Model Diagnostics

Johns Hopkins Engineering

625.461 Statistical Models and Regression

Module 8 – Lecture 8E

Leverage Point and Influential Point

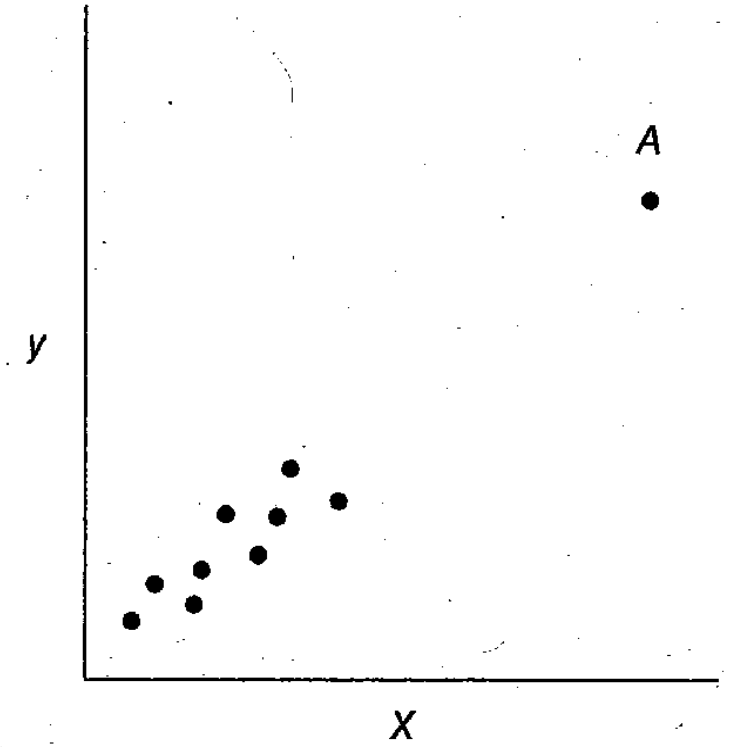

Figure 6.1 An example of a leverage point.

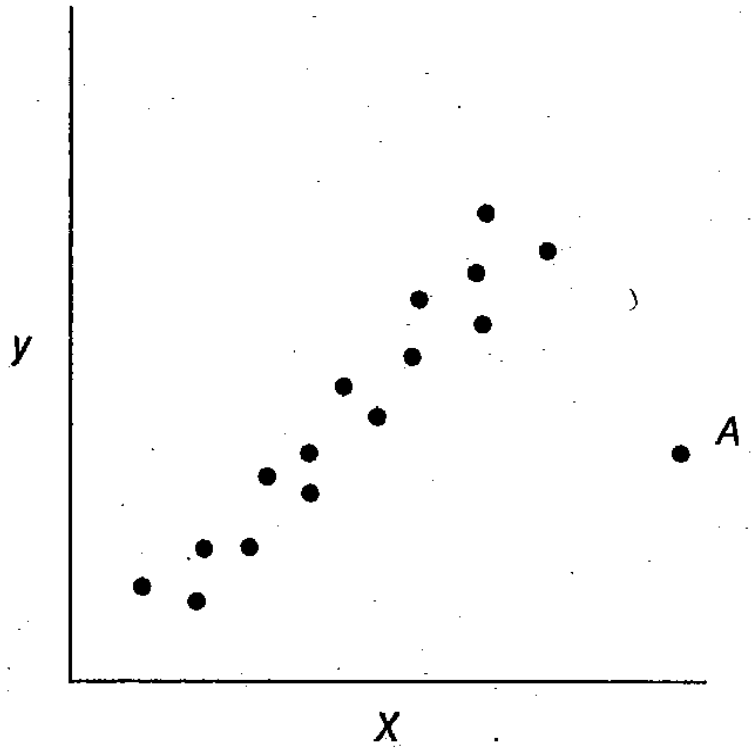

Figure 6.2 An example of an influential observation.

Leverage Point and Influential Point

Leverage point: it has an unusual x value and may control certain model properties

Influence point: it has a noticeable impact on the model coefficients in that it "pulls" the regression model in its direction

Leverage Point and Influential Point

Remote points potentially have disproportionate impact on the parameter estimates, standard errors, predicted values, and model summary statistics.

The hat matrix plays an important role in identifying influential observations.

Leverage Point and Influential Point

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \qquad \hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

$$\text{Var}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$$

$$\text{Var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

The elements h_{ij} of the matrix $\mathbf{H}$ may be interpreted as the amount of leverage exerted by the ith observation y_i on the jth fitted value $\hat{y}_j$

Leverage Point and Influential Point

The diagonal elements h_{ii} of $\mathbf{H}$: $h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$

is a standardized measure of the distance of the ith observation from the center (or centroid) of the x space.

Leverage Point and Influential Point

Large hat diagonals reveal observations that are potentially influential because they are remote in x space from the rest of the sample. Since

$$\sum_{i=1}^{n} h_{ii} = \operatorname{rank}(\mathbf{H}) = \operatorname{rank}(\mathbf{X}) = p$$

the average size of a hat diagonal is p / n.
Any observation with the hat diagonal exceeding $2p / n$ may be remote enough to be considered a leverage point.

Leverage Point and Influential Point

Not all leverage points are going to be influential on regression coefficients (see Figure 6.1). Because the hat diagonals examine only the location of the observations in x space, some analysts like to look at the studentized residuals or R-student in conjunction with the h_{ii}.

Leverage Point and Influential Point

In general, observations with large hat diagonals and large residuals are likely to be influential.

Note that if $2p/n > 1$, then this cut-off does not apply

Delivery Time Data (Ex 6.1, page 213 of Textbook)

$p = 3$ and $n = 25$. $2p/n = 0.24$

Observations 9 and 22 are “leverage” points

Table 4.1 (p.137) containing the studentized residuals and R-student. These residuals are not unusually large for obs 22, indicating it likely has little influence on the fitted model.

Delivery Time Data (Ex 6.1, page 213 of Textbook)

For obs 9, both scaled residuals are moderately large, suggesting this obs may have moderate influence on the model.

Run	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	MS_{Res}	R^2
9 and 22 in	2.341	1.616	0.014	10.624	0.9596
9 out	4.447	1.498	0.010	5.905	0.9487
22 out	1.916	1.786	0.012	10.066	0.9564
9 and 22 out	4.643	1.456	0.011	6.163	0.9072

Deleting obs 9 results in approximately a 28% change in $\hat{\beta}_2$ and a 90% change in $\hat{\beta}_0$, also dramatically affects MS_{Res}.

A Measure of Influence: Cook's D

Use a measure of the squared distance between the LS estimate $\hat{\boldsymbol{\beta}}$ based on all n points and the LS estimate obtained by deleting the ith point, say labeled $\hat{\boldsymbol{\beta}}_{(i)}$

A Measure of Influence: Cook's D

$$D_i(\mathbf{M}, c) = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)' \mathbf{M} \left(\boldsymbol{\beta}_{(i)} - \hat{\boldsymbol{\beta}}\right)}{c}, \quad i = 1, 2, \ldots, n$$

The usual choice is

$$D_i = \frac{(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})'(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})}{pMS_{\text{Res}}}$$

$$D_i(\mathbf{X}'\mathbf{X}, pMS_{\text{Res}}) \equiv D_i = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{X} \left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)}{pMS_{\text{Res}}}, \quad i = 1, 2, \ldots, n$$

Points with large values of D_i have considerable influence on the LS estimate $\hat{\boldsymbol{\beta}}$

A Measure of Influence: Cook's D

Usually consider points with $D_i > 1 \cong F_{0.5,\, p,\, n\text{-}p}$ to be influential

$$D_i = \frac{r_i^2}{p}\frac{\mathrm{Var}(\hat{y}_i)}{\mathrm{Var}(e_i)} = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}, \quad i = 1, 2, \ldots, n$$

Thus, D_i is made up of a component that reflects how well the model fits the ith obs y_i and a component that measures how far that point is from the rest of data.

Delivery Time Data (Ex 6.2, page 216 of Textbook)

For observation 9, Table 6.1 (p.214) shows that $D_9 = 3.41835$.

$$\Pr(D_9 > 3.41835)$$
$$= \Pr(F_{3,22} > 3.41835)$$
$$= 0.035$$

This observation is likely to be an influential value.

JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING