# MA 575 Linear Models:

Cedric E. Ginestet, Boston University

*Detecting Outliers and Influence Analysis*

Week 12, Lecture 1

## 1 Motivation

In this lecture, we are aiming to answer the general question of deciding when a particular case stands out. This issue can be subdivided into two questions:

i. **Outlier:** Is case $i$ strongly different from the overall regression line? What can we say about the distance between the $i^{\text{th}}$ observation and the corresponding $i^{\text{th}}$ fitted value predicted by the model? An *outlier* is therefore an observed value that does not follow the fitted model.

ii. **Influence:** Does case $i$ strongly *influence* the overall regression line? In particular, how is our regression slope influenced by the removal of case $i$? Not all outliers will have the same amount of impact on the values of the OLS estimators.

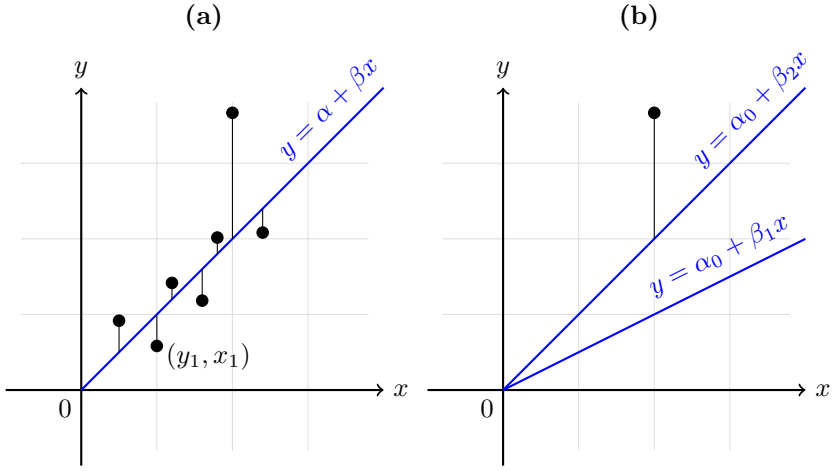


**Figure 1.** Examplary illustrations of an **outlier** in (a) and the **influence** exerted by an outlier in (b).

The main approach to the detection outliers is based on the assumption that there exists an *unaccounted for* amount of bias, denoted by $\delta$, for every data point. Thus, there is an unknown value *corrupting* the model, such that

$$\mathbb{E}[Y_i|\mathbf{x}_i] = \mathbf{x}_i^T\boldsymbol{\beta} + \delta, \qquad i = 1, \ldots, n.$$

We can then test for the size of $\delta$, for every $i$. Here, the difficulty is that we do not want to declare all data points as outliers. As a rule, a data point associated with a small leverage –and hence a large residual

variance– will be a good candidate. The strategy is therefore to run several simultaneous tests, which will require us to consider issues related to **multiple comparisons**.

# 2 Outliers

## 2.1 Standardized Residuals

From the properties of the hat matrix, we can deduce the first two moments of the residuals, such that

$$\mathbb{E}[\widehat{e}|\mathbf{X}] = 0, \quad \text{and} \quad \mathbb{V}\text{ar}[\widehat{e}_i|\mathbf{X}] = \sigma^2(1-h_{ii}),$$

for every $i = 1, \ldots, n$. It is convenient to standardize the variance of these residuals in order to obtain the following

$$\frac{\mathbb{V}\text{ar}[\widehat{e}_i|\mathbf{X}]}{\sigma^2(1-h_{ii})} = \mathbb{V}\text{ar}\left[\frac{\widehat{e}_i}{\sigma\sqrt{1-h_{ii}}}\middle|\mathbf{X}\right] = 1,$$

using the standard properties of the variance operator, which states that $\mathbb{V}\text{ar}[aX] = a^2\,\mathbb{V}\text{ar}[X]$.

Observe, however, that $\sigma^2$ is unknown. Therefore, we will define the **standardized residuals**, as follows,

$$\widehat{r}_i := \frac{\widehat{e}_i}{\widehat{\sigma}\sqrt{1-h_{ii}}},$$

where recall that $\widehat{\sigma}^2$ is the OLS estimator of the variance parameter. Formally,

$$\widehat{\sigma}^2 := \frac{\text{RSS}(\widehat{\boldsymbol{\beta}})}{n-p^*} = \frac{(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})}{n-p^*}. \tag{1}$$

The standardized residuals are useful for graphical representation. However, the variance of each residual is based on the same estimate of the variance. Therefore, they do not reflect the impact of each individual data point on the overall fit of the model. One way to tackle this problem is to compute the **studentized residuals**.

## 2.2 Case Deletion in Linear Regression

We will denote the removal of the $i^{\text{th}}$ observation in the vector $\mathbf{y}$, and in the matrix $\mathbf{X}$, through a parenthesized subscript $i$, such that

$$\mathbf{y}_{(i)} := \begin{vmatrix} y_1 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_n \end{vmatrix}, \qquad \mathbf{X}_{(i)} := \begin{vmatrix} 1 & x_{1,1} & \ldots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i-1,1} & \ldots & x_{i-1,p} \\ 1 & x_{i+1,1} & \ldots & x_{i+1,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \ldots & x_{n,p} \end{vmatrix}. \tag{2}$$

We will use this **reduced data set** to estimate $\boldsymbol{\beta}_{(i)}$. Once a case has been deleted, we can re-formulate all the quantities in the linear model with respect to the data and design matrix in equation (2). Thus, we have

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T\mathbf{y}_{(i)}.$$

Here, interest lies in evaluating the importance of the $i^{\text{th}}$ observation for **predicting itself**. That is, we wish to evaluate whether we can somewhat 'extrapolate' from the model without that case, in order to recover

the value of the missing data point. This is conducted by first estimating the missing observation using the vector of regressors based on the $n-1$ other data points,

$$\widehat{y}_{i(i)} := \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{(i)}.$$

Crucially, this can be done for **every data point** with $i = 1, \ldots, n$, such that we recover a vector of $n$ fitted values, denoted $\widehat{\mathbf{y}}_{(i)}$. It is then of interest to characterize the distribution of the differences, $Y_i - \widehat{Y}_{i(i)}$.

## 2.3 Standard Error for Outlier Test

Firstly, observe that since the definition of $\widehat{Y}_{i(i)}$ is not based on $Y_i$, it follows that these two quantities are (probabilistically) independent. Secondly, the variance of this difference is obtained as follows,

$$\begin{aligned}
\mathbb{V}\text{ar}[Y_i - \widehat{Y}_{i(i)}|\mathbf{X}] &= \mathbb{V}\text{ar}[Y_i|\mathbf{X}] + \mathbb{V}\text{ar}[\widehat{Y}_{i(i)}|\mathbf{X}_{(i)}] \\
&= \sigma^2 + \mathbf{x}_i^T \, \mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}_{(i)}|\mathbf{X}_{(i)}]\mathbf{x}_i \\
&= \sigma^2 + \sigma^2 \mathbf{x}_i^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{x}_i \\
&= \sigma^2 \left(1 + \mathbf{x}_i^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{x}_i\right),
\end{aligned}$$

using the standard properties of the variance operator, $\mathbb{V}\text{ar}[X - Y] = \mathbb{V}\text{ar}[X] + \mathbb{V}\text{ar}[Y] - 2\,\mathbb{C}\text{ov}[X, Y]$. Here, the theoretical variance of $\widehat{\boldsymbol{\beta}}_{(i)}$ is given by

$$\mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}_{(i)}|\mathbf{X}_{(i)}] = \sigma^2 (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1},$$

which follows from the fact that $\mathbb{V}\text{ar}[\mathbf{y}_{(i)}|\mathbf{X}] = \sigma^2 \mathbf{I}_{n-1}$.

Now, since the error terms are assumed to be independently and normally distributed, $e_i \sim N(0, \sigma^2)$ for every $i = 1, \ldots, n$; it immediately follows that the distribution of the difference $y_i - \widehat{y}_{i(i)}$ is given by

$$y_i - \widehat{y}_{i(i)}|\mathbf{X} \sim N\left(\delta, \sigma^2 \left(1 + \mathbf{x}_i^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{x}_i\right)\right). \tag{3}$$

for every $i = 1, \ldots, n$. This is a theoretical distribution, and a sample estimate is obtained for the variance of $y_i - \widehat{y}_{i(i)}$, by replacing $\sigma^2$ with the unbiased sample estimate $\widehat{\sigma}^2_{(i)}$, such that

$$\widehat{\mathbb{V}\text{ar}}[Y_i - \widehat{Y}_{i(i)}|\mathbf{X}] = \widehat{\sigma}^2_{(i)} \left(1 + \mathbf{x}_i^T (\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1} \mathbf{x}_i\right),$$

where $\widehat{\sigma}^2_{(i)}$ is defined as follows (see Sen and Srivastava, 1990, p.156),

$$\widehat{\sigma}^2_{(i)} = \frac{1}{n - 1 - p^*} \sum_{j=1, j\neq i}^{n} (y_j - \widehat{y}_{j(i)})^2,$$

where we are here using the fact that both vectors $\mathbf{y}$ and $\widehat{\mathbf{y}}_{(i)}$ are of order $n \times 1$. Compare this definition of $\widehat{\sigma}^2_{(i)}$ with the standard OLS estimator based on the full sample, in equation (1).

## 2.4 Studentized Residuals

Assuming that the errors are independent and normally distributed, we wish to test for the following null hypothesis,

$$H_0 : \mathbb{E}[Y_i - \widehat{Y}_{i(i)}|\mathbf{X}] = 0, \qquad H_1 : \mathbb{E}[Y_i - \widehat{Y}_{i(i)}|\mathbf{X}] \neq 0.$$

The corresponding $t$-statistic under the null hypothesis is then defined as follows,

$$t_i := \frac{y_i - \widehat{y}_{i(i)}}{\widehat{\sigma}_{(i)}\sqrt{1 + \mathbf{x}_i^T(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{x}_i}}.$$

Moreover, from the standard properties of the $t$-distribution and under the null hypothesis stating that $\delta = 0$, and using equation (3) it can then be shown that this statistic has indeed a $t$-distribution where the parametrization is given by the number of degrees of freedom used when computing the denominator, $\widehat{\sigma}_{(i)}$,

$$T_i \sim t(n - 1 - p^*),$$

for every $i = 1, \ldots, n$.

We are here conducting multiple tests simultaneously and moreover these tests are not independent. Therefore, we address this problem by using a conservative procedure, the **Bonferroni's correction** for multiple comparison, such that our new significance threshold becomes,

$$\alpha_B := \frac{\alpha}{n}.$$

The outlier test is then performed separately for each data point, and the null hypothesis is rejected if we obtain a $p$-value, which is less than the Bonferroni-adjusted $p$-value. Formally, $H_0$ is rejected for a **two-sided** $t$-test if

$$\widehat{p}_i := 2\mathbb{P}\big[T > |t_i|\big] \leq \frac{\alpha}{n},$$

where $T$ is $t$-distributed with $n - 1 - p^*$ degrees of freedom. However, once we have detected an outlier, the test itself does not tell us what to do about it. The ultimate decision of discarding an outlier lies in the hands of the researcher.

# 3 Studentized Residuals in Relation to OLS Estimators

In this section, we will derive an important relationship between the studentized residuals and the OLS estimators. This result will be used in the derivation of the Cook's distance. Here, we will need the following classical formula,

$$(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1 - h_{ii}}, \tag{4}$$

where recall that $h_{ii} := \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$.

Now, from the definition of the vector of regressors after removal of the $i^{\text{th}}$ case, we have

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T\mathbf{y}_{(i)}.$$

Observe that the $j^{\text{th}}$ entry of the $(p^* \times 1)$-dimensional vector $\mathbf{X}_{(i)}^T\mathbf{y}_{(i)}$ can be decomposed as follows,

$$\Big[\mathbf{X}_{(i)}^T\mathbf{y}_{(i)}\Big]_j = \sum_{k=1}^{n} x_{kj}y_k - x_{ij}y_i = \mathbf{x}_j^T\mathbf{y} - x_{ij}y_i,$$

where $\mathbf{x}_j$ is the $j^{\text{th}}$ *column* of $\mathbf{X}$, defined as a column vector. If considering all $p^*$ entries, this becomes

$$\mathbf{X}_{(i)}^T\mathbf{y}_{(i)} = \mathbf{X}^T\mathbf{y} - \mathbf{x}_i y_i,$$

where $\mathbf{x}_i$ represents the $i^{\text{th}}$ *row* of $\mathbf{X}$, also defined as a column vector.

Next, using this result in combination with equation (4), we obtain

$$\widehat{\boldsymbol{\beta}}_{(i)} = \left( (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1-h_{ii}} \right) \left( \mathbf{X}^T\mathbf{y} - \mathbf{x}_i y_i \right),$$

which can be expanded to

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i y_i + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}}{1-h_{ii}} - \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i h_{ii} y_i}{1-h_{ii}},$$

and simplified to

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_{ii}} \left( y_i(1-h_{ii}) - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}} + h_{ii}y_i \right).$$

Finally, this gives

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_{ii}} \left( y_i - \widehat{y}_i \right),$$

and since $\widehat{e}_i = y_i - \widehat{y}_i$,

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_{ii}}\, \widehat{e}_i. \tag{5}$$

This particular formula will aid with the interpretation of the Cook's distance, as described in the next section.

# 4 Influence Analysis

Cases whose removal result in major modification of the analysis are referred to as *influential*. As before, we will be interested in comparing the vector of regressors based on the full data set, $\widehat{\boldsymbol{\beta}}$, and the vector of regressors based on the data after removal of the $i^{\text{th}}$ case, and denoted $\widehat{\boldsymbol{\beta}}_{(i)}$. Here, we will study the **robustness** of the model to **perturbations** of the data.

## 4.1 Multivariate Confidence Region

We wish to compute the CI of the OLS vector $\widehat{\boldsymbol{\beta}}$. First, observe that since, under the standard distributional assumptions on the error terms, we know that $\widehat{\boldsymbol{\beta}}$ follows a multivariate distribution, the normalized quantity,

$$S(\widehat{\boldsymbol{\beta}}) := (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T \left[ \mathbb{V}\mathrm{ar}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] \right]^{-1} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) = \frac{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})}{\sigma^2},$$

follows a $\chi^2$-distribution with $p^*$ degrees of freedom,

$$S(\widehat{\boldsymbol{\beta}}) \sim \chi^2(p^*). \tag{6}$$

Here, we have simply used the standard relationship between the $\chi^2$-distribution and any sum of squared random variables having **unit normal** distribution, such that if

$$X_i \overset{\text{iid}}{\sim} N(0,1), \qquad i = 1, \ldots, d;$$

then

$$\sum_{i=1}^{d} X_i^2 \sim \chi^2(d).$$

Note that in equation (6), we have not yet replaced the unknown population variance, $\sigma^2$, by its empirical estimate. However, it is also known that (see Bain and Engelhardt, 1992, p.273),

$$(n-p^*)\widehat{\sigma}^2/\sigma^2 \sim \chi^2(n-p^*), \tag{7}$$

Moreover, recall that for any ratio of $\chi^2$-distributed random variables, $X_1$ and $X_2$, such that $X_1 \sim \chi^2(d_1)$ and $X_2 \sim \chi(d_2)$, we have

$$\frac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2).$$

Therefore, by taking the ratio of the two random variables in equations (6) and (7), and **normalizing** by their respective degrees of freedom, we obtain

$$\frac{S(\widehat{\boldsymbol{\beta}})/p^*}{\widehat{\sigma}^2/\sigma^2} = \frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{p^*\widehat{\sigma}^2} \sim F(p^*, n-p^*),$$

which follows an $F$-distribution with $p^*$ and $n-p^*$ degrees of freedom. Altogether, the $(1-\alpha)\times 100\%$ confidence region for $\boldsymbol{\beta}$ is thus the following subset of $\mathbb{R}^{p^*}$,

$$\left\{\boldsymbol{\beta} \in \mathbb{R}^{p^*} : \frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{p^*\widehat{\sigma}^2} \leq F(\alpha; p^*, n-p^*)\right\}.$$

## 4.2 Cook's Distance

Cook (1977) used a similar rationale to quantity the influence of each observation on $\widehat{\boldsymbol{\beta}}$. He suggested the following statistic,

$$D_i := \frac{(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(i)})^T(\mathbf{X}^T\mathbf{X})(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(i)})}{p^*\widehat{\sigma}^2},$$

for every $i = 1, \ldots, n$. The Cook's distance has an important desirable property: it is **invariant** under linear transformations of the columns of the design matrix, $\mathbf{X}$. Moreover, if we use $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{y}}_{(i)} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{(i)}$, we obtain a re-scaled version of the (squared) Euclidean metric in $\mathbb{R}^n$,

$$D_i = \frac{(\widehat{\mathbf{y}}-\widehat{\mathbf{y}}_{(i)})^T(\widehat{\mathbf{y}}-\widehat{\mathbf{y}}_{(i)})}{p^*\widehat{\sigma}^2} = \frac{||\widehat{\mathbf{y}}-\widehat{\mathbf{y}}_{(i)}||^2}{p^*\sigma^2}.$$

The Cook's distance can therefore be interpreted both in terms of the values of the $\widehat{\boldsymbol{\beta}}_{(i)}$'s and the ones of the $\mathbf{y}_{(i)}$'s, since deletion of the $i^{\text{th}}$ observation would have substantial consequences on both of these quantities. By using the formula obtained in equation (5), which stated that

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{(1-h_{ii})}\widehat{e}_i;$$

we can further simplify the $D_i$'s in order to obtain

$$D_i = \frac{\widehat{r}_i^2}{p^*}\frac{h_{ii}}{(1-h_{ii})},$$

which shows that the $D_i$'s are **monotonic increasing** in the leverages, since $h_{ii}/(1-h_{ii})$ is monotonic increasing in $h_{ii}$, over $[0,1]$. Similarly, the $D_i$'s are also monotonic increasing in the values of the **standardized residuals**, since we are here taking the square of each $\widehat{r}_i$. Thus, the values of the $D_i$'s are controlled by two distinct sources:

i. Firstly, the value of $\widehat{r}_i$ indicates the *degree of lack of fit* of the $i^{\text{th}}$ point, i.e. its distance from the value predicted by the mean function.

ii. Secondly, the value of $h_{ii}$ quantifies the distance of the $i^{\text{th}}$ point from the *mean profile*, $\bar{\mathbf{x}}$. Indeed, recall the following definition of $h_{ii}$ with respect to the centered design matrix,

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i^* - \bar{\mathbf{x}}^*)^T (\mathcal{X}^T \mathcal{X})^{-1} (\mathbf{x}_i^* - \bar{\mathbf{x}}^*).$$

It can thus be concluded that *influential observations* are special cases of *outliers*. That is, while all influential observations are also outliers, not all outliers are necessarily influential.

### 4.3 Magnitude of the Cook's Distances

The declaration of certain data points as influential or not influential is more an art than a science. We here proceed by analogy with our statistical treatment of $(1-\alpha)\times 100\%$ confidence regions for $\boldsymbol{\beta}$, and compare the $D_i$'s with the values of an $F$-statistic with $p^*$ and $n-p^*$ degrees of freedom. As an extreme case, we may have

$$D_i \geq F(\alpha; p^*, n-p^*).$$

If this inequality were to be satisfied, this would imply that removing the $i^{\text{th}}$ observation would be tantamount to moving $\widehat{\boldsymbol{\beta}}$ outside of its 95%-confidence region.

## References

Bain, L. and Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics*. Duxbury, Pacific Grove, CA.

Sen, A. and Srivastava, M. (1990). *Regression analysis: Theory, methods and applications*. Springer.