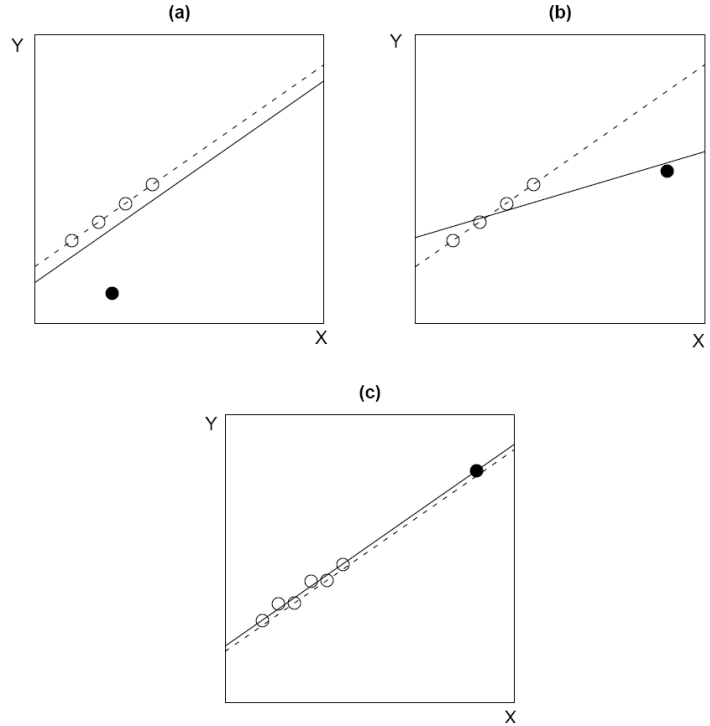
Chapter 11: Unusual and Influential Data

11.1 Outliers, Leverage, and Influence

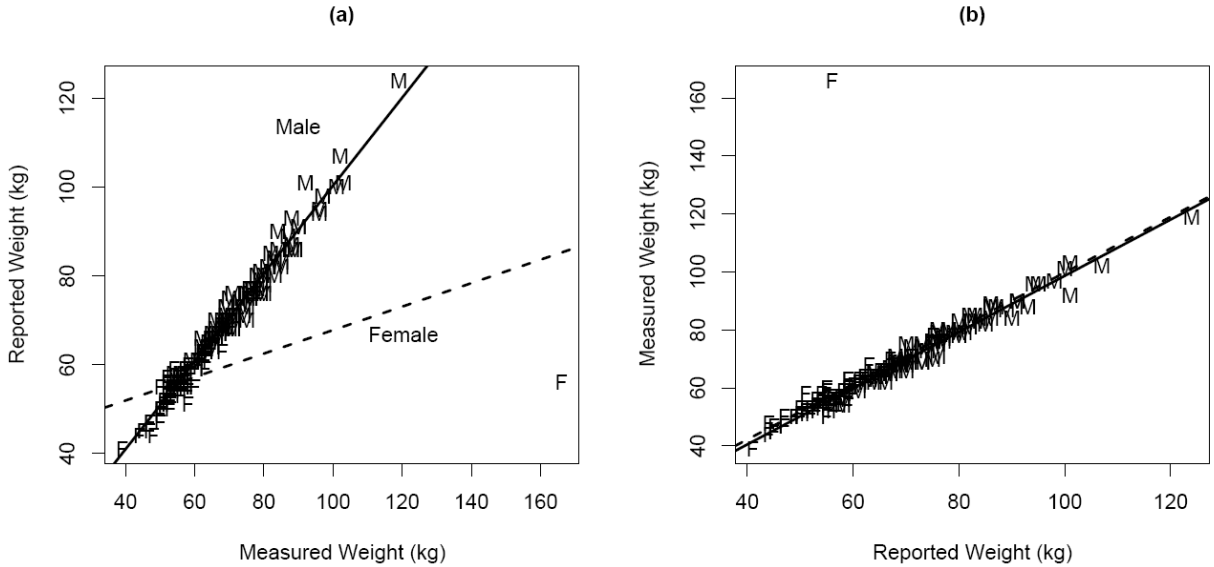
Figure 11.1



In regression, and *outlier* is an observation whose response variable value is *conditionally unusual* given the values of the explanatory variables. (a) point with *low leverage* and *little influence* on regression; (b) point with *high leverage* and *high influence*; (c) point with *high leverage* but *low influence*.

Influence on coefficients = Leverage × Discrepancy

Figure 11.2



11.2 Assessing Leverage: the hat values

Recall the *Hat Matrix*:

- The *Hat Matrix*: $H = X(X^tX)^{-1}X^t$
- It's a projection matrix: $\hat{Y} = X\hat{\beta} = X(X^tX)^{-1}X^tY = HY$
- So, it is idempotent ($HH = H$) and symmetric ($H^t = H$)
- And, $E = Y - \hat{Y} = Y - HY = (I - H)Y$, where $(I - H)$ is also a projection matrix

$\hat{Y}_j = \sum_{i=1}^{n} h_{ij} Y_i$, so h_{ij} tells us the contribution of the i^{th} observation to the j^{th} fitted value. Because $H = HH$, the diagonal elements of *H* are $h_i \equiv h_{ii} = \mathbf{h}_i' \mathbf{h}_i = \sum_{j=1}^{n} h_{ij}^2$, which summarizes the potential influence or *leverage* of Y_i on all the fitted values. These *hat-values* satisfy $(1/n \leq h_i \leq 1)$ and the average hat-value is $\bar{h} = (k+1)/n$ where k is the number of explanatory variables (excluding the intercept).

[Note: the upper bound on the h_i is actually $1/c_i$, in general, where c_i is the number of times that the ith row of *X*, $\boldsymbol{x}_i$, is replicated.]

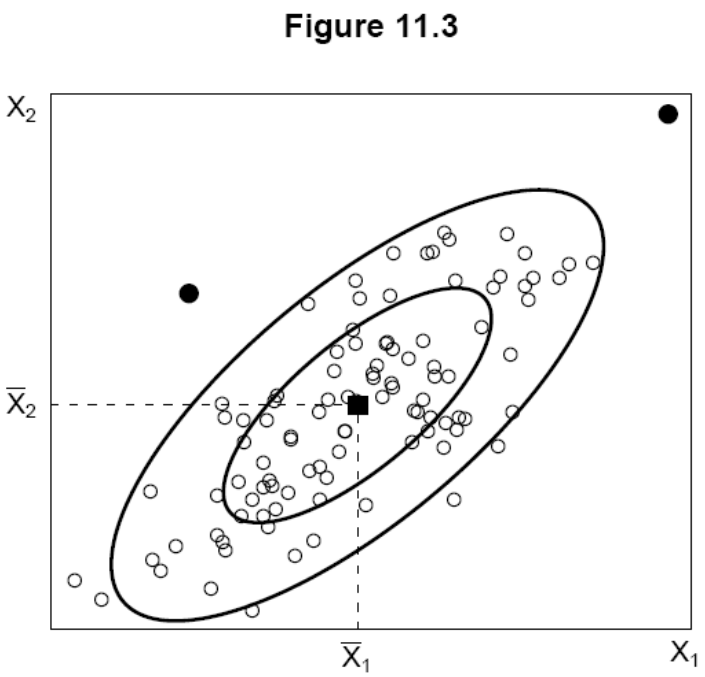
For simple linear regression it is easy to show that

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

Using the multiple regression model in matrix notation in mean-centered form: $\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ where $\mathbf{y}^* \equiv \left\{ Y_i - \bar{Y} \right\}$ and $\mathbf{X}^* \equiv \left\{ X_{ij} - \bar{X}_j \right\}$ and $\boldsymbol{\beta}_1$ is the vector of regression coefficients without the intercept, the hat-value for the i^{th} observation is

$$h_i^* = \mathbf{h}_i^{*'} \mathbf{h}_i^* = \mathbf{x}_i^{*'} (\mathbf{X}^{*'} \mathbf{X}^*) \mathbf{x}_i^* = h_i - \frac{1}{n}$$

Figure 11.3 Elliptical contours of constant leverage (constant hat-values h_i) for k=2 explanatory variables.

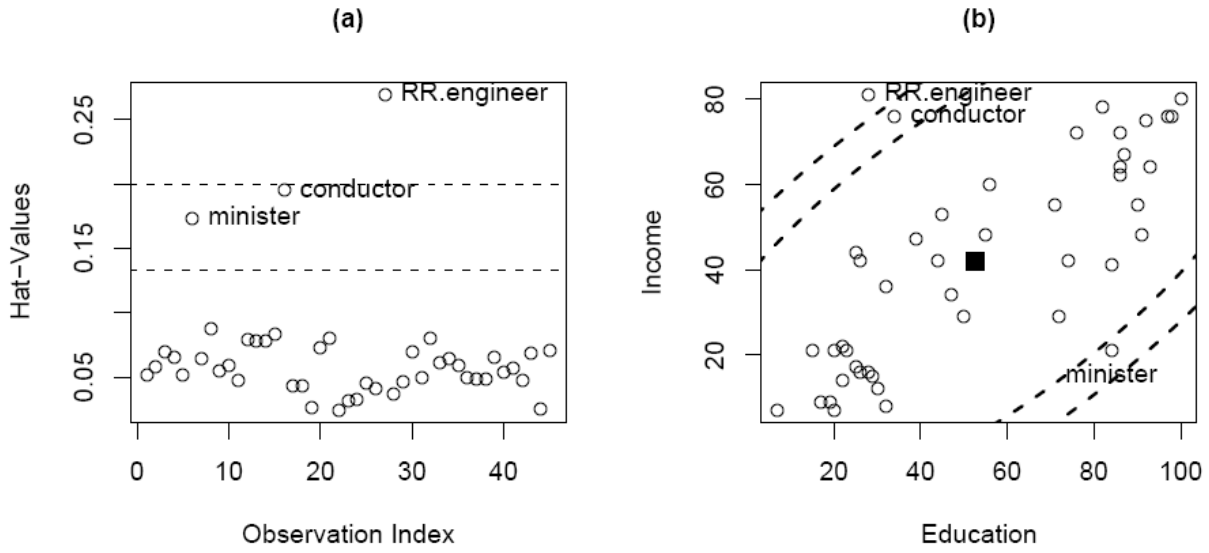


Note the differences between the two high leverage points:

- one is has the highest value on both X_1 and X_2, although it is not too extreme on either
- the other is not extreme on either X_1 or X_2

So, leverage cannot be judged by examining simple histograms. In fact you might not be able to see influential points even in scatterplots of two variables at a time.

Figure 11.4 For the regression of Duncan's prestige data on education and income: (a) index plot of hat-values, (b) contours of constant leverage, with reference lines at $2\times\overline{h}$ and $3\times\overline{h}$. In (b) the dashed lines are part of elliptical contours.



11.3 Detecting Outliers: "Studentized" residuals?

Another use of the hat matrix.

- $E = Y - \hat{Y} = Y - HY = (I - H)Y$, where $(I - H)$ is also a projection matrix, so
- $\text{var}(E) = \text{var}((I - H)Y) = (I - H)\,\text{var}(Y)(I - H)^T = (I - H)\sigma^2$
- *Standardized Residuals*:

$$E_i' = \frac{E_i}{s\sqrt{1 - h_{ii}}},$$

so we see that the residuals do not have equal variances even though we assume that the true errors ε_i do have equal variances. Standardized residuals are useful, but the numerator is not independent of the denominator, so E_i' cannot follow a t-distribution, which would be nice to judge the magnitude of a residual

- *Studentized Residuals*:

$$E_i^* = \frac{E_i}{s_{-i}\sqrt{1-h_{ii}}}$$

Text provides an equation showing that $\left(E_i^*\right)^2$ is a monotonic transformation of $\left(E_i^{'}\right)^2$.

Text also notes that these studentized residuals can be derived from a model with a mean-shift for the ith observation and $E_i^* \sim t_{n-k-2}$ which we can use for testing the significance of a particular residual, but we must correct for multiple testing---picking the biggest residual to test---using a Bonferroni correction. I.e. test with an α-level of, for example, .05/*k*, or equivalently report the p-value as *k*×(the usual p-value).
Note that although these studentized residuals are based on a fit to all the data; they are related to the

- *Leave-one-out Residuals*:

$$E_{(-i)} = y_i - \mathbf{x}_i^{'}\hat{\boldsymbol{\beta}}_{(-i)} = \frac{E_i}{1-h_{ii}}$$

So the (externally) studentized residuals also represent the "studentization" of these leave-one-out residuals that go into the predictive error sum of squares (PRESS).

$$E_i^* = \frac{E_{(-i)}}{s_{-i} / (1-h_{ii})^{1/2}}$$

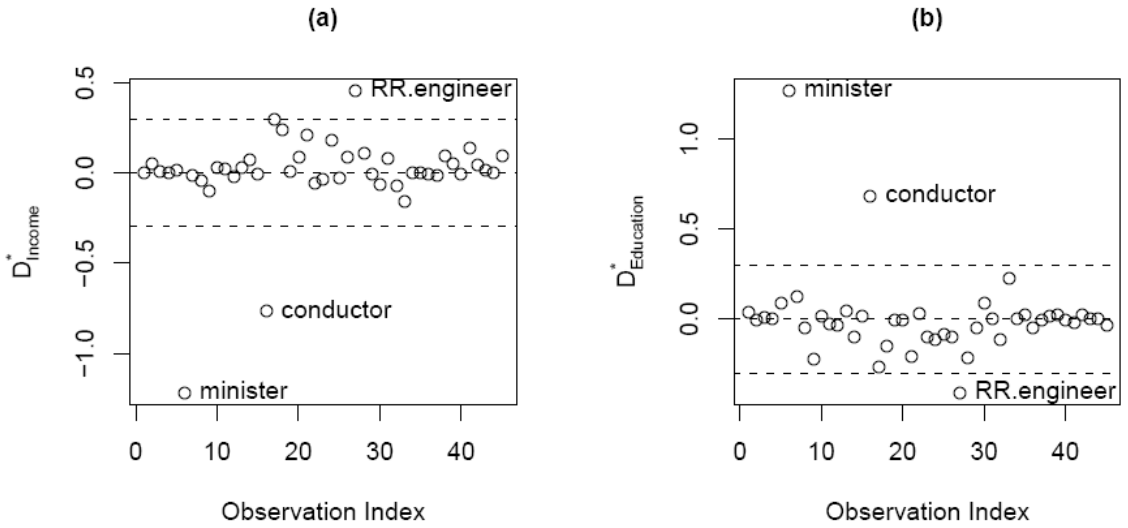
11.4 Measuring Influence

$$D_{ij} = DFBETA_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)},\ i = 1,\ldots,n \text{ and } j = 1,\ldots,k.$$

$$D^*_{ij} = DFBETAS_{ij} = \frac{D_{ij}}{SE_{(-i)}(\beta_j)}$$

See sect 11.5 for rule-of-thumb cutoffs of $\pm 2/\sqrt{n}$

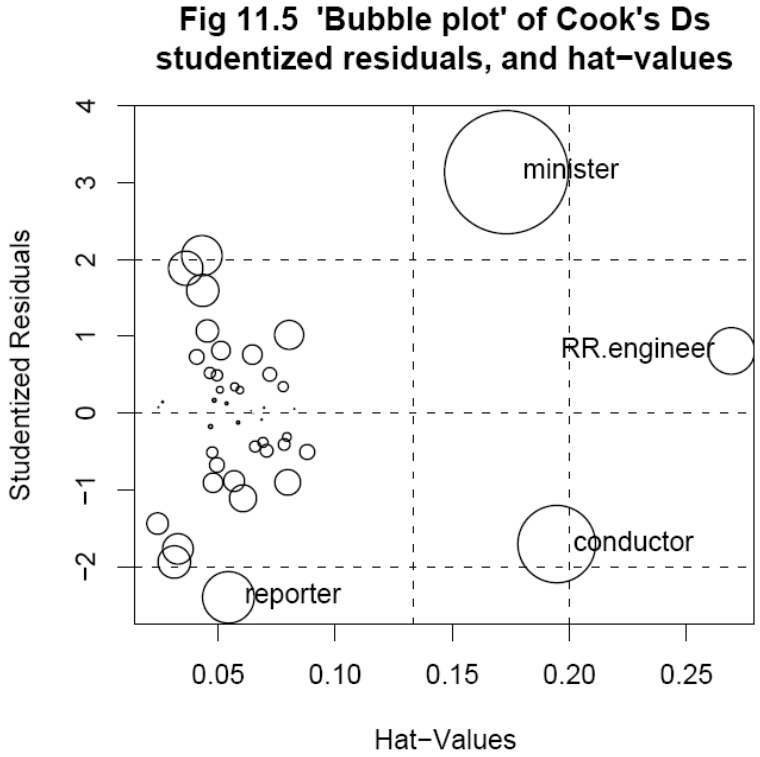
Fig 11.6 Index plots of D^*_{ij}



Cook's measure of "distance" derives from the form of an F-statistic comparing $\hat{\beta}$ with $\hat{\beta}_{(-i)}$, or the evaluation of $\hat{\beta}_{(-i)}$ with respect to confidence ellipsoids for $\hat{\beta}$:

$$
\begin{aligned}
D_i &= \frac{(\hat{\beta}_{(-i)} - \hat{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\beta}_{(-i)} - \hat{\beta})}{(k+1) S_E^2} \\
&= \frac{(\hat{y}_{(-i)} - \hat{y})^T (\hat{y}_{(-i)} - \hat{y})}{(k+1) S_E^2} \\
&= \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}
\end{aligned}
$$

So we see that high "influence" according to Cook's D comes from a combination of a large residual with high leverage. Belsley, Kuh and Welsh proposed a similar measure defined in terms of the studentized residuals $DFFITS_i = E_i^* \Big/ \sqrt{h_i/(1-h_i)}$

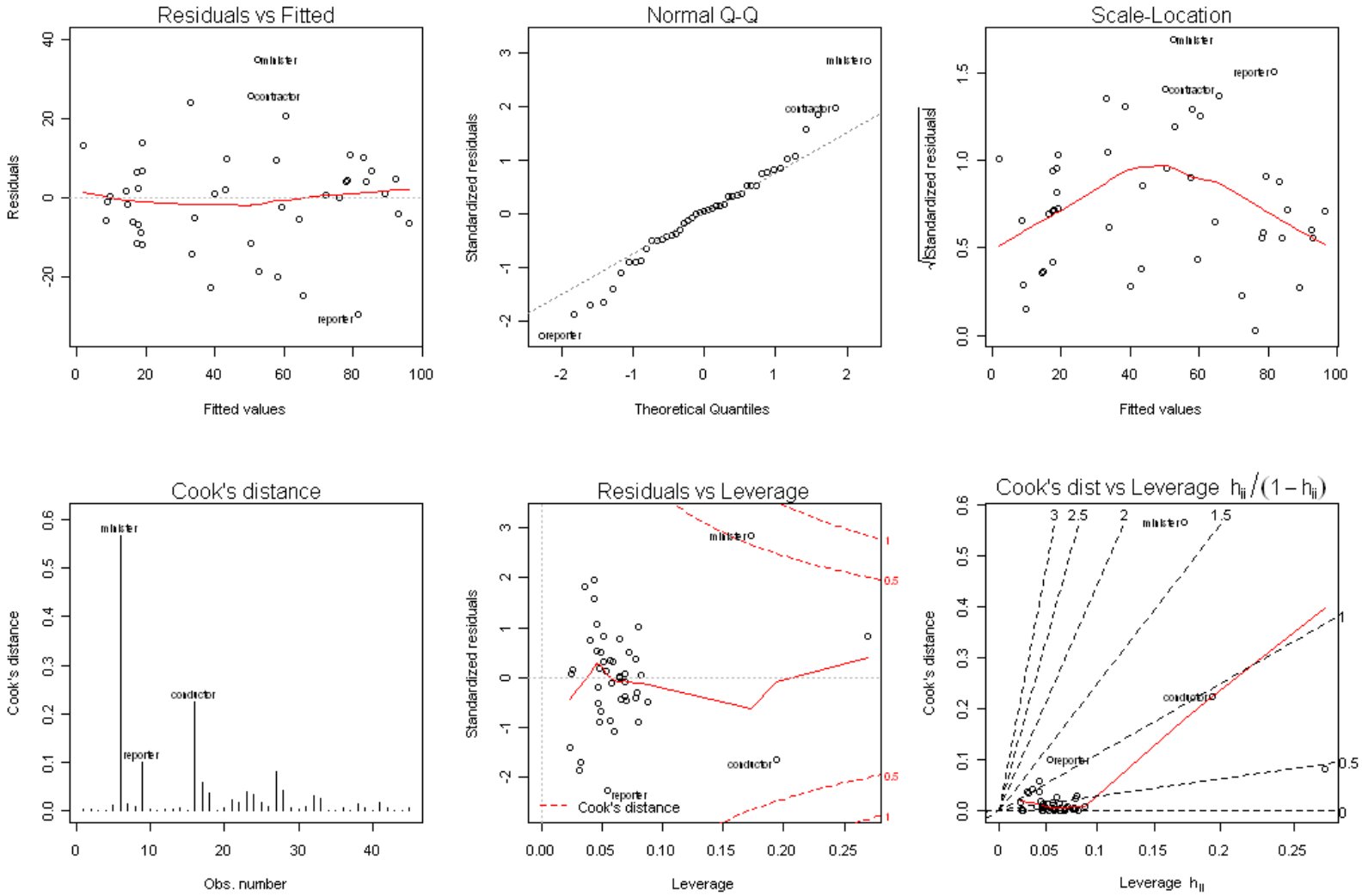


[*See R code*]

There are also measures called "COVRATIO" to measure influence on collinearity, but we won't bother with these.

See section 11.5.3 for discussion of suggested cutoffs for paying attention to large values of influence measures. One for Cook's D, based on the analogy with F-statistics, is $D_i > 4/(n-k-1)$.

```
## Standard R diagnostic plots
windows()
par(mfrow=c(2,3))
plot(duncan.mod,which=1:6)
# All the diagnostic plots; default: which=c(1:3,5)
# Contours of standardized residuals drawn on 6th plot.
```



11.5 Joint Influence

All of these influence measures refer to the influence of individual points. It is much more complicated when there are groups of influential points (where the influence of one point in a group is masked by the others). Generalizations to multiple influential points are possible but not very practical. At this point Fox introduces "Added Variable Plots" for graphical assessment (which I would normally have introduced much earlier).

The idea: we can graphically assess the influence of extreme observations in 2D scatters. Look at 2D scatters that represent the computation of partial regression coefficients.

In Fox's notation (not my preference):

$Y_i^{(1)} = \text{Residual}(Y_i | X_2, \ldots, X_k)$, residuals of the regression on all but X_1.

$X_i^{(1)} = \text{Residual}(X_{i1} | X_2, \ldots, X_k)$, residuals of regression of X_1 on the rest

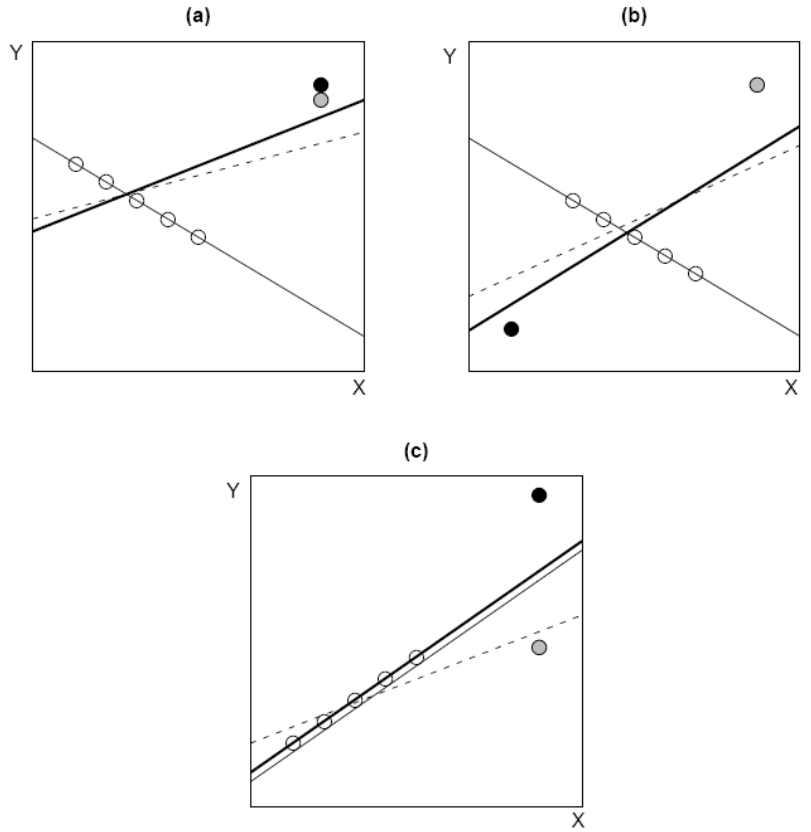


Fig 11.7 Jointly influential data in simple linear regression.

The scatterplot and regression of $Y_i^{(1)}$ on $X_i^{(1)}$ provides a complete understanding of the partial regression coefficient $\hat{\beta}_1$.

1. Slope of regression of $Y_i^{(1)}$ on $X_i^{(1)}$ (without intercept) is $\hat{\beta}_1$
2. Residuals of regression of $Y_i^{(1)}$ on $X_i^{(1)}$ are the same as the residuals from the full multiple regression
3. Standard error of $\hat{\beta}_1$ from this simple ("auxiliary") regression, $SE(\hat{\beta}_1) = S_E \Big/ \sqrt{\sum X_i^{(1)2}}$, is the same as the std error of $\hat{\beta}_1$ in the full multiple regression.

So, we can examine these *added variable* (or *partial-regression*) plots to assess influence as we would in examining Fig 11.7.

Fig 11.8

[*See R code*]