Joint Influence

Thus far we have examined ways to detect individual data points that can change a fitted regression model, whether via a large residual, or high leverage, or from a combination of both. However, with many data sets outliers and leverage points may exist as groups instead of individual points. In these situations many of the measures that we have examined may not be able to detect such a group of data. In this lecture we discuss two ways to investigate the possibility of a set of data jointly influencing the regression fit.

Added-Variable or Partial Regression plots

As we saw in an earlier chapter, we can learn more about the effect of a single covariate X_i in a multiple regression model on the dependent variable Y with the following approach. Suppose that the covariate of interest is the first covariate. Then we fit two separate models to predict Y and X_1 from all of the other covariates:

$$\begin{aligned} Y_i &= A^{(1)} + B_2^{(1)} X_{i2} + \cdots + B_k^{(1)} X_{ik} + Y_i^{(1)}, \text{ and} \\ X_{i1} &= C^{(1)} + D_2^{(1)} X_{i2} + \cdots + D_k^{(1)} X_{ik} + X_i^{(1)}. \end{aligned}$$

Now the residuals $Y_i^{(1)}$ and $X_i^{(1)}$ have the properties that i) the least-squares slope from the regression of $Y_i^{(1)}$ on $X_i^{(1)}$ equals the least-squares slope B_1 from the full multiple regression, ii) the residuals from the regression of $Y_i^{(1)}$ on $X_i^{(1)}$ are the same as those from the full regression, and iii) the variation in $X_i^{(1)}$ is the conditional variation in X_1 holding the other X's constant, thus the standard error of B_1 from the regression of $Y_i^{(1)}$ on $X_i^{(1)}$ is the same as the multiple regression error of B_1.

We can plot $Y_i^{(1)}$ against $X_i^{(1)}$ to examine leverage and influence of data points on B_1, and we can make similar plots for other covariates X_i.

Forward Search

A very different approach was suggested by Atkinson and Riani (2000) (with extensions in subsequent papers) to examine the joint influence of data points on a fitted regression model. The approach is to start with a highly resistant fit such as a least median of squares regression to obtain a small subset of the data, which is free of outliers and leverage points. Linear regression models are then fitted iteratively, starting with the minimum number of data

points from the outlier and leverage free set. The data are added one-by-one in the following way: Suppose that m observations are in the current model. The $m+1$ observations in the next step are those with the smallest residuals from the current fit, which typically will be the current m observations plus one new data point. By following how the regression coefficients and other outputs of interest change as we add data we can learn a great deal about joint influence on the regression model.

How should we handle unusual data?

Some considerations on how to deal with data points identified in this chapter.