10.4: So Now What? Implications of Residual Analysis

What should you do if you observe patterns in the residuals that seem to violate the assumptions of OLS? If you find deviant cases – outliers that are shown to be highly influential – you need to first evaluate the specific cases (observations). Is it possible that the data were miscoded? We hear of many instances in which missing value codes (often "-99") were inadvertently left in the dataset. `R` would treat such values as if they were real data, often generating glaring and influential outliers. Should that be the case, recode the offending variable observation as missing ("NA") and try again.

But what if there is no obvious coding problem? It may be that the influential outlier is appropriately measured, but that the observation is different in some theoretically important way. Suppose, for example, that your model included some respondents who – rather than diligently answering your questions – just responded at random to your survey questions. They would introduce noise and error. If you could measure these slackers, you could either exclude them or include a control variable in your model to account for their different patterns of responses. We will discuss inclusion of model controls when we turn to multiple regression modeling in later chapters.

What if your residual analysis indicates the presence of heteroscedasticity? Recall that this will undermine your ability to do hypothesis tests in OLS. There are several options. If the variation in fit over the range of the predicted value of YY could plausibly result from the omission of an important explanatory variable, you should respecify your model accordingly (more on this later in this book). It is often the case that you can improve the distribution of residuals by including important but previously omitted variables. Measures of income, when left out of consumer behavior models, often have this effect.

Another approach is to use a different modeling approach that accounts for the heteroscedasticity in the estimated standard error. Of particular utility are robust estimators, which can be employed using the `rlm` (robust linear model) function in the `MASS` package. This approach increases the magnitude of the estimated standard errors, reducing the t-values and resulting p-values. That means that the "cost" of running robust estimators is that the precision of the estimates is reduced.

Evidence of non-linearity in the residuals presents a thorny problem. This is a basic violation of a central assumption of OLS, resulting in biased estimates of AA and BB. What can you do? First, you can respecify your model to include a polynomial; you would include both the XX variable and a square of the XX variable. Note that this will require you to recode XX. In this approach, the value of XX is constant, while the value of the square of XX increases exponentially. So a relationship in which YY decreases as the square of XX increases will provide a progressively steeper slope as XX rises. This is the kind of pattern we observed in the example in which political ideology was used to predict the perceived risk posed by climate change.