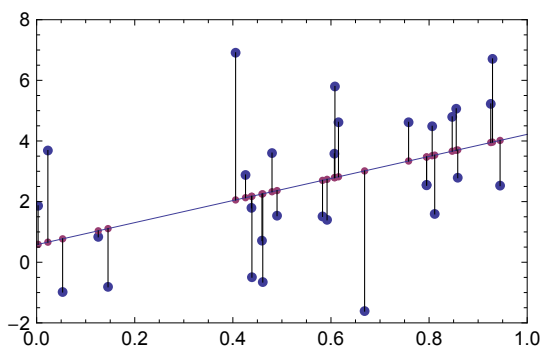
If you look at the sums of squares above, you should notice that $TSS = MSS + RSS$. In other words, the fitted model *partitions* the overall variability in the response variable, y_i , into two components: that which is explained by the model, and that which remains as residual variability. One way of classifying a model as a ‘good fit’ is if the model explains a large fraction of the variability of the response (meaning that residual variability is small). This fraction, $\frac{MSS}{TSS}$ is called the r^2 (“*r*-squared”, or coefficient of determination) of the model. Many software packages don’t calculate the MSS, but r^2 can also be calculated as $1 - \frac{RSS}{TSS}$.

0.376475

■ A data/model combination with a high r^2 value



■ A data/model combination with a low r^2 value



Again, make note that these concepts only apply to models with Normally-distributed errors!

Populations, samples and inference

We need to revisit the question of what a fitted model represents. It is a simplified description of two underlying processes — a deterministic one and a probabilistic one — that combine to produce some observed aspect of the real world (i.e., some data). But why go to all the trouble?

Consider our data on class heights. We saw that they were not particularly Normally distributed, and attributed that in discussion to gender differences in height. It might make more sense to fit models to the male and female students separately. We could then ask if the means are different. But how does our model help us do that? We could simply calculate the means the simple way, and compare them. Why should we care that the means are also the MLEs of the parameter μ of a fitted Normal distribution?

If all we are interested in doing is comparing the average heights of students in the class, then there is *no need for a statistical approach*. The means are all the information we need. The distribution of the values is irrelevant to the question "are the males in the class taller than the females?"

This, however, is a rather unusual situation, because we were easily able to collect data from every individual in the **population** we were interested in (the class). Most interesting questions are more general, and therefore apply to much larger populations. For example, we might ask if students at technical schools like NJIT are, on average, taller or shorter than students at liberal arts colleges. In those cases, it is not feasible to collect heights from everyone in each of the two populations. Instead we collect a **sample** of heights from a subset of individuals.

The distinction between populations and samples is crucial to understanding statistics, which might be described as making **inferences** about populations based on samples. Let's return to our class data. Suppose we wanted to know the mean height of students at NJIT. If the students in this class are a representative sample, then the mean height of the students in class tells us *something* about the height of students in general, but doesn't give the exact answer (which would only be obtainable by collecting heights from everybody). The class mean, or sample mean, is the only information we have. We say that the **sample mean is an estimate of the population, or "true" mean**.

The obvious follow-up question is: how good an estimate is it? You should be able to see that this has to do, in part with the size of the sample. If we measured all but one student in the whole university, then our sample mean would be very close indeed to the population mean. If we measured just a couple of students, chosen at random, then we might easily get a pair who are much taller or shorter than the mean, and we should have correspondingly less confidence in our estimate. It also has to do with the variability in the data, and whether the sample is, indeed representative (for example, we wouldn't want to sample just the basketball team to get an idea of average heights across all students).

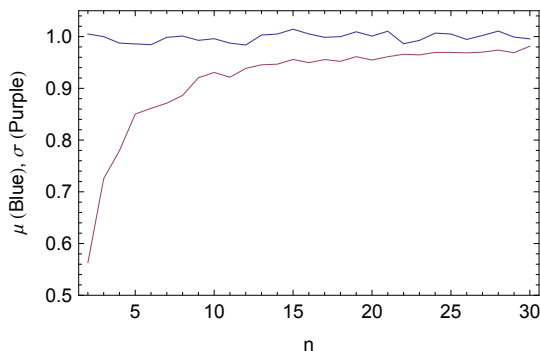
Making an assumption about the underlying distribution of heights allows us to assess how close our estimate is likely to be to the true value.

We won't investigate this until the next class, but this ability to assess estimates is crucial to asking statistical questions such as whether two samples have different heights. If our data come from small *samples* with a lot of variability, then our uncertainty about the mean of each *population* will be large, and it will be difficult to say with any confidence whether one population mean is larger than the other.

Biased and unbiased estimators

The distinction between samples and populations is also important in this section on fitting, because of something called bias. Returning to the Normal distribution, we can do a computer experiment about sampling. We will generate random numbers from a Normal distribution with known μ and σ . The random numbers represent a sample. We will take samples of different sizes, and calculate the maximum likelihood estimators of the mean and standard deviation of the two parameters, $\hat{\mu}$ and $\hat{\sigma}$, for each sample.

Here are the results for a distribution with $\mu = 1$ and $\sigma = 1$, doing this 1000 times and taking the average $\hat{\mu}$ and $\hat{\sigma}$ for each sample size n .



While the average estimate of μ , the mean, is consistently around the true value of 1, the average estimate of σ is consistently *below* 1 until the sample size gets quite large. This tells us that the standard deviation of the data is a *biased* estimator of σ , the standard deviation of the underlying distribution, meaning that its error is consistently positive or negative (in this case, negative). The mean, on the other hand, is an *unbiased* estimator of μ .

Can you see why estimates of σ are biased? It is sensitive to the tails of the distribution, and values in the tails of the Normal distribution are, by definition, rare (that's why they are called tails), which means that they tend not to occur in small samples. So small samples rarely have *any* large deviations from the mean, and so the standard deviation of small samples underestimates the true standard deviation.

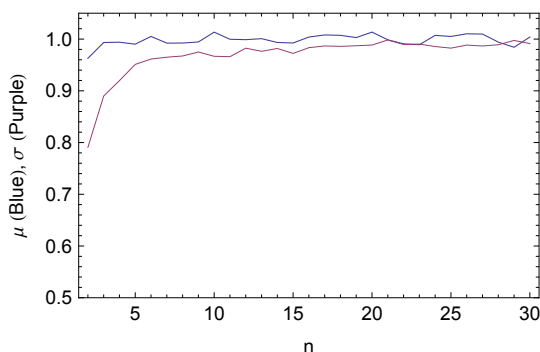
How can we deal with this? It turns out that there is an easy method. The normal formula for a standard deviation is

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

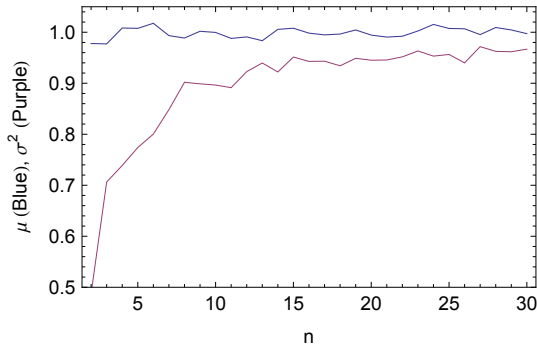
If we make a correction by dividing by $n - 1$ instead of n :

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

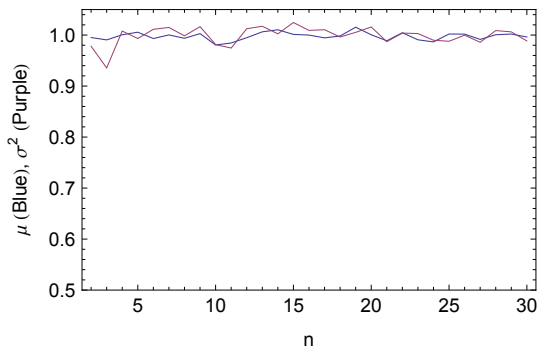
we get an improved estimator of the standard deviation:



It's not perfect, but actually estimates of the variance, which is simply σ^2 , are truly unbiased with this correction. Uncorrected:



and corrected:



So, some estimators are unbiased, but many are biased, especially if they have something to do with the ‘spread’ of data. Many of the biased estimators have ‘bias-corrected’ variants.

Note also that we have shown that a model fit by maximum likelihood may not necessarily be the ‘best’ fit. It depends on your goals. In this case, it depends in whether you want to make inferences simply about your data, or about a larger population of which your data are a sample.

■ Dubious language

The uncorrected standard deviation of the sample is often called, unsurprisingly, the "standard deviation of the sample."

The corrected standard deviation of the sample is often called the "sample standard deviation."

Statisticians make some great choices...

■ Take-home message

When you calculate something like a standard deviation in a software package, make sure you know which version it is using, and whether it is right for your question. For example, if I simply wanted to know the standard deviation of the heights of students in my class, and I had *all* the heights (i.e., the whole population), then I wouldn't use the bias-corrected version. If I was using the same data to estimate the standard deviation of the heights of students in NJIT as a whole, (i.e., I'm treating my data as a sample of a larger population), then I *would* use bias-correction.

Maximum likelihood fits only give uncorrected estimates, but you can convert from an MLE estimate to a bias-corrected estimate if you know the sample size.

Take-home messages

Fitting a model by maximum likelihood is a general method that applies to any functional form and any distributional form. It has the disadvantage that it requires numerical optimization, which can sometimes be slow and difficult to achieve.

Linear models with Normally-distributed errors can be fit quickly and accurately by a fast matrix calculation, no matter how complex they are.

Usually, in statistics, we are attempting to infer something about a population that is larger than our sample of data. Once we do that, the differences between a sample and the population it comes from show that some estimated parameters will be biased estimates of their population-level versions.