Basics of Statistical Models

Statistics is about building models. Models vary in how much they are like the thing they are trying to model, and in how simple they are. These two considerations are inversely related, i.e., making a model more complex makes it more similar to the thing being modelled.

Suppose I set out to build a model of an airplane. I could begin with something like this:

![Diagram of an airplane model]

The horizontal part represents the airplane’s wings, and the vertical part represents its body. This model has the virtue of being very simple, but the disadvantage that it really isn’t similar enough to an airplane to teach us anything. So if I wished, I could making the model more and more elaborate, until it became three dimensional, then a working model of an airplane, until finally, it wouldn’t really be a model anymore, it would be an airplane. At this point, the model isn’t very useful because it is as complex as the thing we are trying to model.

This class is basically about building models of data that find some kind of ideal point between models that are too simple and models that are too complex. The basic equation describing statistical models is:

$\text{DATA} = \text{MODEL} + \text{ERROR}$

That is, the model approximates the actual data, and whatever is left over is error. We talk about the complexity of our models in terms of the number of parameters ($p$) that the model has. A parameter is a variable term that is optimized to fit our data. You might say that the plus sign model had two parameters: a cross and a body. The number of data points is called $n$. Notice that the point of modelling is to have $n \geq p$, otherwise the model is as complex as the data.

So let’s get down to business. Here are some data:

$$y_i = 1, 3, 5, 9, 14$$

That is, measurements have been collected from $n=5$ subjects. How can we build a model of these data? Here is a model with 0 parameters:

$$Y_i = 0 + e_i$$
This model has 0 parameters because 0 wasn’t estimated from the data. I just picked it out of the air; it might just as easily been 100 or 17. We will be interested in 0 parameter models shortly, but first it will be more instructive to think about a 1 parameter model for a while:

\[ Y_i = b_0 + e_i \]

\( b_0 \) is a parameter. That is, we are going to find a way to estimate it from the data. Another way to write this model is

\[ \hat{Y}_i = b_0 \]

where \( \hat{Y} \) denotes a \( Y \) with a pointy hat over it, and means "estimated \( Y \)." Estimated \( Y \)s are what the model predicts, as opposed to the actual \( Y \)s, which are what the model predicts plus any error that is left over. I will (lazily) tend to switch back and forth between the two. Most of the time it doesn’t matter.

Anyway, how do we go about deciding on a value for \( b_0 \)? We do it by finding a way to minimize error. According to our basic equation,

\[ \text{ERROR} = \text{DATA} - \text{MODEL}. \]

Or, more algebraically,

\[ e_i = Y_i - \hat{Y}_i \]

The total error for any given set of estimates is given by,

\[ \sum(Y_i - \hat{Y}) \]

The only problem with this measure of error is that positive errors and negative errors will tend to cancel each other out. Really we need to get rid of the signs and just deal with the magnitude of the errors. The two obvious ways to do this (there are an infinity of ways, and many more than two that are actually used; we will only discuss two, and only emphasize one). One way is to square the errors to get rid of the signs:

\[ \sum(Y_i - \hat{Y}_i)^2 \]

This is called least squared error, and it is the most common error function we will use. The other obvious way is called absolute error, and involves taking the absolute values of the errors to get rid of the signs:
The choice between these two error functions (and many others) is not obvious, and ultimately depends on what you are doing. We will focus on least squared error because that is the traditional way, and leave other methods to a more advanced course.

Getting back to our data, how do we select a value for $b_0$? The technique is to try different values, and compute the total amount of error for each one. The value of $b_0$ that produces the smallest amount of error is our choice.

[See SAS Program Minimizing Error]

We’ll talk about what this program does in class. Note that you can cut and paste it into your SAS program window to run it. Here is the graph that results:

Notice that as the values of X get closer to a reasonable value for $b_0$, the total amount of error gets smaller and smaller, reaches a minimum between 6 and 7, then starts to get larger again. The same is true for the red line, but it reaches a minimum around 5.

The value of $b_0$ at which error reaches a minimum is our least squares estimate for the parameter. What is this value actually equal to? To answer this question we need to revert
briefly to calculus. If you don’t know how to do this, don’t worry about it. Before we do that, however, I do want to review a little about the algebra of sums, which will be useful to have throughout the course. Here are three rules:

\[ \Sigma(y_1 + y_2) = \Sigma y_1 + \Sigma y_2 \]
\[ \Sigma cy = c \Sigma y \]
\[ \Sigma c = nc \]

Now let’s proceed to the derivation of the minimum of the least squares error function for a single parameter model. We want to minimize the expression:

\[ \Sigma(Y_i - b_0)^2 \]
\[ \Sigma(Y_i^2 - 2b_0 \Sigma Y_i + nb_0^2) \]

Taking the first derivative with respect to \( b_0 \) and setting equal to 0, we get:

\[ -2 \Sigma Y_i + 2n b_0 = 0 \]
\[ b_0 = \frac{\Sigma Y_i}{n} \]

Which is, of course, the familiar formula for the mean of a set of numbers. We learn something very important from this. The mean has the value it does for a reason. It is the least squares estimate of \( b_0 \) in a one parameter model, the value that minimizes the sum of squared errors in using the one parameter model to estimate the actual data. We can use a similar argument to show that the value of \( b_0 \) that minimizes absolute error is the median, i.e. the value that has the same number of values less than it as greater than it.

**Description of Error Distributions**

Once we have fit a model like our one parameter model to some data, the next step is to quantify error. This tells us in a general way how well our model fits. So when we fit a model like,

\[ Y_i = b_0 + e_i \]

to the data, error is equal to

\[ \Sigma(Y_i - \hat{Y}_i)^2 \]
For a one parameter model, $Y$hat is equal to $b_0$, so we can write the formula for the total amount of error, called sum of squared error or SSE, as

$$\Sigma(Y_i - b_0)^2$$

The average amount of error, known as mean squared error, is the average amount of squared error per observation, or,

$$\frac{\Sigma(Y_i - b_0)^2}{n}$$

Note that this formula only applies to populations as opposed to samples. We'll understand what that means next week. Finally, we sometimes want to know the amount of error in units that are not squared. We do this by taking the square root of mean squared error, obtaining something called root mean squared error, or RMSE.

Of course, $b_0$, MSE and RMSE all have traditional interpretations as part of what is called descriptive statistics. $b_0$ as we have already seen is equal to the mean, which along with the median and the mode (The score occurring most frequently in the distribution) are usually sometimes called "measures of central tendency". [Question: What error function would give you the mode as an estimate of $b_0$?]. In a one parameter model, MSE is known as the variance and RMSE is known as the standard deviation. The variance, written as $\sigma^2$ in populations and $s^2$ in samples (again, more on this distinction next week) and standard deviation, written as $s$ in populations and $s$ in samples, are measures of how much error there is around $b_0$, that is to say, how well the mean estimates all the scores in the distribution. So they are a measure of dispersion around the mean. If all the scores are the same: 5 5 5 5 5, the mean is 5 and a 1 parameter model fits perfectly. The more different the scores are from each other, the worse the fit of a single mean becomes, and $s$ and $s^2$ increase.

We can use what we learned above about summation to understand what happens to means and standard deviations when we change variables. So, for example, what about:

$$\overline{(y + c)}$$

Working through the formula, we have,

$$\sum\frac{(y + c)}{n}$$

Which based on the above is,

$$\sum\frac{y + nc}{n}$$
Which is equal to:

\[ \overline{y} + c \]

Try for yourself \((\overline{cy})\).

Changing variables by adding constants to them and multiplying them by constants are called **linear transformations**, because the new variable that is created is a linear function of the old one. Two linear transformations will be particularly important to us as we proceed. Note that based on the above:

\[
\text{Mean}(Y_i - \text{Mean}(Y_i)) = 0
\]

\[
\text{SD}(Y_i / \text{SD}(Y_i)) = 1
\]

Therefore, if we transform a variable by subtracting the mean from each score and dividing by the standard deviation, we get a variable with a mean of zero and a standard deviation of 1. These are called **z scores**:

\[
z_i = \frac{Y_i - \text{Mean}(Y_i)}{\text{SD}(Y_i)}
\]

Now let’s start looking at some semi-real data.

[Refer to SAS Program Automobile Data]

auto.dat gives data on drinking age and automobile fatality rates in the fifty states. Note that this program runs on my computer. You will have to make some changes to get it to run on yours. The first part is PROC UNIVARIATE, which provides basic descriptive statistics for 0 and 1 parameter models. Here are the results:

**Variable=FATRATE**

**Moments**

<table>
<thead>
<tr>
<th>Moment</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>50</td>
</tr>
<tr>
<td>Mean</td>
<td>3.572</td>
</tr>
<tr>
<td>Std Dev</td>
<td>0.857628</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.720735</td>
</tr>
<tr>
<td>USS</td>
<td>674</td>
</tr>
<tr>
<td>CV</td>
<td>24.00975</td>
</tr>
<tr>
<td>T:Mean=0</td>
<td>Pr&gt;</td>
</tr>
<tr>
<td>Num ^= 0</td>
<td>Num &gt; 0 50</td>
</tr>
<tr>
<td>M(Sign)</td>
<td>25</td>
</tr>
<tr>
<td>Sgn Rank</td>
<td>637.5</td>
</tr>
</tbody>
</table>

\(N\), mean and standard deviation should be self explanatory. The field CSS (corrected sums of squares, "corrected" for the mean) is the sum of squared error for the one parameter model. The USS, or uncorrected sum of squares, is the SSE for a zero
parameter model in which $b_0$ has been set to zero. What, then, is USS equal to? Can you write down a formula? The next part of the output gives the percentiles:

```
Quantiles (Def=5)

100% Max  5.8 99%  5.8
75% Q3  4.95%  5.4
50% Med 3.45 90%  4.8
25% Q1  3.1 10%  2.7
 0% Min 1.7  5%  2.3
 1% 1.7
 Range 4.1
 Q3-Q1 1
 Mode  3.5
```

The best way to define quantiles is to say that the median is the 50$^{th}$ quantile. The 25$^{th}$ quantile is the score that is greater than 25% of the scores, the 75$^{th}$ quantile is greater than 75%, the 95$^{th}$ is greater than 95% and so forth. The range is the difference between the highest and lowest score; Q3-Q1 is called the interquartile range and is a measure of dispersion related to the median.