Lecture Notes

7. Unusual and Influential Data

1. Introduction

- Linear statistical models make strong assumptions about the structure of data, which often do not hold in applications.
- The method of least-squares is very sensitive to the structure of the data, and can be markedly influenced by one or a few unusual observations.
- We could abandon linear models and least-squares estimation in favor of nonparametric regression and robust estimation.
- Alternatively, we can adapt and extend methods for examining and transforming data to diagnose problems with a linear model, and to suggest solutions.

2. Goals:

- To distinguish among regression outliers, high-leverage observations, and influential observations.
- To show how outlyingness, leverage, and influence can be measured.
- To introduce added-variable ('partial-regression') plots as a means of displaying leverage and influence on particular coefficients.

3. Outliers, Leverage, and Influence

- Unusual data are problematic in linear models fit by least squares because they can unduly influence the results of the analysis, and because their presence may be a signal that the model fails to capture important characteristics of the data.
- Some central distinctions are illustrated in Figure 1 for the simple regression model $Y = \alpha + \beta X + \varepsilon$.
  - In simple regression, an *outlier* is an observation whose response-variable value is conditionally unusual given the value of the explanatory variable.
  - In contrast, a univariate outlier is a value of Y or X that is unconditionally unusual; such a value may or may not be a regression outlier.

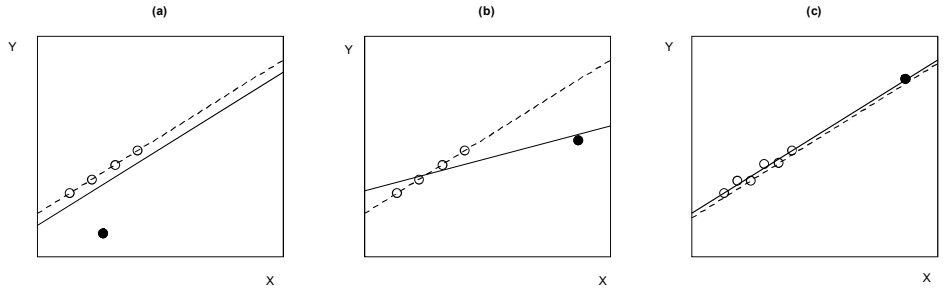

Figure 1. Unusual data in regression: (a) a low-leverage and hence uninfluential outlier; (b) a high-leverage and hence influential outlier; (c) a high-leverage in-line observation. In each case, the solid line is the least--squares line for all of the data; the broken line is the least-squares line with the unusual observation omitted.

- Regression outliers appear in (a) and (b).
  - In (a), the outlying observation has an X-value that is at the center of the X distribution; deleting the outlier has little impact on the least-squares fit.
  - In (b), the outlier has an unusual X-value; its deletion markedly affects both the slope and the intercept. Because of its unusual X-value, the outlying last observation in (b) exerts strong *leverage* on the regression coefficients, while the outlying middle observation in (a) is at a low-leverage point. The combination of high leverage with a regression outlier produces substantial *influence* on the regression coefficients.
  - In (c), the last observation has no influence on the regression coefficients even though it is a high-leverage point, because this observation is in line with the rest of the data.

- The following heuristic formula helps to distinguish among the three concepts of influence, leverage and discrepancy ('outlyingness'):

  Influence on Coefficients = Leverage $\times$ Discrepancy

► A simple example with real data from Davis (1990) appears in Figure 2. The data record the measured and reported weight of 183 male and female subjects who engage in programs of regular physical exercise. Davis's data can be treated in two ways:

1. We could regress reported weight (RW) on measured weight (MW), a dummy variable for sex (F, coded 1 for women and 0 for men), and an interaction regressor (formed as the product $MW \times F$):

$$\begin{array}{rcccccc} \widehat{RW} & = & 1.36 & + & 0.990MW & + & 40.0F & - & 0.725(MW \times F) \\ & & (3.28) & & (0.043) & & (3.9) & & (0.056) \end{array}$$

$$R^2 = 0.89 \quad S_E = 4.66$$

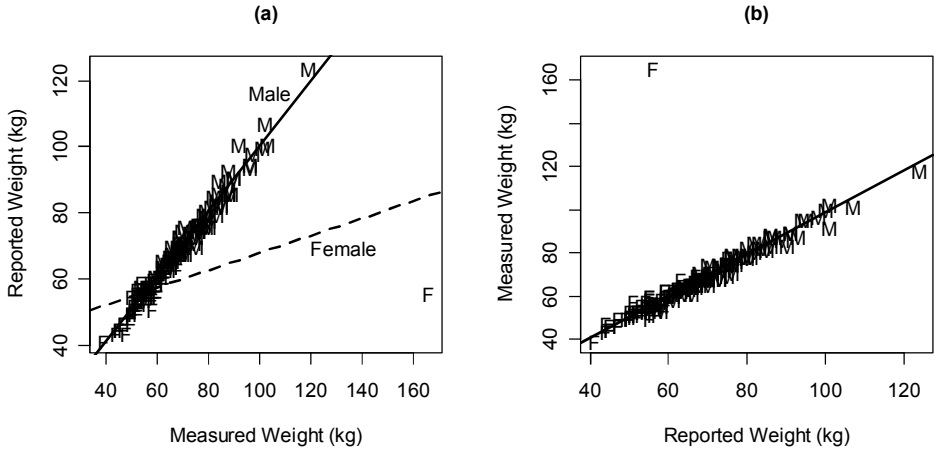

Figure 2. (a) Regressing reported weight on measured weight, sex, and their interaction; (b) regressing measured weight on reported weight, sex, and their interaction.

- Were these results taken seriously, we would conclude that men are unbiased reporters of their weights (because $A \approx 0$ and $B_1 \approx 1$), while women tend to over-report their weights if they are relatively light and under-report if they are relatively heavy.
- The figure makes it clear that the differential results for women and mean are due to one erroneous data point.
- Correcting the data produces the regression

$$\begin{array}{rcccccc} \widehat{RW} & = & 1.36 & + & 0.990MW & + & 1.98F & - & 0.0567(MW \times F) \\ & & (1.58) & & (0.021) & & (2.45) & & (0.0385) \end{array}$$
$$R^2 = 0.97 \quad S_E = 2.24$$

2. We could regress measured weight on reported weight, sex, and their interaction:

$$\begin{array}{rcccccc} \widehat{MW} & = & 1.79 & + & 0.969RW & + & 2.07F & - & 0.00953(MW \times F) \\ & & (5.92) & & (0.076) & & (9.30) & & (0.147) \end{array}$$
$$R^2 = 0.70 \quad S_E = 8.45$$

- The outlier does not have much impact on the regression coefficients because the value of RW for the outlying observation is near $\overline{RW}$ for women.
- There is, however, a marked effect on the multiple correlation and standard error: For the corrected data, $R^2 = 0.97$ and $S_E = 2.25$.

4. Assessing Leverage: Hat-Values

▶ The *hat-value* h_i is a common measure of leverage in regression. These values are so named because it is possible to express the fitted values $\widehat{Y}_j$ ('Y-hat') in terms of the observed values Y_i:

$$\widehat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{jj}Y_j + \cdots + h_{nj}Y_n = \sum_{i=1}^{n} h_{ij}Y_i$$

- Thus, the weight h_{ij} captures the contribution of observation Y_i to the fitted value $\widehat{Y}_j$: If h_{ij} is large, then the ith observation can have a substantial impact on the jth fitted value.

▶ Properties of the hat-values:
- $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, and so the hat-value $h_i \equiv h_{ii}$ summarizes the potential influence (the leverage) of Y_i on *all* of the fitted values.
- $1/n \leq h_i \leq 1$

- The average hat-value is $\overline{h} = (k+1)/n$.
- In simple-regression analysis, the hat-values measure distance from the mean of X:

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}$$

- In multiple regression, h_i measures distance from the centroid (point of means) of the X's, taking into account the correlational and variational structure of the X's, as illustrated for $k = 2$ in Figure 3. Multivariate outliers in the X-space are thus high-leverage observations. The response-variable values are not at all involved in determining leverage.

▶ For Davis's regression of reported weight on measured weight, the largest hat-value by far belongs to the 12*th* subject, whose measured weight was wrongly recorded as 166 kg.: $h_{12} = 0.714$. This quantity is many times the average hat-value, $\overline{h} = (3+1)/183 = 0.0219$.

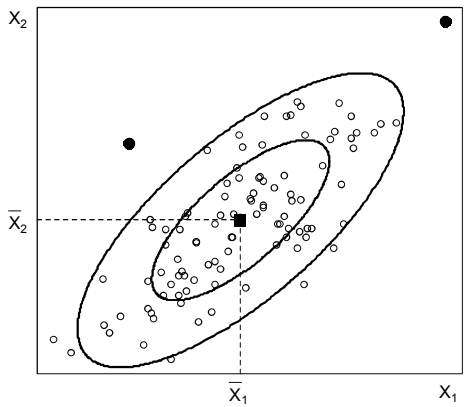

Figure 3. Contours of constant leverage in multiple regression with two explanatory variables, X_1 and X_2. The two observations marked with solid black dots are have equal hat-values.

- Recall Duncan's regression of occupational prestige on income and education for 45 U. S. occupations in 1950:

$$\widehat{\text{Prestige}} = \underset{(4.27)}{-6.06} + \underset{(0.120)}{0.599} \times \text{Income} + \underset{(0.098)}{0.546} \times \text{Education}$$

  - An index plot of hat-values for the observations in Duncan's regression is shown in Figure 4 (a), with a scatterplot for the explanatory variables in Figure 4 (b).

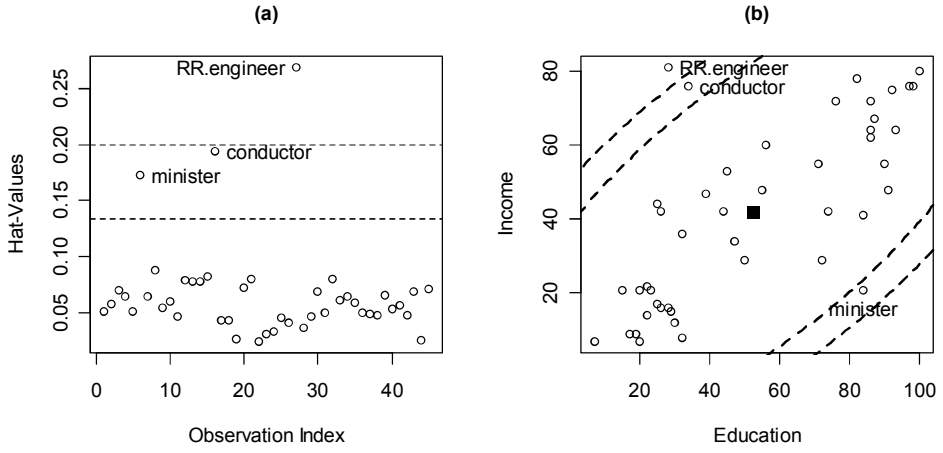

Figure 4. Duncan's occupational prestige regression: (a) hat-values; (b) scatterplot for education and income, showing contours of constant leverage at $2 \times \overline{h}$ and $3 \times \overline{h}$.

5. Detecting Outliers: Studentized Residuals

- Discrepant observations usually have large residuals, but even if the errors ε_i have equal variances (as assumed in the general linear model), the residuals E_i do not:

$$V(E_i) = \sigma_\varepsilon^2(1 - h_i)$$

  - High-leverage observations tend to have small residuals, because these observations can coerce the regression surface to be close to them.

- Although we can form a *standardized residual* by calculating

$$E_i' = \frac{E_i}{S_E\sqrt{1 - h_i}}$$

this measure is slightly inconvenient because its numerator and denominator are not independent, preventing E_i' from following a t-distribution: When $|E_i|$ is large, $S_E = \sqrt{\sum E_i^2/(n - k - 1)}$, which contains E_i^2, tends to be large as well.

- Suppose that we refit the model deleting the ith observation, obtaining an estimate $S_{E(-i)}$ of σ_ε that is based on the remaining $n-1$ observations.
  - Then the *studentized residual*

$$E_i^* = \frac{E_i}{S_{E(-i)}\sqrt{1-h_i}}$$

    has independent numerator and denominator, and follows a t-distribution with $n-k-2$ degrees of freedom.
  - An equivalent procedure for finding the studentized residuals employs a 'mean-shift' outlier model

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \gamma D + \varepsilon$$

    where D is a dummy regressor set to one for observation i and zero for all other observations:

$$D = \begin{cases} 1 & \text{for obs. } i \\ 0 & \text{otherwise} \end{cases}$$

- Thus

$$\begin{aligned} E(Y_i) &= \alpha + \beta_1 X_{i1} + \cdots \beta_k X_{ik} + \gamma \\ E(Y_j) &= \alpha + \beta_1 X_{j1} + \cdots \beta_k X_{jk} \text{ for } j \neq i \end{aligned}$$

  - It would be natural to specify this model if, before examining the data, we suspected that observation i differed from the others.
  - Then to test H_0: $\gamma = 0$, we can calculate $t_0 = \widehat{\gamma}/\text{SE}(\widehat{\gamma})$. This test statistic is distributed as t_{n-k-2} under H_0, and is the studentized residual E_i^*.

5.1 Testing for Outliers

- In most applications we want to look for *any* outliers that may occur in the data; we can in effect refit the mean-shift model n times, producing studentized residuals $E_1^*, E_2^*, ..., E_n^*$. (It is not literally necessary to perform n auxiliary regressions.)
  - Usually, our interest then focuses on the largest absolute E_i^*, denoted $E_{\max}^*$.
  - Because we have picked the biggest of n test statistics, it is not legitimate simply to use t_{n-k-2} to find a p-value for $E_{\max}^*$.
- One solution to this problem of simultaneous inference is to perform a *Bonferroni adjustment* to the p-value for the largest absolute E_i^*: Let $p' = \Pr(\, t_{n-k-2} > E_{\max}^*)$.
  - Then the Bonferroni p-value for testing the statistical significance of $E_{\max}^*$ is $p = 2np'$.

  - Note that a much larger $E_{\max}^*$ is required for a statistically significant result than would be the case for an ordinary individual t-test.
- Another approach is to construct a quantile-comparison plot for the studentized residuals, plotting against either the t or normal distribution.
- In Davis's regression of reported weight on measured weight, the largest studentized residual by far belongs to the incorrectly coded 12*th* observation, with $E_{12}^* = -24.3$.
  - Here, $n-k-2 = 183-3-2 = 178$, and $\Pr(t_{178} > 24.3) \approx 10^{-58}$.
  - The Bonferroni p-value for the outlier test is $p << 2 \times 183 \times 10^{-58} = 4 \times 10^{-56}$, an unambiguous result.
- For Duncan's occupational prestige regression, the largest studentized residual belongs to *ministers*, with $E_{\text{minister}}^* = 3.135$.
  - The Bonferroni p-value is $2 \times 45 \times \Pr(t_{45-2-2} > 3.135) = .143$.

6. Measuring Influence

- Influence on the regression coefficients combines leverage and discrepancy.
- The most direct measure of influence simply expresses the impact on each coefficient of deleting each observation in turn:

$$D_{ij} = B_j - B_{j(-i)} \text{ for } i = 1, ..., n \text{ and } j = 0, 1, ..., k$$

where the B_j are the least-squares coefficients calculated for all of the data, and the $B_{j(-i)}$ are the least-squares coefficients calculated with the ith observation omitted. (So as not to complicate the notation here, I denote the least-squares intercept A as B_0.)
- One problem associated with using the D_{ij} is their large number — $n(k+1)$.
  - It is useful to have a single summary index of the influence of each observation on the least-squares fit.

- Cook (1977) has proposed measuring the 'distance' between the B_j and the corresponding $B_{j(-i)}$ by calculating the F-statistic for the 'hypothesis' that $\beta_j = B_{j(-i)}$, for $j = 0, 1, ..., k$.
  - This statistic is recalculated for each observation $i = 1, ..., n$.
  - The resulting values should not literally be interpreted as F-tests, but rather as a distance measure that does not depend upon the scales of the X's.
  - Cook's statistic can be written (and simply calculated) as

$$D_i = \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

  - In effect, the first term in the formula for Cook's D is a measure of discrepancy, and the second is a measure of leverage.
  - We look for values of D_i that are substantially larger than the rest.

- Because all of the deletion statistics depend on the hat-values and residuals, a graphical alternative is to plot the E_i^* against the h_i and to look for observations for which both are big. A slightly more sophisticated version of this plot that incorporates Cook's D is given below.
- For Davis's regression of reported weight on measured weight, Cook's D points to the obviously discrepant 12*th* observation:

Cook's $D_{12} = 85.9$ (next largest, $D_{21} = 0.065$)

- For Duncan's regression, the largest Cook's D is for ministers, $D_6 = 0.566$.
  - Figure 5 displays a plot of studentized residuals versus hat-values, with the areas of the plotted circles proportional to values of Cook's D. The lines on the plot are at $E^* = \pm 2$ (on the vertical axis), and at $h = 2\overline{h}$ and $3\overline{h}$ (on the horizontal axis).
  - Four observations that exceed these cutoffs are identified on the plot.

- Notice that *reporters* have a relatively large residual but are at a low-leverage point, while *railroad engineers* have high leverage but a small studentized residual.
- In developing the concept of influence in regression, I have focused on changes in regression coefficients. Other regression outputs, such as the set of coefficient sampling variances and covariances, are also subject to influence.

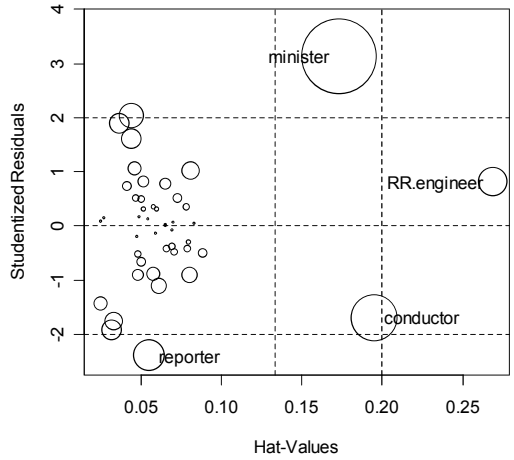

Figure 5. Influence plot for Duncan's occupational prestige regression. The areas of the circles are proportional to Cook's distance.

7. Numerical Cutoffs for Diagnostic Statistics

- I have refrained from suggesting specific numerical criteria for identifying noteworthy observations on the basis of measures of leverage and influence: I believe that it is generally more effective to examine the distributions of these quantities directly to locate unusual values.
  - For studentized residuals, outlier-testing provides a numerical cutoff, but even this is no substitute for graphical examination of the residuals.
- Nevertheless, numerical cutoffs can be of some use, as long as they are not given too much weight, and especially when they are employed to enhance graphical displays.
  - A line can be drawn on a graph at the value of a numerical cutoff, and observations that exceed the cutoff can be identified individually.
- Cutoffs for a diagnostic statistic may be derived from statistical theory, or they may result from examination of the sample distribution of the statistic.

- Cutoffs may be absolute, or they may be adjusted for sample size.
  - For some diagnostic statistics, such as measures of influence, absolute cutoffs are unlikely to identify noteworthy observations in large samples.
  - In part, this characteristic reflects the ability of large samples to absorb discrepant data without changing the results substantially, but it is still often of interest to identify *relatively* influential points, even if no observation has strong *absolute* influence.
  - The cutoffs presented below are derived from statistical theory:

7.1 Hat-Values

- Belsley, Kuh, and Welsch suggest that hat-values exceeding about twice the average $\overline{h} = (k+1)/n$ are noteworthy.
- In small samples, using $2 \times \overline{h}$ tends to nominate too many points for examination, and $3 \times \overline{h}$ can be used instead.

7.2 Studentized Residuals

- Beyond the issue of 'statistical significance,' it sometimes helps to call attention to residuals that are relatively large.
- Under ideal conditions, about five percent of studentized residuals are outside the range $|E_i^*| \leq 2$. It is therefore reasonable to draw attention to observations outside this range.

7.3 Measures of Influence

- Many cutoffs have been suggested for different measures of influence, including the following size-adjusted cutoff for *Cook's D*, due to Chatterjee and Hadi:

$$D_i > \frac{4}{n-k-1}$$

- Absolute cutoffs for D, such as $D_i > 1$, risk missing relatively influential data.

8. Joint Influence: Added-Variable Plots

- As illustrated in Figure 6, subsets of observations can be *jointly influential* or can offset each other's influence.
  - Influential subsets or multiple outliers can often be identified by applying single-observation diagnostics, such as Cook's D and studentized residuals, sequentially.
  - It can be important to refit the model after deleting each point, because the presence of a single influential value can dramatically affect the fit at other points, but the sequential approach is not always successful.
- Although it is possible to generalize deletion statistics to subsets of several points, the very large number of subsets usually renders this approach impractical.
- An attractive alternative is to employ graphical methods, and a particularly useful influence graph is the *added-variable plot* (also called a *partial-regression plot* or an *partial-regression leverage plot*).

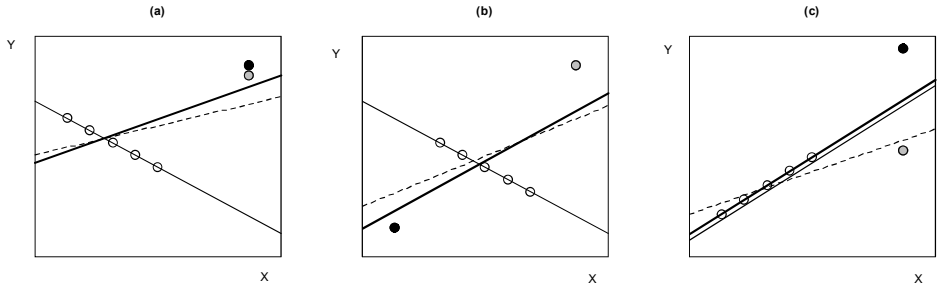

Figure 6. Jointly influential observations: (a) a pair of jointly influential points; (b) a widely separated jointly infuential pair; (c) two points that offset each other's influence. In each case the heavier solid line is the least-squares line for all of the data, the broken line deletes the black point, and the lighter solid line deletes both the gray and the black points.

- Let $Y_i^{(1)}$ represent the residuals from the least-squares regression of Y on all of the X's with the exception of X_1:

$$Y_i = A^{(1)} + B_2^{(1)}X_{i2} + \cdots + B_k^{(1)}X_{ik} + Y_i^{(1)}$$

- Likewise, $X_i^{(1)}$ are the residuals from the least-squares regression of X_1 on all the other X's:

$$X_{i1} = C^{(1)} + D_2^{(1)}X_{i2} + \cdots + D_k^{(1)}X_{ik} + X_i^{(1)}$$

- The notation emphasizes the interpretation of the residuals $Y^{(1)}$ and $X^{(1)}$ as the parts of Y and X_1 that remain when the effects of $X_2, ..., X_k$ are 'removed.'
- The residuals $Y^{(1)}$ and $X^{(1)}$ have the following interesting properties:

1. The slope from the least-squares regression of $Y^{(1)}$ on $X^{(1)}$ is simply the least-squares slope B_1 from the full multiple regression.

2. The residuals from the simple regression of $Y^{(1)}$ on $X^{(1)}$ are the same as those from the full regression:

$$Y_i^{(1)} = B_1X_i^{(1)} + E_i$$

   No constant is required, because both $Y^{(1)}$ and $X^{(1)}$ have means of 0.

3. The variation of $X^{(1)}$ is the conditional variation of X_1 holding the other X's constant and, as a consequence, the standard error of B_1 in the auxiliary simple regression

$$\text{SE}(B_1) = \frac{S_E}{\sqrt{\sum X_i^{(1)^2}}}$$

   is (except for df) the multiple-regression standard error of B_1. Unless X_1 is uncorrelated with the other X's, its conditional variation is smaller than its marginal variation — much smaller, if X_1 is strongly collinear with the other X's.

- Plotting $Y^{(1)}$ against $X^{(1)}$ permits us to examine leverage and influence on B_1. Because of properties 1–3, this plot also provides a visual impression of the precision of estimation of B_1.
- Similar added-variable plots can be constructed for the other regression coefficients:

  Plot $Y^{(j)}$ versus $X^{(j)}$ for each $j = 0, ..., k$

▶ Illustrative added-variable plots are shown in Figure 7, using data from Duncan's regression of occupational prestige on the income and educational levels of 45 U.S. occupations:

$$\widehat{\text{Prestige}} = \underset{(4.27)}{-6.06} + \underset{(0.120)}{0.599} \times \text{Income} + \underset{(0.098)}{0.546} \times \text{Education}$$

$$R^2 = 0.83 \quad S_E = 13.4$$

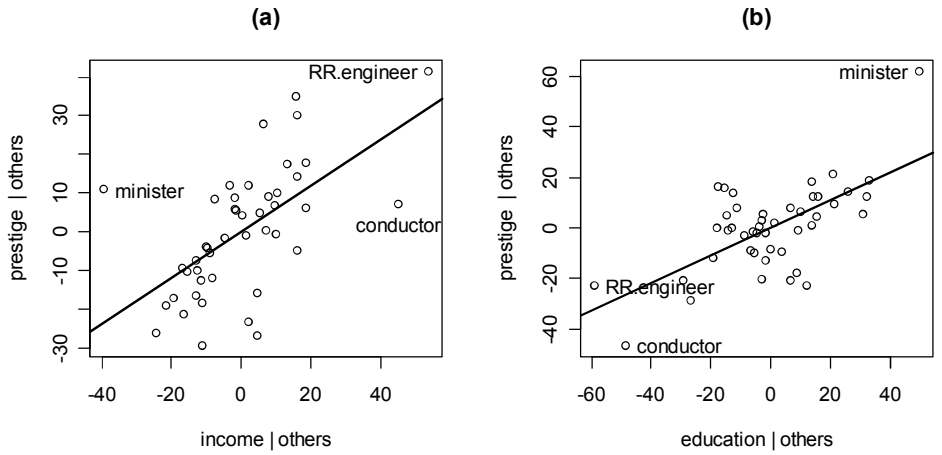

Figure 7. Added-variable plots for Duncan's occupational prestige regression, (a) for income, and (b) for education.

- The added-variable plot for income (a) reveals three unusual data points:
  - *ministers*, whose income is unusually low given the educational level of the occupation; and
  - *railroad conductors* and *railroad engineers*, whose incomes are unusually high given education.
  - Together, *ministers* and *railroad conductors* reduce the income slope; *railroad engineers*, while a high-leverage point, are more in line with the rest of the data.
  - Remember that the horizontal variable in this added-variable plot is the residual from the regression of income on education, and thus values far from 0 in this direction are for occupations with incomes that are unusually high or low given their levels of education.

- The added-variable plot for education (b) shows that the same three observations have relatively high leverage on the education coefficient:
  - *ministers* and *railroad conductors* tend to increase the education slope;
  - *railroad engineers* appear to be closer in line with the rest of the data.

▶ Deleting *ministers* and *conductors* produces the fitted regression

$$\widehat{\text{Prestige}} = \underset{(3.65)}{-6.41} + \underset{(0.122)}{0.867} \times \text{Income} + \underset{(0.099)}{0.332} \times \text{Education}$$

$$R^2 = 0.88 \quad S_E = 11.4$$

which has a larger income slope and smaller education slope than the original regression.

- The estimated standard errors are likely optimistic, because relative outliers have been trimmed away.

- Deleting *railroad engineers*, along with *ministers* and *conductors*, further increases the income slope and decreases the education slope, but the change is not dramatic: $B_{\text{Income}} = 0.931, B_{\text{Education}} = 0.285$.

9. Should Unusual Data Be Discarded?

- 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., as in Davis's regression) can be corrected or thrown away.
  - When a discrepant data-point is correct, we may be able to understand why the observation is unusual.
    - For Duncan's regression, for example, it makes sense that ministers enjoy prestige not accounted for by the income and educational levels of the occupation.
    - In a case like this, we may choose to deal separately with an outlying observation.

- Outliers or influential data may motivate model respecification.
  - For example, the pattern of outlying data may suggest the introduction of additional explanatory variables.
    - If, in Duncan's regression, we can identify a variable that produces the unusually high prestige of ministers (net of their income and education), and if we can measure that variable for other observations, then the variable could be added to the regression.
  - In some instances, transformation of the response variable or of an explanatory variable may draw apparent outliers towards the rest of the data, by rendering the error distribution more symmetric or by eliminating nonlinearity.
  - We must, however, be careful to avoid 'over-fitting' the data — permitting a small portion of the data to determine the form of the model.

- Except in clear-cut cases, we are justifiably reluctant to delete observations or to respecify the model to accommodate unusual data.
  - Some researchers reasonably adopt alternative estimation strategies, such as robust regression, which continuously downweights outlying data rather than simply including or 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.

10. Summary

- Unusual data are problematic in linear models fit by least squares because they can substantially influence the results of the analysis, and because they may indicate that the model fails to capture important features of the data.
- Observations with unusual combinations of explanatory-variables values have high *leverage* in a least-squares regression. The hat-values h_i provide a measure of leverage. A rough cutoff for noteworthy hat-values is $h_i > 2\overline{h} = 2(k+1)/n$.

- A regression *outlier* is an observation with an unusual response-variable value given its combination of explanatory-variable values. The studentized residuals E_i^* can be used to identify outliers, through graphical examination or a Bonferroni test for the largest absolute E_i^*. If the model is correct (and there are no true outliers), then each studentized residual follows a t-distribution with $n-k-2$ degrees of freedom.
- Observations that combine high leverage with a large studentized residual exert substantial *influence* on the regression coefficients. Cook's D-statistic provides a summary index of influence on the coefficients. A rough cutoff is $D_i > 4/(n-k-1)$.
- Subsets of observations can be jointly influential. Added-variable plots are useful for detecting joint influence on the regression coefficients. The added-variable plot for the regressor X_j is formed using the residuals from the least-squares regressions of X_j and Y on all of the other X's.

- Outlying and influential data should not be ignored, but they also should not simply be deleted without investigation. 'Bad' data can often be corrected. 'Good' observations that are unusual may provide insight into the structure of the data, and may motivate respecification of the statistical model used to summarize the data.