Lecture 14

October 3, 2018

Regression Diagnostics

Regression Diagnostics

- The estimates for β and their confidence intervals etc

Regression Diagnostics

- The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model.

Regression Diagnostics

- The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model. In particular, note that we have assumed:

Regression Diagnostics

- The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model. In particular, note that we have assumed:
  1. The errors $e_1, \ldots, e_n$ are independent,

Regression Diagnostics

- The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model. In particular, note that we have assumed:
  1. The errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2

Regression Diagnostics

- The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model. In particular, note that we have assumed:
  1. The errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2 and are normally distributed.

Regression Diagnostics

- The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model. In particular, note that we have assumed:
  1. The errors $e_1, \dots, e_n$ are independent, have equal variance σ^2 and are normally distributed.
  2. We have assumed that the expected value of the response vector Y equals $X\beta$.

Regression Diagnostics

- The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model. In particular, note that we have assumed:
  1. The errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2 and are normally distributed.
  2. We have assumed that the expected value of the response vector Y equals $X\beta$.
  3. We have assumed that all the subjects obey the same linear model.

Regression Diagnostics

- The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model. In particular, note that we have assumed:
  1. The errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2 and are normally distributed.
  2. We have assumed that the expected value of the response vector Y equals $X\beta$.
  3. We have assumed that all the subjects obey the same linear model. In practice, it may happen that a few subjects do not obey the model.

Regression Diagnostics

- The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model. In particular, note that we have assumed:
  1. The errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2 and are normally distributed.
  2. We have assumed that the expected value of the response vector Y equals $X\beta$.
  3. We have assumed that all the subjects obey the same linear model. In practice, it may happen that a few subjects do not obey the model. These few observations might change the choice and fit of the model.

Regression Diagnostics

- The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model. In particular, note that we have assumed:
  1. The errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2 and are normally distributed.
  2. We have assumed that the expected value of the response vector Y equals $X\beta$.
  3. We have assumed that all the subjects obey the same linear model. In practice, it may happen that a few subjects do not obey the model. These few observations might change the choice and fit of the model.

  Fortunately, we can do regression diagnostics to check if the data show any evidence of deviating from the assumptions.

Regression Diagnostics

- The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model. In particular, note that we have assumed:
  1. The errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2 and are normally distributed.
  2. We have assumed that the expected value of the response vector Y equals $X\beta$.
  3. We have assumed that all the subjects obey the same linear model. In practice, it may happen that a few subjects do not obey the model. These few observations might change the choice and fit of the model.

  Fortunately, we can do regression diagnostics to check if the data show any evidence of deviating from the assumptions.
- Regression diagnostics should always be performed after regression analysis.

Regression Diagnostics

- We will first discuss the third part above.

Regression Diagnostics

- We will first discuss the third part above. How to detect the presence of unusual observations.

Regression Diagnostics

- We will first discuss the third part above. How to detect the presence of unusual observations. We will mainly use leverages,

Regression Diagnostics

- We will first discuss the third part above. How to detect the presence of unusual observations. We will mainly use leverages, jackknife residuals (or standardized predicted residuals)

Regression Diagnostics

- We will first discuss the third part above. How to detect the presence of unusual observations. We will mainly use leverages, jackknife residuals (or standardized predicted residuals) and Cook's distance for this purpose.

Leverages and Mahalanobis Distance

- For the bootstrap confidence intervals to be accurate, we need that there are no points with very high leverage.

.

Leverages and Mahalanobis Distance

- For the bootstrap confidence intervals to be accurate, we need that there are no points with very high leverage. What does high leverage mean?

.

Leverages and Mahalanobis Distance

- For the bootstrap confidence intervals to be accurate, we need that there are no points with very high leverage. What does high leverage mean?
- It turns out that **the *i*th subject has high leverage if and only if**

.

Leverages and Mahalanobis Distance

- For the bootstrap confidence intervals to be accurate, we need that there are no points with very high leverage. What does high leverage mean?
- It turns out that **the *i*th subject has high leverage if and only if** the explanatory variable values for the *i*th subject are far from the explanatory variable values of the remaining subjects.

Leverages and Mahalanobis Distance

- For the bootstrap confidence intervals to be accurate, we need that there are no points with very high leverage. What does high leverage mean?
- It turns out that **the *i*th subject has high leverage if and only if** the explanatory variable values for the *i*th subject are far from the explanatory variable values of the remaining subjects.
- To understand this, we need to know how to define "far from the explanatory variable values of the remaining subjects".

Leverages and Mahalanobis Distance

- For the bootstrap confidence intervals to be accurate, we need that there are no points with very high leverage. What does high leverage mean?
- It turns out that **the *i*th subject has high leverage if and only if** the explanatory variable values for the *i*th subject are far from the explanatory variable values of the remaining subjects.
- To understand this, we need to know how to define "far from the explanatory variable values of the remaining subjects". This is done via the notion of *Mahalanobis distance*.

Leverages and Mahalanobis Distance

- For the bootstrap confidence intervals to be accurate, we need that there are no points with very high leverage. What does high leverage mean?
- It turns out that **the *i*th subject has high leverage if and only if** the explanatory variable values for the *i*th subject are far from the explanatory variable values of the remaining subjects.
- To understand this, we need to know how to define "far from the explanatory variable values of the remaining subjects". This is done via the notion of *Mahalanobis distance*.
- Given a set of points (also called a point cloud) $v_1, \dots, v_n$ on the real line

Leverages and Mahalanobis Distance

- For the bootstrap confidence intervals to be accurate, we need that there are no points with very high leverage. What does high leverage mean?
- It turns out that **the *i*th subject has high leverage if and only if** the explanatory variable values for the ith subject are far from the explanatory variable values of the remaining subjects.
- To understand this, we need to know how to define "far from the explanatory variable values of the remaining subjects". This is done via the notion of *Mahalanobis distance*.
- Given a set of points (also called a point cloud) $v_1, \ldots, v_n$ on the real line and a test point v, can we quantify how far v is from the set?

Leverages and Mahalanobis Distance

- For the bootstrap confidence intervals to be accurate, we need that there are no points with very high leverage. What does high leverage mean?
- It turns out that **the *i*th subject has high leverage if and only if** the explanatory variable values for the *i*th subject are far from the explanatory variable values of the remaining subjects.
- To understand this, we need to know how to define "far from the explanatory variable values of the remaining subjects". This is done via the notion of *Mahalanobis distance*.
- Given a set of points (also called a point cloud) $v_1, \dots, v_n$ on the real line and a test point v, can we quantify how far v is from the set?
- A natural idea is to look at $|v - \bar{v}|^2$ where $\bar{v} = (v_1 + \cdots + v_n)/n$.

Leverages and Mahalanobis Distance

- This makes sense but it does not give a sense of scale.

Leverages and Mahalanobis Distance

- This makes sense but it does not give a sense of scale. For example if I get $|v - \bar{v}|^2 = 1$, then does this mean that v is far from the point cloud or near it.

Leverages and Mahalanobis Distance

- This makes sense but it does not give a sense of scale. For example if I get $|v - \bar{v}|^2 = 1$, then does this mean that v is far from the point cloud or near it.
- A better idea is to look at

Leverages and Mahalanobis Distance

- This makes sense but it does not give a sense of scale. For example if I get $|v - \bar{v}|^2 = 1$, then does this mean that v is far from the point cloud or near it.
- A better idea is to look at

$$\frac{(v - \bar{v})^2}{s^2}$$

Leverages and Mahalanobis Distance

- This makes sense but it does not give a sense of scale. For example if I get $|v - \bar{v}|^2 = 1$, then does this mean that v is far from the point cloud or near it.
- A better idea is to look at

$$\frac{(v - \bar{v})^2}{s^2}$$

where

$$s^2 := \frac{1}{n-1} \sum_{j=1}^{n} (v_j - \bar{v})^2$$

Leverages and Mahalanobis Distance

- This easily generalizes to the multivariate case. Suppose $v_1, \dots, v_n$ are all in $\mathbb{R}^p$. Then for a test point $v \in \mathbb{R}^p$, its distance to the point cloud $\{v_1, \dots, v_n\}$ is measured by the *Mahalanobis distance* defined by

  $$(v - \bar{v})^T S^{-1} (v - \bar{v}).$$

  where

  $$S := \frac{1}{n-1} \sum_{j=1}^{n} (v_j - \bar{v})(v_j - \bar{v})^T$$

Leverages and Mahalanobis Distance

- Let us now see how this applies to our regression setup with Y and X.

Leverages and Mahalanobis Distance

- Let us now see how this applies to our regression setup with Y and X.
- Let the ith row of the X matrix be denoted by x_i^T.

Leverages and Mahalanobis Distance

- Let us now see how this applies to our regression setup with Y and X.
- Let the ith row of the X matrix be denoted by x_i^T. Therefore x_i is a $(p+1) \times 1$ vector whose first entry is one

Leverages and Mahalanobis Distance

- Let us now see how this applies to our regression setup with Y and X.
- Let the ith row of the X matrix be denoted by x_i^T. Therefore x_i is a $(p+1) \times 1$ vector whose first entry is one and the rest of the entries carry the values of the explanatory variables for the ith subject.

Leverages and Mahalanobis Distance

- Let us now see how this applies to our regression setup with Y and X.
- Let the ith row of the X matrix be denoted by x_i^T. Therefore x_i is a $(p+1) \times 1$ vector whose first entry is one and the rest of the entries carry the values of the explanatory variables for the ith subject.
- We can write $x_i^T = [1, z_i^T]$

Leverages and Mahalanobis Distance

- Let us now see how this applies to our regression setup with Y and X.
- Let the ith row of the X matrix be denoted by x_i^T. Therefore x_i is a $(p+1) \times 1$ vector whose first entry is one and the rest of the entries carry the values of the explanatory variables for the ith subject.
- We can write $x_i^T = [1, z_i^T]$ where z_i^T just contains the values of the explanatory variables (without 1) for the ith subject.

Leverages and Mahalanobis Distance

- Let us now see how this applies to our regression setup with Y and X.
- Let the ith row of the X matrix be denoted by x_i^T. Therefore x_i is a $(p+1) \times 1$ vector whose first entry is one and the rest of the entries carry the values of the explanatory variables for the ith subject.
- We can write $x_i^T = [1, z_i^T]$ where z_i^T just contains the values of the explanatory variables (without 1) for the ith subject.
- The Mahalanobis distance for the ith subject is defined as

Leverages and Mahalanobis Distance

- Let us now see how this applies to our regression setup with Y and X.
- Let the ith row of the X matrix be denoted by x_i^T. Therefore x_i is a $(p+1) \times 1$ vector whose first entry is one and the rest of the entries carry the values of the explanatory variables for the ith subject.
- We can write $x_i^T = [1, z_i^T]$ where z_i^T just contains the values of the explanatory variables (without 1) for the ith subject.
- The Mahalanobis distance for the ith subject is defined as

$$\Gamma_i := (z_i - \bar{z})^T S^{-1} (z_i - \bar{z}),$$

Leverages and Mahalanobis Distance

- Let us now see how this applies to our regression setup with Y and X.
- Let the ith row of the X matrix be denoted by x_i^T. Therefore x_i is a $(p+1) \times 1$ vector whose first entry is one and the rest of the entries carry the values of the explanatory variables for the ith subject.
- We can write $x_i^T = [1, z_i^T]$ where z_i^T just contains the values of the explanatory variables (without 1) for the ith subject.
- The Mahalanobis distance for the ith subject is defined as

$$\Gamma_i := (z_i - \bar{z})^T S^{-1} (z_i - \bar{z}),$$

where

$$S := \frac{1}{n-1} \sum_{j=1}^{n} (z_j - \bar{z})(z_j - \bar{z})^T$$

Leverages and Mahalanobis Distance

- Let us now see how this applies to our regression setup with Y and X.
- Let the ith row of the X matrix be denoted by x_i^T. Therefore x_i is a $(p+1) \times 1$ vector whose first entry is one and the rest of the entries carry the values of the explanatory variables for the ith subject.
- We can write $x_i^T = [1, z_i^T]$ where z_i^T just contains the values of the explanatory variables (without 1) for the ith subject.
- The Mahalanobis distance for the ith subject is defined as

$$\Gamma_i := (z_i - \bar{z})^T S^{-1} (z_i - \bar{z}),$$

where

$$S := \frac{1}{n-1} \sum_{j=1}^{n} (z_j - \bar{z})(z_j - \bar{z})^T$$

and

$$\bar{z} = \sum^{n} z_j / n.$$

- Γ_i clearly measures how far the ith subject is from the rest of the subjects

$$ \tag{1} $$

- Γ_i clearly measures how far the ith subject is from the rest of the subjects in terms of the values of the explanatory variables.

$$ \tag{1} $$

- Γ_i clearly measures how far the ith subject is from the rest of the subjects in terms of the values of the explanatory variables.
- It turns out that the leverage for the ith subject

(1)

- Γ_i clearly measures how far the *i*th subject is from the rest of the subjects in terms of the values of the explanatory variables.
- It turns out that the leverage for the *i*th subject , h_{ii},

(1)

- Γ_i clearly measures how far the *i*th subject is from the rest of the subjects in terms of the values of the explanatory variables.
- It turns out that the leverage for the *i*th subject , h_{ii}, is related to Γ_i by the following simple expression:

(1)

- Γ_i clearly measures how far the *i*th subject is from the rest of the subjects in terms of the values of the explanatory variables.
- It turns out that the leverage for the *i*th subject , h_{ii}, is related to Γ_i by the following simple expression:

$$\Gamma_i = (n-1)h_{ii} - \frac{n-1}{n}. \tag{1}$$

- Γ_i clearly measures how far the *i*th subject is from the rest of the subjects in terms of the values of the explanatory variables.
- It turns out that the leverage for the *i*th subject , h_{ii}, is related to Γ_i by the following simple expression:

$$\Gamma_i = (n-1)h_{ii} - \frac{n-1}{n}. \tag{1}$$

- This explains the second interpretation of leverage.

- Γ_i clearly measures how far the *i*th subject is from the rest of the subjects in terms of the values of the explanatory variables.
- It turns out that the leverage for the *i*th subject , h_{ii}, is related to Γ_i by the following simple expression:

$$\Gamma_i = (n-1)h_{ii} - \frac{n-1}{n}. \tag{1}$$

- This explains the second interpretation of leverage. For the above formula to hold, it is necessary that there be an intercept term in the model.

- Γ_i clearly measures how far the ith subject is from the rest of the subjects in terms of the values of the explanatory variables.
- It turns out that the leverage for the ith subject , h_{ii}, is related to Γ_i by the following simple expression:

$$\Gamma_i = (n-1)h_{ii} - \frac{n-1}{n}. \tag{1}$$

- This explains the second interpretation of leverage. For the above formula to hold, it is necessary that there be an intercept term in the model.
- In other words, this formula does not hold without an intercept term.

- Γ_i clearly measures how far the ith subject is from the rest of the subjects in terms of the values of the explanatory variables.
- It turns out that the leverage for the ith subject , h_{ii}, is related to Γ_i by the following simple expression:

$$\Gamma_i = (n-1)h_{ii} - \frac{n-1}{n}. \tag{1}$$

- This explains the second interpretation of leverage. For the above formula to hold, it is necessary that there be an intercept term in the model.
- In other words, this formula does not hold without an intercept term. We will not go over the proof of this formula. It can be easily verified in R.

- Note that one consequence of (1) is the fact that $h_{ii} \geq 1/n$ for every i.

- Note that one consequence of (1) is the fact that $h_{ii} \geq 1/n$ for every i. This is because Γ_i is always nonnegative.

- Note that one consequence of (1) is the fact that $h_{ii} \geq 1/n$ for every i. This is because Γ_i is always nonnegative. The leverages therefore always lie between $1/n$ and 1 whenever there is an intercept term in the linear model.