STAT 574 Linear and Nonlinear Mixed Models

Lecture 10: Diagnoses and Influence Analysis

Chencheng Cai

Washington State University

Influence analysis

- Influence analysis is a set of techniques used to identify and assess the impact of individual data points on the overall model fit and parameter estimates.
- In this lecture, we consider the **influence analysis** as a **sensitivity analysis** of the model fit to the data.
- Data influence: the sensitivity of the model to a infinitesimal purturbation in the data.
- Model influence: the sensitivity of the model to the assumptions.

Linear Regression Model

Consider a linear regression model:

$$y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i \quad \forall i.$$

The OLS estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} = \left(\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \sum_i \boldsymbol{x}_i y_i,$$

where $\boldsymbol{X}$ is the design matrix and $\boldsymbol{y}$ is the response vector.

Leverage

The **leverage** of the i-th observation is defined as the i-th diagonal element of the hat matrix

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

That is

$$h_i = (\boldsymbol{H})_{ii} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i.$$

Leverage

The sum of the leverages is equal to the number of parameters in the model:

$$\begin{aligned}\sum_{i=1}^{n} h_i &= \mathrm{tr}(\boldsymbol{H}) \\ &= \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\ &= \mathrm{tr}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}) \\ &= \mathrm{tr}(\boldsymbol{I}) \\ &= m,\end{aligned}$$

where m is the number of parameters in the model.

- Observations with high leverage are called **influential observations**.
- It measures how the predicted value is influenced by the i-th observation.

Leave-one-out

Another measure of influence is to check the change in the estimates of the parameters when the i-th observation is removed from the data set.
The estimated parameter $\hat{\boldsymbol{\beta}}_{(i)}$ when the i-th observation is removed is given by

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

where $\boldsymbol{X}_{(i)}$ and $\boldsymbol{y}_{(i)}$ are the design matrix and response vector with the i-th observation removed.

Leave-one-out

We notice that

$$(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1} = (\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i^T)^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{x}_i\boldsymbol{x}_i^T)(\boldsymbol{X}^T\boldsymbol{X})^{-1}}{1 - \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i}$$

Then

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(i)} &= \left(\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}(\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{x}_i y_i) \\
&= \left[(\boldsymbol{X}^T\boldsymbol{X})^{-1} + (1-h_i)^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right](\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{x}_i y_i) \\
&= \hat{\boldsymbol{\beta}} + (1-h_i)^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i \\
&\quad - (1-h_i)^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i \\
&= \hat{\boldsymbol{\beta}} + (1-h_i)^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{y}_i - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i - \frac{h_i}{1-h_i}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i y_i \\
&= \hat{\boldsymbol{\beta}} - \frac{y_i - \hat{y}_i}{1-h_i}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i.
\end{aligned}$$

Cook's distance

The confidence region for the estimator $\hat{\boldsymbol{\beta}}$ is given by

$$\left\{\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq m s^2 F_{\alpha,m,n-m}\right\}$$

The **Cook's distance** is defined as

$$D_i = \frac{1}{ms^2} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})$$

By previous result, we have

$$D_i = \frac{(y_i - \hat{y}_i)^2}{ms^2} \frac{h_i}{(1 - h_i)^2}$$

Large values of D_i indicate that the i-th observation has a large influence on the fitted model.

Infinitesimal Influence (I-influence)

- $\boldsymbol{D}$: data vector including all the observed values.
- $\boldsymbol{t}(\boldsymbol{D})$: a statistic of interest.
- The **infinitesimal data influence** is

$$\lim_{\Delta D \to 0} \frac{\boldsymbol{t}(\boldsymbol{D} + \Delta D \boldsymbol{e}_i) - \boldsymbol{t}(\boldsymbol{D})}{\Delta D} = \frac{\partial \boldsymbol{t}(\boldsymbol{D})}{\partial D_i}$$

Infinitesimal Influence (I-influence)

- Let $\ell(\boldsymbol{\theta})$ be the log-likelihood function of the model.
- Consider a more general model $\ell(\boldsymbol{\theta} \mid \boldsymbol{\omega})$ such that $\boldsymbol{\omega} = 0$ corresponds to the model of interest.
- The **infinitesimal model influence** is defined as

$$\left.\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{\omega}}\right|_{\boldsymbol{\omega}=0}$$

Influence of the Dependent Variable

We consider the influence of y_i on the estimated parameters $\hat{\boldsymbol{\beta}}$.

- The OLS solution is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sum_i \boldsymbol{x}_i y_i.$$

- The influence of y_i on $\hat{\boldsymbol{\beta}}$ is given by

$$\frac{\partial \hat{\boldsymbol{\beta}}}{\partial y_i} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i.$$

- Two possibilities that the size of the influence is large:
  - $|\boldsymbol{x}_i|$ is large: the i-th observation is far from the center of the data.
  - The direction of $\boldsymbol{x}_i$ is close to the direction of the eigenvector of $\boldsymbol{X}^T\boldsymbol{X}$ corresponding to the smallest eigenvalue.

Influence of the Continuous Explanatory Variable

In particular, we are interested in

$$\frac{\partial \hat{\boldsymbol{\beta}}}{\partial x_{ik}},$$

where x_{ik} is the k-th element of $\boldsymbol{x}_i$.
Use matrix calculus, we have (by considering $\boldsymbol{x}_i$ as the only variable)

$$\begin{aligned}
d\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1} d(\boldsymbol{X}^T\boldsymbol{y}) + \left(d(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)(\boldsymbol{X}^T\boldsymbol{y}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1} y_i d\boldsymbol{x}_i - (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{x}_i d\boldsymbol{x}_i^T + (d\boldsymbol{x}_i)\boldsymbol{x}_i^T)(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{y}) \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1} y_i d\boldsymbol{x}_i - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i (d\boldsymbol{x}_i)^T\hat{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}(d\boldsymbol{x}_i)\hat{y}_i \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}(y_i - \hat{y}_i) d\boldsymbol{x}_i - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\hat{\boldsymbol{\beta}}^T d\boldsymbol{x}_i.
\end{aligned}$$

Therefore,

$$\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{x}_i} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\left((y_i - \hat{y}_i)\boldsymbol{I} - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}^T\right)$$

Influence of the Continuous Explanatory Variable

Now we have

$$\frac{\partial \hat{\boldsymbol{\beta}}}{\partial x_{ik}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\left((y_i - \hat{y}_i)\boldsymbol{e}_k - \hat{\beta}_k \boldsymbol{x}_i\right)$$

- First component: normalized residual.
- Second component: the influence through the dependent variable.

It is also connected to Cook's local influence, which is measured by

$$\text{local influence of } x_{ik} \text{ on } \hat{\beta}_k = y_i - \hat{y}_i - \hat{\beta}_k q_i$$

where q_i is the residual of the regression of $\boldsymbol{x}^{(k)}$ on the other variables.

Influence of the Binary Explanatory Variable

Now we assume x_{ik} is a binary variable.

- A binary variable can be misclassified.
- We assume x_{ik} is an observation for the true binary variable z_{ik} such that miscalssification occurs with probability q_i.
- The true model should be

$$E(y_i \mid z_{ik}) = \alpha + \beta_k z_{ik}$$

- But now

$$E(y_i \mid x_{ik}) = \alpha + \beta_k(x_{ik} + (1 - 2x_{ik})q_i)$$

- The influence of the misclassification is given by

$$\left.\frac{\partial \hat{\boldsymbol{\beta}}}{\partial q_i}\right|_{q_i=0} = (1 - 2x_{ik})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\left((y_i - \hat{y}_i)\boldsymbol{e}_i - \hat{\beta}_k \boldsymbol{x}_i\right)$$

Influence on the Predicted Value

The predicted value is connected to the estimated parameters by

$$\hat{y}_i = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i.$$

The influence can be transferred to the influence on the predicted value. That is

$$\begin{aligned}
\frac{\partial \hat{y}_i}{\partial x_{ik}} &= \frac{\partial \hat{y}_i}{\partial \hat{\boldsymbol{\beta}}} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial x_{ik}} + \left.\frac{\partial \hat{y}_i}{\partial x_{ik}}\right|_{\hat{\boldsymbol{\beta}}} \\
&= \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \left((y_i - \hat{y}_i) \boldsymbol{e}_k - \hat{\beta}_k \boldsymbol{x}_i \right) + \hat{\beta}_k \\
&= \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{x})^{-1} \boldsymbol{e}_k (y_i - \hat{y}_i) + (1 - h_i) \hat{\beta}_k.
\end{aligned}$$

Influence on Regression Characteristics

Y-influence on

- **Coefficient of determination** R^2:

$$\frac{\partial R^2}{\partial y_i} = \frac{2}{\mathrm{SST}} \left[(1 - R^2)(y_i - \bar{y}) - (y_i - \hat{y}_i) \right]$$

- **t-statistics** with $\boldsymbol{t} = s^{-1}\boldsymbol{D}^{-1/2}\hat{\boldsymbol{\beta}}$ where $\boldsymbol{D} = \mathrm{diag}((\boldsymbol{X}^T\boldsymbol{X})^{-1})$:

$$\frac{\partial \boldsymbol{t}}{\partial y_i} = s^{-1}\boldsymbol{D}^{-1/2}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i - \frac{y_i - \hat{y}_i}{\mathrm{RSS}}\boldsymbol{t}_i$$