MAT 3375
Regression Analysis

Chapter 6
Outliers and Influential Observations

P. Boily (uOttawa)

Summer – 2023

Outline

6 – Outliers and Influential Observations

When we are working with a single predictor, we can usually tell quite quickly if a prediction or a response is unusual, in some sense.

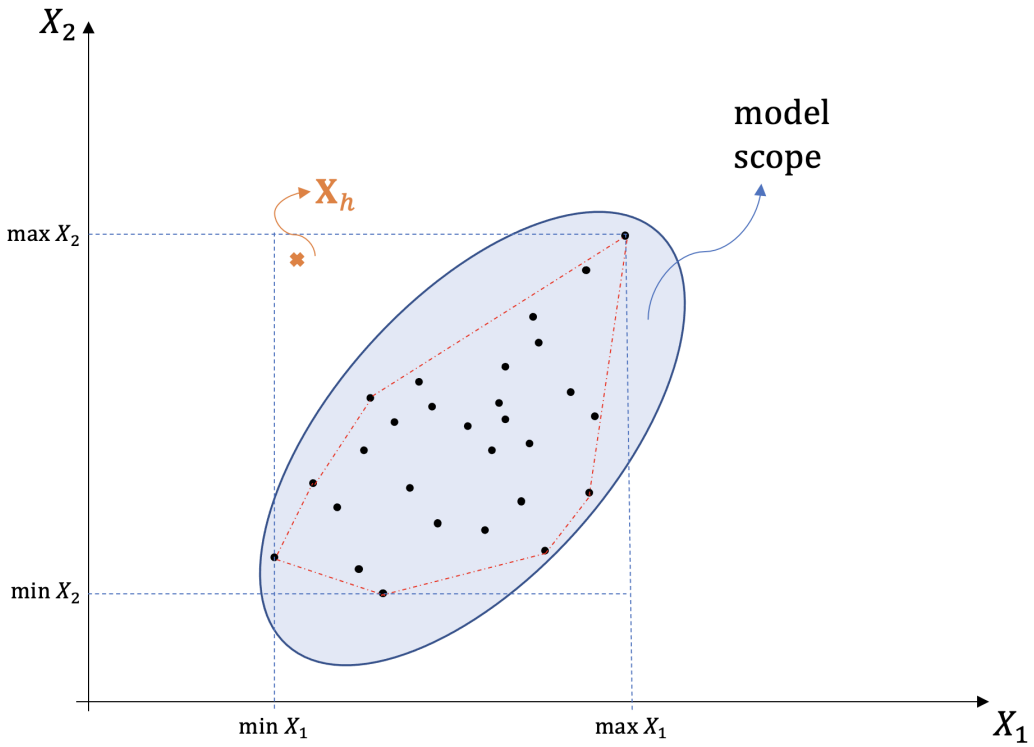
If a predictor value is much smaller/much larger than the other predictor values, we might be hesitant to use the regression model to fit the value because no similar values were used to "train" the model.

When $p > 1$, finding the anomalous observations (predictors and/or responses) is not as obvious.

In this chapter, we introduce a small number of methods to do so (there are a lot more, see DUDADS).

6.1 – Leverage and Hidden Extrapolation

Consider a dataset with two predictors X_1, X_2, as shown below.



Regression models are typically only useful when we are working within the **model scope**; if regression is an attempt to **interpolate** the data, then we must avoid situations where we are **extrapolating** from the data.

The problem is that we cannot always easily tell if a predictor $\mathbf{X}_h$ is in the scope or not; in the previous image, each component of $\mathbf{X}_h$ is in the range of the predictors used to build the model, but $\mathbf{X}_h$ as a whole is **not**. When p is large, this **visual** approach fails.

The **leverage of the ith case** is:

$$h_{ii} = \mathbf{X}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i, \quad \mathbf{X}_i \text{ is the } i\text{th row of } \mathbf{X};$$

in other words, h_{ii} is the ith diagonal element of $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$.

The **leverage** determines if a predictor level $\mathbf{X}_h$ is in the **model scope**: if

$$\mathbf{X}_h^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_h > \max\{h_{ii} \mid i = 1, \ldots, n\},$$

$\mathbf{X}_h$ is **outside the scope** and $\hat{Y}_h = \mathbf{X}_h\mathbf{b}$ contains a **hidden extrapolation**.

Note that $0 \leq h_{ii} \leq 1$, for $i = 1, \ldots, n$. Indeed, since:

1. $\mathbf{0} \leq \sigma^2\{\hat{\mathbf{Y}}\} = \sigma^2\{\mathbf{HY}\} = \mathbf{H}\sigma^2\{\mathbf{Y}\}\mathbf{H}^\top = \sigma^2\mathbf{H} \implies h_{ii} \geq 0$ for all i

2. $\mathbf{0} \leq \sigma^2\{\mathbf{e}\} = \sigma^2\{(\mathbf{I}_n - \mathbf{H})\mathbf{Y}\} = \sigma^2(\mathbf{I}_n - \mathbf{H}) \implies 1 - h_{ii} \geq 0$ for all i

Generally-speaking, the surface of $\mathbf{X}_h^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_h = c$ is an ellipsoid centred around $\overline{\mathbf{X}} = (1, \overline{X}_1, \ldots, \overline{X}_p)$ (the larger c, the larger the "distance" to $\overline{\mathbf{X}}$).

An $X-$**outlier** is an observation which is **atypical** with respect to the **predictor levels**.

We note that

$$\overline{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{1}{n}\,\mathrm{tr}(\mathbf{H}) = \frac{p}{n} \quad (p \leq n);$$

1. if $h_{ii} \leq 0.2$, then the leverage of the ith case is **low** (very near $\overline{\mathbf{X}}$);
2. if $0.2 < h_{ii} < 0.5$, then the leverage is **moderate**;
3. if $h_{ii} \geq 0.5$, then the leverage is **high** (potential $X-$**outlier**);
4. when n is large, if $h_{ii} > 3\overline{h} = \frac{3p}{n}$, then the ith case is an $X-$**outlier**.

Example: we wish to fit the multiple linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

to a dataset with n observations, with

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{pmatrix} 1.17991 & -0.00731 & 0.00073 \\ -0.00731 & 0.00008 & -0.00012 \\ 0.00073 & -0.00012 & 0.00046 \end{pmatrix} \quad \text{and} \quad \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} 220 \\ 36768 \\ 9965 \end{pmatrix}$$

What are the point estimates for the regression coefficients $\boldsymbol{\beta}$? We would like to predict the value of Y_h when $X_1 = 200$ and $X_2 = 50$, i.e., at the point $\mathbf{X}_h = (1, 200, 50)^\top$. What is the leverage of $\mathbf{X}_h$? Is this case of hidden extrapolation? If not, what is the predicted value Y_h?

Solution: the LS estimates of the regression coefficients are

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} -1.91943 \\ 0.13744 \\ 0.33234 \end{pmatrix}.$$

The leverage of $\mathbf{X}_h$ is

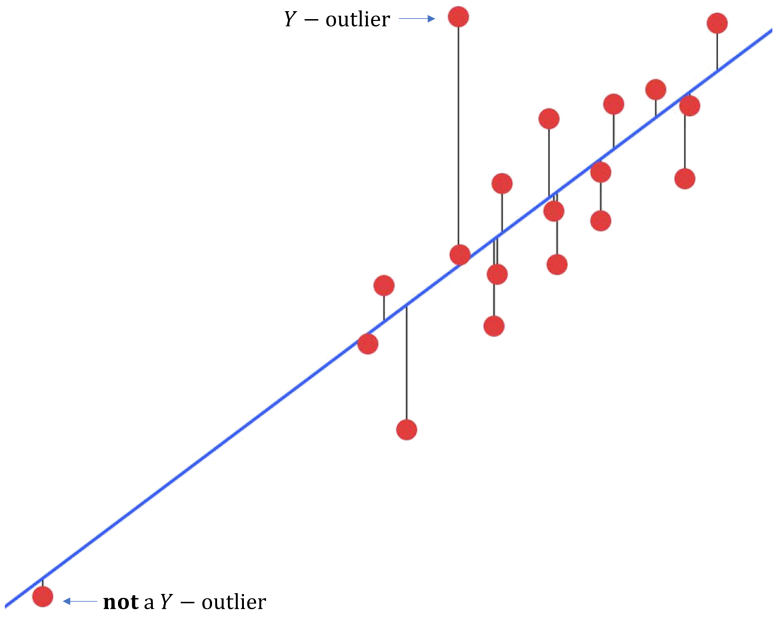
$$\mathbf{X}_h^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}_h = 0.27891;$$

it is small enough to suggest that we are not in a hidden extrapolation situation (although n is unknown, so we cannot compare it against $\frac{3p}{n}$).

The predicted response at $\mathbf{X}_h$ is thus $\hat{Y}_h = \mathbf{X}_h^\top \mathbf{b} = 42.18557$.

6.2 – Deleted Studentized Residuals

While $X-$outliers can be determined without reference to a **regression surface** $\hat{Y}(\mathbf{x}) = \mathbf{xb}$, we can also look for observations whose response values are **unexpectedly distant** from $\hat{Y}(\mathbf{x})$.



A $Y-$**outlier** is an observation which yields a **large** regression residual.

1. If the **(internal) studentized residual** is large enough,

$$|r_i| = \left|\frac{e_i}{\text{s}\{e_i\}}\right| = \left|\frac{e_i}{\sqrt{\text{MSE}}\sqrt{1-h_{ii}}}\right| \geq 3,$$

   say, then the ith point is a $Y-$outlier; or

2. Another approach: delete the ith case from the model and refit

$$\mathbf{b}_{(i)} = \left(\mathbf{X}_{(i)}^{\top}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^{\top}\mathbf{Y}_{(i)},$$

   yielding an expected value for the ith case, $\hat{Y}_{i(i)}$.

For $i = 1, \ldots, n$, the **deleted residual** is $d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1-h_{ii}}$ and the **external studentization** is

$$t_i = \frac{d_i}{\mathrm{s}\{d_i\}} = e_i \sqrt{\frac{n-p-1}{\mathrm{SSE}(1-h_{ii}) - e_i^2}} \sim t(n-p-1),$$

where

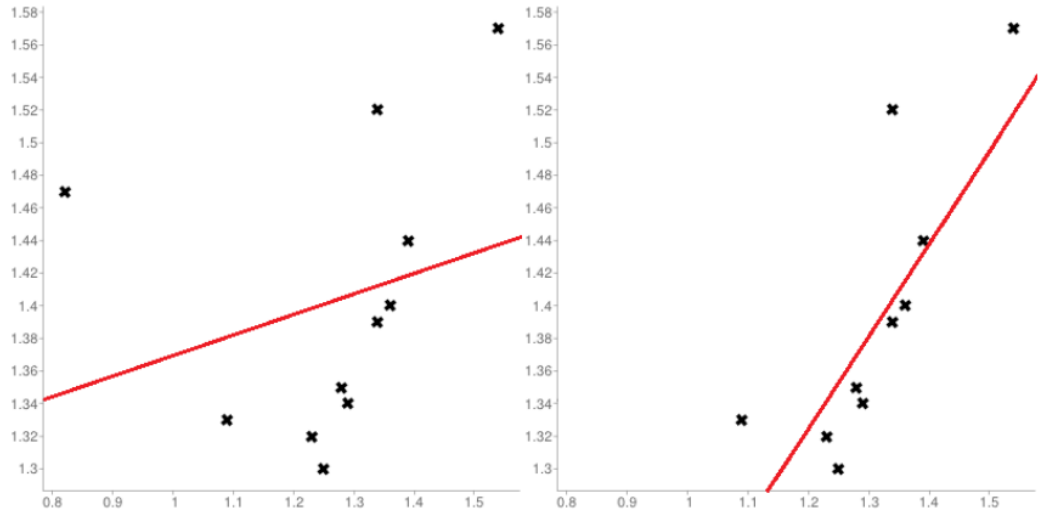
$$\mathrm{s}^2\{d_i\} = \mathrm{MSE}_{(i)} \left[1 + \mathbf{X}_i \left(\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)}\right)^{-1} \mathbf{X}_i^\top\right].$$

Decision Rule: if $|t_i| > t(1 - \frac{\alpha/n}{2}; n-p-1)$, then the ith case is a $Y-$outlier at significance level α.

Note that it is possible for an observation to be an $X-$outlier without being an $Y-$outlier, and *vice-versa* (see previous chart).

6.3 – Influential Observations

In the regression context, we may also be interested in determining which observations are **influential** – observations whose absence from (or presence in) the data significantly change the **nature of the fit** (qualitatively).



Influential observations need not be outliers (but may be!), and *vice-versa*.

For the ith case, DFFITS_i is a measure of the **influence** of the ith case on the $\hat{Y}$ in a neighbourhood of $\mathbf{X}_i$. The **difference from the fitted value** is

$$\mathrm{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\mathrm{MSE}_{(i)}\, h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}.$$

For small and moderately-sized samples, if $|\,\mathrm{DFFITS}_i\,| > 2$, then the ith case is **likely influential**. For larger samples, if $|\,\mathrm{DFFITS}_i\,| > 2\sqrt{\frac{p}{n}}$, then the ith case is **influential**.

A similar measure can be determined to see if case i has a lot of influence on the value of the **fitted parameter** b_k:

$$\mathrm{DFBETAS}_i^k = \frac{b_k - b_{k(i)}}{\sqrt{\mathrm{MSE}_{(i)}\,[(\mathbf{X}^\top\mathbf{X})^{-1}]_{k,k}}}.$$

6.4 – Cook's Distance

We can also use **Cook's distance** to measure observation i's influence:

$$D_i = \frac{1}{p \cdot \text{MSE}} \sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2 = \frac{e_i^2}{p \cdot \text{MSE}} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right] \sim F(p, n-p).$$

Decision Rule:

- if $D_i < F(0.2; p; n-p)$, then the ith case **has little influence**;
- if $D_i > F(0.5; p; n-p)$, then the ith case **is very influential**.

Regressions based on LS framework are convenient, but they are not **robust** against outliers and influential observations (median, absolute value).

Example: let

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 4 \\ 1 & 2 & 5 \\ 1 & 3 & 3 \\ 1 & 4 & 3 \\ 1 & 4 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} 2.1 \\ 24.2 \\ 29.5 \\ 27.6 \\ 30.5 \\ 27.5 \end{pmatrix}.$$

Find the data's $X-$outliers, $Y-$outliers, and influential observations.

Solution: since $n = 6$, the sample is small. The LS estimates are

$$\mathbf{b} = \begin{pmatrix} -7.3 \\ 5.51 \\ 5.70 \end{pmatrix},$$

from which

$$\mathbf{e} = \mathbf{Y} - \mathbf{Xb} = (-1.8, 3.2, -2.7, 1.28, -1.32, 1.37)^\top.$$

The external residuals are $(-18.47, 2.40, -1.99, 0.41, -0.5, 0.57)^\top$. Since

$$t\left(1 - \frac{\alpha/n}{2}; n - p - 1\right) = t\left(1 - \frac{0.1/6}{2}; 6 - 3 - 1\right) = 7.65,$$

only the first case is a $Y-$outlier at $\alpha = 0.1$; conservatively, when $|t_i|$ is large, we should further study the influence of case i, so we will be sure to look into **case** 1 in detail.

(Note the Bonferroni correction term).

For $X-$outliers, we seek cases with leverages above 0.5. The leverages are

$$\mathbf{h} = (0.87, 0.45, 0.58, 0.19, 0.41, 0.48)^\top.$$

Cases 1,3 are **high** leverage points, suggesting that they are potential $X-$outliers, whereas cases 2,5,6 have **moderate** leverages (but are unlikely to be $X-$outliers, lest 5/6 observations be so).

The **differences in fitted values** are

$$\text{DFFITS} = (-48.7, 2.29, -2.33, 0.2, -0.42, 0.54)^\top,$$

suggesting that only the first 3 cases are influential. The **Cook distances** are $\mathbf{D} = (6.9, 0.67, 0.91, 0.02, 0.08, 0.13)^\top$**;** since D_1 is the only distance larger than than $F(0.5; p, n-p) = 1$, only the **first** case is likely to be influential.