14 Residuals and Influence

Let $\mathbf{X}$ be a $n \times \tilde{p}$ design matrix of full rank ($\tilde{p} = p + 1$ if we have an intercept, and $\tilde{p} = p$ otherwise), and let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\operatorname{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. Define $\mathbf{H} = \mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$.

14.1 Theorem: Under the above assumptions, $E[\hat{\boldsymbol{\varepsilon}}] = \mathbf{0}$ and $\operatorname{var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$. In particular, the variance for residual i is $\operatorname{var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$, where $h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$.

14.2 Note: Since $\operatorname{rank}(\mathbf{H}) = \operatorname{trace}(\mathbf{H})$, it follows that $\sum h_{ii} = \tilde{p}$.

14.3 Note: If we assume there is an intercept in the model, then $\mathbf{H}\mathbf{1}_n = \mathbf{1}_n$, and hence $\sum_i h_{ij} = \sum_j h_{ij} = 1$.

14.4 Theorem: Assume we have an intercept in the model. Let $\mathcal{X}$ be the $n \times p$ matrix of the original data (without the intercept) with the column averages substracted off. Let $\mathcal{X}'\mathcal{X}$ be the corrected cross-product matrix, i. e. $(\mathcal{X}'\mathcal{X})_{jk} = \sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$, and redefine $\mathbf{x}_i'$ to be the ith row of $\mathbf{X}$ without the one for the intercept. Then the following holds:

(a) $h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathcal{X}'\mathcal{X})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$. This means in particular that each h_{ii} is bounded below by $\frac{1}{n}$.

(b) Let r_i be the number of rows of $\mathbf{X}$ that are identical to its ith row. Then $h_{ii} \leq \frac{1}{r_i}$.

14.5 Example: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Then $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}$, and $h_{ii} = \frac{1}{n}$ iff $x_i = \bar{x}$.

14.6 Definition: The internally studentized residuals are defined as

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}},$$

where $\hat{\sigma}$ is the estimated standard error including case i.

14.7 Definition: The externally studentized residuals are defined as

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}},$$

where $\hat{\sigma}_{(i)}$ is the estimated standard error excluding case i.

14.8 Theorem: The internally and externally studentized residuals are monotonically related through

$$t_i = r_i\sqrt{\frac{n - \tilde{p} - 1}{n - \tilde{p} - r_i^2}}.$$

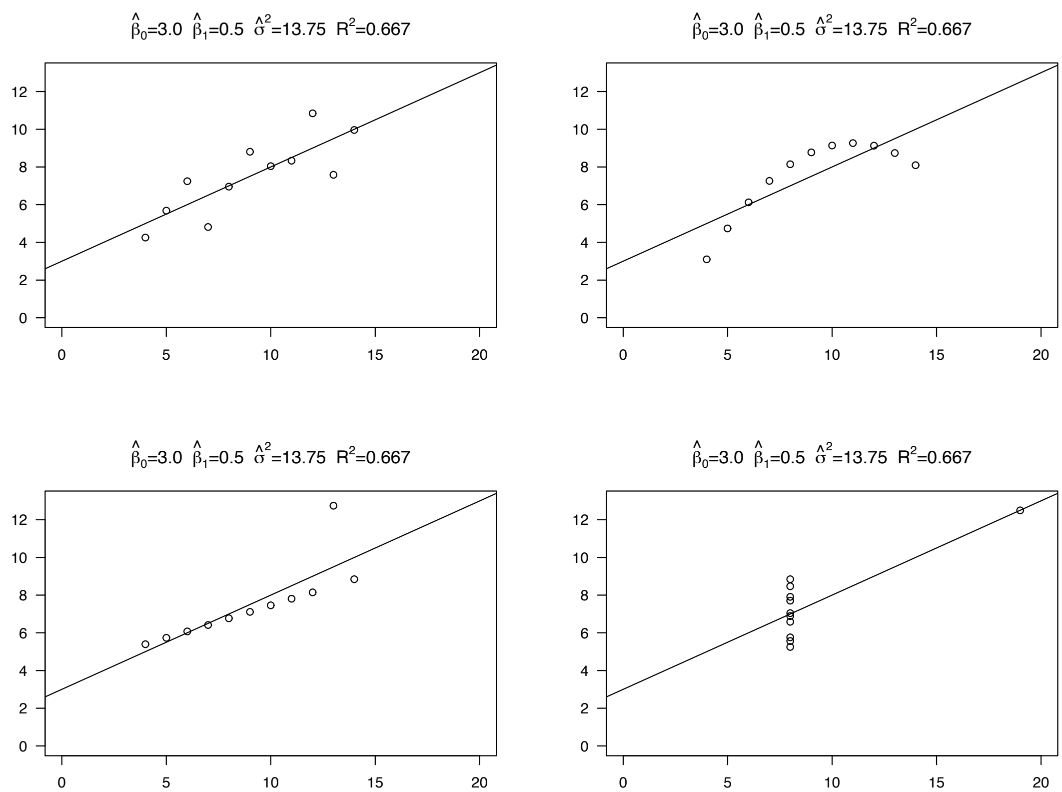
14.9 Definition: Cook's distance is defined by

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{\tilde{p}\hat{\sigma}^2} = \frac{(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})}{\tilde{p}\hat{\sigma}^2} = \frac{1}{\tilde{p}} \times r_i^2 \times \left(\frac{h_{ii}}{1 - h_{ii}}\right),$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ are the parameter estimates obtained after deleting observation i, and $\hat{\mathbf{Y}}_{(i)}$ are the corresponding fitted values.

14.10 Theorem: An alternative expression for the Cook's distance is $D_i = \frac{1}{\tilde{p}} \times r_i^2 \times \left(\frac{h_{ii}}{1-h_{ii}}\right)$.

14.11 Note: The importance of plotting the data and checking model assumptions is illustrated below.



Anscombe, 1973.