1 Models and Errors

The fundamental equations of modeling are

\[
\text{Observation} = \text{Model} + \text{Error} \quad \text{Observation} = \text{Fitted} + \text{Residual}
\]

Our observation is a quantitative response variable – we’ll usually call it \( y \). Our model depends on several explanatory variables – we’ll call them \( x_1, \ldots, x_k \). Each of these could be quantitative or categorical. Our model also depends on unknown parameters – we’ll use Greek letters \( \beta_0, \ldots, \beta_m \) for these. We will estimate the parameters using our data – the fitted values of the parameters are called \( \hat{\beta}_0, \ldots, \hat{\beta}_m \). Finally, our error will be called \( \varepsilon \). We usually estimate the parameters by choosing them so that the sum of the squares of the residuals is minimized.

In the case of linear regression with a single explanatory variable \( x \), these equations amount to:

\[
y = \beta_0 + \beta_1 x + \varepsilon \quad y = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon.
\]

We are eventually going to introduce wide variety of models. However we are going to continue to use minimizing the sums of squares of residuals as our criteria for a good fit. In order for this to be a good strategy for fitting the model, we need to make some statistical assumptions of the errors.

We will continue to assume that we have \( n \) observations. In particular, a typical case of our data looks like \( (x_{1,i}, \ldots, x_{k,i}, y_i) \) where \( x_{j,i} \) is the value of the \( j^{th} \) explanatory variable for the \( i^{th} \) observation. This in turn means that there really are \( n \) errors, \( \varepsilon_1, \ldots, \varepsilon_n \), one for each observation. In order to decide what method of fitting the model makes sense, we have to make some assumptions about these errors. There are five important assumptions about the errors – not all of which will be exactly true in every case that we use the method.

\( \varepsilon \) is a random variable

Here we are assuming that we are producing data by some process and that the errors in this model produced by this process are random. In other words, the errors are unpredictable but have a particular distribution over all possible data values. This distribution is determined by our sampling process. As usual in this course, randomness arises either because we choose random samples from a finite population or because we are modelling a process as a random variable. So what we are assuming is that each observation has a deterministic component (determined by the model) and a random component.

\( \varepsilon \) has mean 0

This assumption means that we sometimes expect that \( y \) is greater than the value predicted by the model and sometimes less. However over all possible data points (whatever that means in the particular case at hand), the errors of the model have mean 0.

Another way of saying this is that we assume that our model is really a model for the mean value of all possible observations with fixed values of the explanatory variable. In the simple case of a line, we are assuming that for every value \( x_0 \) of \( x \), the “true” (rather than “fitted”) regression line goes through the point \((x_0, y_0)\) where \( y_0 \) is the mean value for all \( y \)’s with that fixed \( x = x_0 \).
\( \varepsilon \) has a constant standard deviation \( \sigma \) over all possible values of the explanatory variables

The standard deviation of \( \varepsilon \) is a measure of how far off the model predictions are on average (even if we know the true model). The assumption that these errors have approximately the same size, regardless of the values of the explanatory variable is really quite a strong assumption. To see why we might want such an assumption, consider the simple case of linear regression and suppose that instead that \( \varepsilon \) has a larger standard deviation if \( x \) is large. This means that data values for large \( x \) are likely to be more spread out than those for small \( x \) and so these larger \( x \) values might tend to influence the regression line more. There is a way to correct for this (instead of minimizing the sums of squares of residuals we minimize a certain weighted sum of squares) but we will not pursue this in this course.

The random variables \( \varepsilon_i \) are independent

If we do not collect data in a way such that the errors are independent, the dependency might affect our regression line in unpredictable ways. (That’s about all we can say here! Data can be dependent in a variety of different ways.)

The random variable \( \varepsilon \) has a normal distribution

This is of course a very strong assumption and one most likely to be false (if it even makes sense to talk about the possibility that it might be true). But it is an attractive assumption. It says that the errors have a nice bell-shaped curve. Some kinds of measurement errors actually fit this model but for the most part it is an approximation which is most useful when we want to make inferences about the parameters in the model (in much the same way that it is useful when we use the \( t \)-test in the case of making inferences about a single mean).

Fitting the model

If all the above assumptions are true, fitting the model by choosing the parameters \( \beta_1, \ldots, \beta_m \) to minimize the sums of squares of residuals is a good thing to do. Good here means several things and the precise meaning of good in this case is beyond the scope of the course. We can give an example of what good means in the case of fitting a single line \( y = \beta_0 + \beta_1 x \) however. In this case, the least squares estimates \( \hat{\beta}_i \) of the parameters will over all possible data sets have average \( \beta_i \). In other words, though for any particular set of data we would never expect \( \hat{\beta}_1 = \beta_1 \), it is the case that over all possible datasets, the average of \( \hat{\beta}_1 \) is \( \beta_1 \). It is also the case that, on average, \( \hat{\beta}_1 \) is likely to be closer to \( \beta_1 \) than if we would use some other procedure to estimate \( \beta_1 \).