

## 17 Lecture 17: Feb 26

### Last time

- Unusual and influential data (JF chapter 11)

### Today

- Added-variable plots
- Should unusual data be discarded

### Added-variable plots

Unlike the case of SLR, the scatterplot with the response variable and one predictor gives only the marginal effect in MLR. Instead, the added-variable plot (also called a partial-regression plot or a partial-regression leverage plot) gives a graphical inspection over each dimension.

Let  $\hat{Y}_i^{(1)}$  represent the residuals from the least-squares regression of  $Y$  on all the  $X$ s except  $X_1$ , in other words, the residuals from the following fitted regression equation:

$$Y_i = \tilde{\beta}_0^{(1)} + \tilde{\beta}_2^{(1)} X_{i2} + \cdots + \tilde{\beta}_p^{(1)} X_{ip} + \tilde{Y}_i^{(1)}$$

where the parenthetical superscript (1) indicates the omission of  $X_1$  from the right-hand side of the regression equation. Likewise,  $\check{X}_i^{(1)}$  is the residual from the least-squares regression of  $X_1$  on all the other  $X$ s:

$$X_{i1} = \check{\beta}_0^{(1)} + \check{\beta}_2^{(1)} X_{i2} + \cdots + \check{\beta}_p^{(1)} X_{ip} + \check{X}_i^{(1)}$$

Then, the residuals  $\tilde{Y}_i^{(1)}$  and  $\check{X}_i^{(1)}$  have the following interesting properties:

1. The slope from the least-squares regression of  $\tilde{Y}_i^{(1)}$  on  $\check{X}_i^{(1)}$  is simply the least-squares slope  $\hat{\beta}_1$  from the *full* multiple regression.
2. The residuals from the simple regression of  $\tilde{Y}_i^{(1)}$  on  $\check{X}_i^{(1)}$  are the same as those from the full regression, that is

$$\tilde{Y}_i^{(1)} = \hat{\beta}_1 \check{X}_i^{(1)} + \hat{\epsilon}_i$$

3. The variation of  $\check{X}_i^{(1)}$  is the *conditional variation* of  $X_1$  holding the other  $X$ s constant.

Figure 17.1 shows that the conditional variation is smaller than its marginal variation – much smaller when  $X_1$  is strongly collinear with other  $X$ s,

Figure 17.2 illustrates the added-variable plots using the Duncan's data.

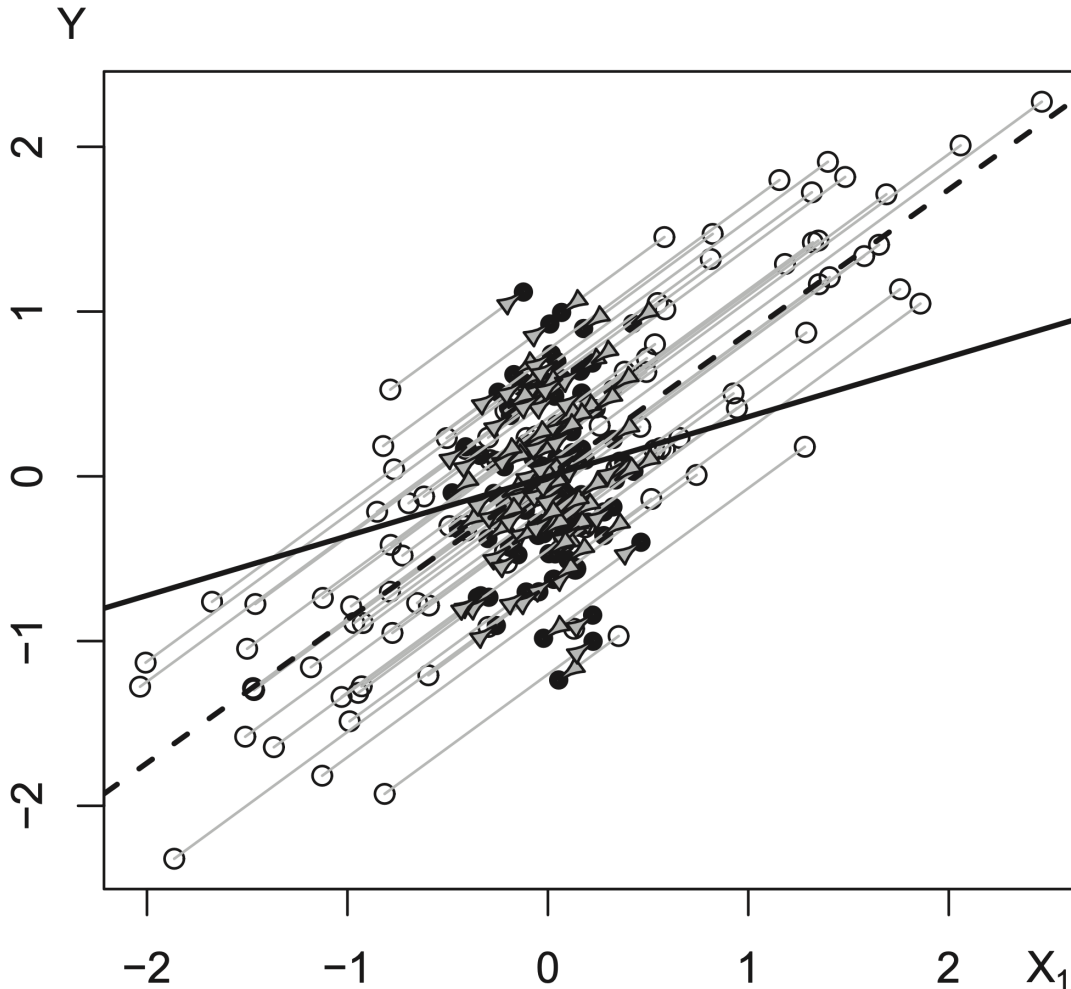


Figure 17.1: The marginal scatterplot (open circles) for  $Y$  and  $X_1$  superimposed on the added-variable plot (filled circles) for  $X_1$  in the regression of  $Y$  on  $X_1$  and  $X_2$ . The variables  $Y$  and  $X_1$  are centered at their means to facilitate the comparison of the two sets of points. The arrows show how the points in the marginal scatterplot map into those in the AV plot. In this contrived data set,  $X_1$  and  $X_2$  are highly correlated ( $r_{12} = 0.98$ ), and so the conditional variation in  $X_1$  (represented by the horizontal spread of the filled points) is much less than its marginal variation (represented by the horizontal spread of the open points). The broken line gives the slope of the marginal regression of  $Y$  on  $X_1$  alone, while the solid line gives the slope  $\hat{\beta}_1$  of  $X_1$  in the MLR of  $Y$  on both  $X$ s. JF Figure 11.9.

### Should unusual data be discarded?

In practice, although problematic data should not be ignored, they also should not be deleted automatically and without reflection:

- It is important to investigate *why* an observation is unusual. Truly “bad” data (e.g., an error in data entry ) can often be corrected or, if correction is not possible, thrown away. When a discrepant data point is correct, we may be able to understand why the

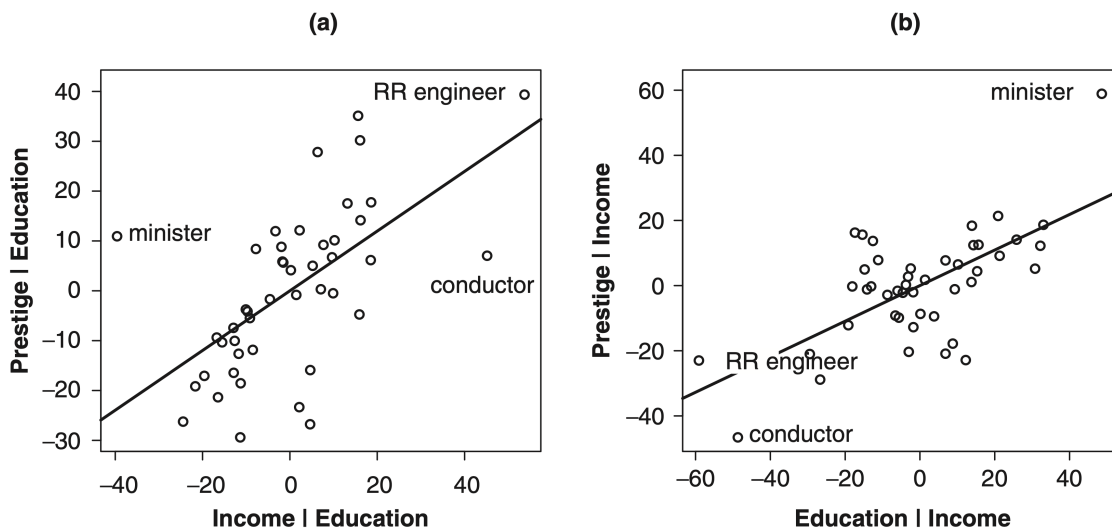


Figure 17.2: Added-variable plots for Duncan's regression of occupational prestige on the (a) income and (b) education levels of 45 US occupations in 1950. Three unusual observations, *ministers*, *conductors*, and *railroadengineers*, are identified on the plots. The added-variable plot for the intercept  $\hat{\beta}_0$  is not shown. JF Figure 11.10.

observation is unusual. For Duncan's data, for example, it makes sense that ministers enjoy prestige not accounted for by the income and educational levels of the occupation and for a reason not shared by other occupations. In a case like this, where an outlying observation has characteristics that render it unique, we may choose to set it aside from the rest of the data.

- Alternatively, outliers, high-leverage points, or influential data may motivate model respecification, and the pattern of unusual data may suggest the introduction of additional explanatory variables. We noticed, for example, that both conductors and railroad engineers had high leverage in Duncan's regression because these occupations combined relatively high income with relatively low education. Perhaps this combination of characteristics is due to a high level of unionization of these occupations in 1950, when the data were collected. If so, and if we can ascertain the levels of unionization of all of the occupations, we could enter this as an explanatory variable, perhaps shedding further light on the process determining occupational prestige.
- Except in clear-cut cases, we are justifiably reluctant to delete observations or to re-specify the model to accommodate unusual data. Some researchers reasonably adopt alternative estimation strategies, such as robust regression, which continuously down-weights outlying data rather than simply discarding them. Because these methods assign zero or very small weight to highly discrepant data, however, the result is generally not very different from careful application of least squares, and, indeed, robust-regression weights can be used to identify outliers.
- Finally, in large samples, unusual data substantially alter the results only in extreme instances. Identifying unusual observations in a large sample, therefore, should be

regarded more as an opportunity to learn something about the data not captured by the model that we have fit, rather than as an occasion to reestimate the model with the unusual observations removed.