Chapter 8: Multiple Regression: Model Validation and Diagnostics

1 Residuals

Consider the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ again. The residual is defined as

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} := \boldsymbol{y} - \hat{\boldsymbol{y}},$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. The fitted value is given by

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is called the hat matrix. The hat matrix has the following properties:

$$\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X},$$

$$\boldsymbol{j} = \boldsymbol{H}\boldsymbol{j}, \quad \boldsymbol{x}_i = \boldsymbol{H}\boldsymbol{x}_i, \quad i = 1, 2, \ldots, k,$$

and

$$\hat{\boldsymbol{\epsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}.$$

In addition,

$$\hat{\boldsymbol{\epsilon}} = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}.$$

The following are some of the properties of $\hat{\boldsymbol{\epsilon}}$:

$$\begin{aligned}
E(\hat{\boldsymbol{\epsilon}}) &= 0, \\
\mathrm{Cov}(\hat{\epsilon}) &= \sigma^2(I - H), \\
\mathrm{Cov}(\hat{\boldsymbol{\epsilon}}, \boldsymbol{y}) &= \sigma^2(\boldsymbol{I} - \boldsymbol{H}), \\
\mathrm{Cov}(\hat{\boldsymbol{\epsilon}}, \hat{\boldsymbol{y}}) &= 0, \\
\bar{\hat{\boldsymbol{\epsilon}}} = \sum_{i=1}^{n} \hat{\epsilon}_i/n &= \hat{\boldsymbol{\epsilon}}^T\boldsymbol{j}/n = 0, \\
\hat{\boldsymbol{\epsilon}}^T\boldsymbol{y} &= SSE = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}, \\
\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{y}} &= 0, \\
\hat{\boldsymbol{\epsilon}}^T\boldsymbol{X} &= 0.
\end{aligned}$$

The above formulas imply that $\hat{\boldsymbol{\epsilon}}$ is orthogonal to $\hat{\boldsymbol{y}}$ and each column of $\boldsymbol{X}$, while it is correlated with $\boldsymbol{y}$. Therefore, we use the scatter plots $\hat{\boldsymbol{\epsilon}}$ versus $\hat{\boldsymbol{y}}$ to examine the pattern of the residual.

If the model is incorrect, various plots involving residuals may show departures from the fitted model such as outliers. curvature, or nonconstant variance.

2 The Hat Matrix

For the centered model,

$$\boldsymbol{y} = \alpha \boldsymbol{j} + \boldsymbol{X}_c \boldsymbol{\beta}_1 + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}_c = (\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{X}_1$, and $\hat{\boldsymbol{y}}$ becomes

$$\hat{\boldsymbol{y}} = \hat{\alpha} \boldsymbol{j} + \boldsymbol{X}_c \hat{\boldsymbol{\beta}}_1,$$

and we can define

$$\boldsymbol{H}_c = \boldsymbol{X}_c(\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^T.$$

Therefore,

$$\hat{\boldsymbol{y}} = \bar{y}\boldsymbol{j} + \boldsymbol{X}_c(\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^T \boldsymbol{y} = (\frac{1}{n}\boldsymbol{j}^T \boldsymbol{y})\boldsymbol{j} + \boldsymbol{H}_c \boldsymbol{y} = (\frac{1}{n}\boldsymbol{J} + \boldsymbol{H}_c)\boldsymbol{y}.$$

By arbitrariness of $\boldsymbol{y}$, we have

$$\boldsymbol{H} = \frac{1}{n}\boldsymbol{J} + \boldsymbol{H}_c = \frac{1}{n}\boldsymbol{J} + \boldsymbol{X}_c(\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^T.$$

Theorem 2.1. *If $\boldsymbol{X}$ is $n \times (k+1)$ of rank $k+1 < n$, and if the first column of $\boldsymbol{X}$ is $\boldsymbol{j}$, then the element h_{ij} of $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$ have the following properties:*

- $1/n \le h_{ii} \le 1$ *for* $i = 1, 2, \ldots, n$.
- $-0.5 \le h_{ij} \le 0.5$ *for all* $j \neq i$.
- $h_{ii} = 1/n + (\boldsymbol{x}_{1i} - \bar{\boldsymbol{x}}_1)^T (\boldsymbol{X}_c \boldsymbol{X}_c)^{-1} (\boldsymbol{x}_{1i} - \bar{\boldsymbol{x}}_1)$, *where* $\boldsymbol{x}_{1i}^T = (x_{i1}, x_{i2}, \ldots, x_{ik})$, $\bar{\boldsymbol{x}}^T = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)$, *and* $(\boldsymbol{x}_{1i} - \bar{\boldsymbol{x}}_1)^T$ *is the* i*th row of the centered matrix* $\boldsymbol{X}_c$.
- $tr(\boldsymbol{H}) = \sum_{i=1}^n h_{ii} = k + 1$.

Proof. (i) The lower bound follows from the relationship $\boldsymbol{H} = \frac{1}{n}\boldsymbol{J} + \boldsymbol{H}_c = \frac{1}{n}\boldsymbol{J} + \boldsymbol{X}_c(\boldsymbol{X}_c^T \boldsymbol{X}_c)^{-1} \boldsymbol{X}_c^T$, where $\boldsymbol{X}_c^T \boldsymbol{X}_c$ is positive definite. The upper bound follows from the property $H = H^2$, which implies

$$h_{ii} = \boldsymbol{h}_i^T \boldsymbol{h}_i = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2,$$

or, equivalently,

$$1 = h_{ii} + \sum_{j \neq i} h_{ij}^2 / h_{ii},$$

which implies $h_{ii} \leq 1$.

(ii) Since $h_{ii} = \boldsymbol{h}_i^T \boldsymbol{h}_i = h_{ii}^2 + h_{ij}^2 + \sum_{r \neq i,j} h_{ir}^2$, we have

$$h_{ii} - h_{ii}^2 = h_{ij}^2 + \sum_{r \neq i,j} h_{ir}^2,$$

and thus $h_{ij}^2 \leq h_{ii} - h_{ii}^2$. Since the maximum value of $h_{ii} - h_{ii}^2$ is $1/4$, we have $-0.5 \leq h_{ij} \leq 0.5$. □

3 Outliers

For outlier analysis, we need to keep in mind that the variance of the residuals is not constant:

$$\mathrm{Var}(\hat{\epsilon}) = \sigma^2(1 - h_{ii}).$$

An additional verification that large values of h_{ii} are accompanied by small residuals is provided by the property:

$$\frac{1}{n} \leq h_{ii} + \frac{\hat{\epsilon}_i^2}{\hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}} \leq 1.$$

There are two common methods of scaling of the residuals:

- Standardized residual:

  $$r_i = \frac{\hat{\epsilon}_i}{s\sqrt{1 - h_{ii}}},$$

  where $s = \sqrt{SSE/(n-k-1)}$.

- Studentized residual:

  $$t_i = \frac{\hat{\epsilon}_i}{s_{(i)}\sqrt{1 - h_{ii}}},$$

  where $s_{(i)}$ is the standard error computed with the $n-1$ samples remaining after omitting the ith sample. Such a residual is also called a studentized deleted residual or externally studentized residual. Alternatively, t_i can be calculated as

  $$t_i = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2},$$

  where $p = k + 1$ denotes the number of columns of $\boldsymbol{X}$.

For deleted residuals, we have the following relationships:

$$\hat{\epsilon}_{(i)} = y_i - \hat{y}_{(i)} = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(i)},$$

where $\hat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} \boldsymbol{X}_{(i)}^T \boldsymbol{y}_{(i)}$, $\boldsymbol{X}_{(i)}$ is an $(n-1) \times (k+1)$ matrix. In addition, we have

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{\hat{\epsilon}_i}{1 - h_{ii}} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i.$$

The deleted residual can also be expressed as

$$\hat{\epsilon}_{(i)} = \frac{\hat{\epsilon}_i}{1 - h_{ii}},$$

and

$$t_i = \frac{\hat{\epsilon}_{(i)}}{s_{(i)}}.$$

The deleted sample variance $s^2_{(i)}$ can be expressed as

$$s^2_{(i)} = \frac{SSE_{(i)}}{n - k - 2} = \frac{SSE - \hat{\epsilon}_i^2/(1 - h_{ii})}{n - k - 2}.$$

The n deleted residuals can be used for model validation or selection by defining the prediction sum of squares (PRESS):

$$PRESS = \sum_{i=1}^{n} \hat{\epsilon}^2_{(i)} = \sum_{i=1}^{n} \left(\frac{\hat{\epsilon}_i}{1 - h_{ii}} \right)^2.$$

To use PRESS to compare alternative models when the objective is prediction, preference would be shown to models with small values of PRESS.

4 Influential Observations and Leverage

4.1 Leverage

To investigate the influence of each observation, we begin with $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, the element of which are

$$\hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j.$$

Therefore, if h_{ii} is large (close to 1), then h_{ij}'s, $j \neq i$, are small, and y_i contributes much more than others to $\hat{y}_i$. Hence, h_{ii} is called the leverage of y_i.

By Theorem 2.1, the average value of h_{ii}'s is $(k + 1)/n$. Hoaglin and Welsch (1978) suggest that a point with $h_{ii} > 2(k + 1)/n$ is a high leverage point. Alternatively, we can simply examine any observations whose value of h_{ii} is unusually lareg relative to the other values of h_{ii}.

4.2 Cook's Distance

To formalize the influence of an observation, we consider the effect of its deletion on $\boldsymbol{\beta}$ and $\hat{\boldsymbol{y}}$. This is measured by Cook's distance

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T \boldsymbol{X}^T \boldsymbol{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(k + 1)s^2} = \frac{(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})^T (\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})}{(k + 1)s^2}.$$

Therefore, if D_i is large, the observation i has substantial influence on both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{y}}$. A more computationally convenience form of D_i is given by

$$D_i = \frac{r_i^2}{k + 1} \left(\frac{h_{ii}}{1 - h_{ii}} \right).$$