Lecture 15

October 11, 2018

Finding Unusual Observations in Regression

Finding Unusual Observations in Regression

- Consider the regression setup with Y and X.

Finding Unusual Observations in Regression

- Consider the regression setup with Y and X. Let the ith row of the X matrix be denoted by x_i^T.

Finding Unusual Observations in Regression

- Consider the 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.

Finding Unusual Observations in Regression

- Consider the 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.

Finding Unusual Observations in Regression

- Consider the 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

$$S := \frac{1}{n-1} \sum_{j=1}^{n} (z_j - \bar{z})(z_j - \bar{z})^T, \quad \text{and} \;\; \bar{z} = \sum_{j=1}^{n} z_j / n.$$

Finding Unusual Observations in Regression

- Consider the 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

  $$S := \frac{1}{n-1} \sum_{j=1}^{n} (z_j - \bar{z})(z_j - \bar{z})^T, \quad \text{and} \quad \bar{z} = \sum_{j=1}^{n} z_j / n.$$

- 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.

Leverages

- It turns out that the leverage for the *i*th subject, h_{ii}, is related to Γ_i by the following simple expression:

$$h_{ii} = \frac{\Gamma_i}{n-1} + \frac{1}{n} \tag{1}$$

Leverages

- It turns out that the leverage for the *i*th subject, h_{ii}, is related to Γ_i by the following simple expression:

$$h_{ii} = \frac{\Gamma_i}{n-1} + \frac{1}{n} \tag{1}$$

- For the above formula to hold, it is necessary that there be an intercept term in the model.

Leverages

- It turns out that the leverage for the ith subject, h_{ii}, is related to Γ_i by the following simple expression:

$$h_{ii} = \frac{\Gamma_i}{n-1} + \frac{1}{n} \tag{1}$$

- 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.

Leverages

- It turns out that the leverage for the *i*th subject, h_{ii}, is related to Γ_i by the following simple expression:

$$h_{ii} = \frac{\Gamma_i}{n-1} + \frac{1}{n} \tag{1}$$

- 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.

Leverages

- It turns out that the leverage for the ith subject, h_{ii}, is related to Γ_i by the following simple expression:

$$h_{ii} = \frac{\Gamma_i}{n-1} + \frac{1}{n} \tag{1}$$

- 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.

Leverages

- It turns out that the leverage for the *i*th subject, h_{ii}, is related to Γ_i by the following simple expression:

$$h_{ii} = \frac{\Gamma_i}{n-1} + \frac{1}{n} \tag{1}$$

- 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.
- This is because Γ_i is always nonnegative.

Leverages

- It turns out that the leverage for the ith subject, h_{ii}, is related to Γ_i by the following simple expression:

$$h_{ii} = \frac{\Gamma_i}{n-1} + \frac{1}{n} \tag{1}$$

- 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.
- 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.

Leverages

- What do we do if we find that some of the subjects have high leverages?

Leverages

- What do we do if we find that some of the subjects have high leverages?
- We should remove these observations, perform the regression again and check by how much the results have changed.

Leverages

- What do we do if we find that some of the subjects have high leverages?
- We should remove these observations, perform the regression again and check by how much the results have changed.
- Two notions are useful here: Predicted Residuals and Cook's distance.

Outliers – Outlying Y

- A common way to detect outlying points in the Y direction is to look at the residuals.

Outliers – Outlying Y

- A common way to detect outlying points in the Y direction is to look at the residuals.
- The residuals are also fundamental in looking at other assumptions of the model.

Outliers – Outlying Y

- A common way to detect outlying points in the Y direction is to look at the residuals.
- The residuals are also fundamental in looking at other assumptions of the model. However, the residuals $\hat{e}$ satisfy $var(\hat{e}) = \sigma^2(I - H)$.

Outliers – Outlying Y

- A common way to detect outlying points in the Y direction is to look at the residuals.
- The residuals are also fundamental in looking at other assumptions of the model. However, the residuals $\hat{e}$ satisfy $var(\hat{e}) = \sigma^2(I - H)$. In particular, it is important to know that the residuals are correlated and have different variances.

Outliers – Outlying Y

- A common way to detect outlying points in the Y direction is to look at the residuals.
- The residuals are also fundamental in looking at other assumptions of the model. However, the residuals $\hat{e}$ satisfy $var(\hat{e}) = \sigma^2(I - H)$. In particular, it is important to know that the residuals are correlated and have different variances.
- **Standardized Residuals** For diagnostics, it is useful to look at standardized residuals; defined as

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_i}}.$$

Outliers – Outlying Y

- A common way to detect outlying points in the Y direction is to look at the residuals.
- The residuals are also fundamental in looking at other assumptions of the model. However, the residuals $\hat{e}$ satisfy $var(\hat{e}) = \sigma^2(I - H)$. In particular, it is important to know that the residuals are correlated and have different variances.
- **Standardized Residuals** For diagnostics, it is useful to look at standardized residuals; defined as

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_i}}.$$

- Under the assumption of normality on $e_1, \ldots, e_n$, we know that the residuals $\hat{e} \sim N(0, \sigma^2(I - H))$.

Outliers – Outlying Y

- A common way to detect outlying points in the Y direction is to look at the residuals.
- The residuals are also fundamental in looking at other assumptions of the model. However, the residuals $\hat{e}$ satisfy $var(\hat{e}) = \sigma^2(I - H)$. In particular, it is important to know that the residuals are correlated and have different variances.
- **Standardized Residuals** For diagnostics, it is useful to look at standardized residuals; defined as

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_i}}.$$

- Under the assumption of normality on $e_1, \ldots, e_n$, we know that the residuals $\hat{e} \sim N(0, \sigma^2(I - H))$. Does the standardized residual r_i have a t-distribution?

Outliers – Outlying Y

- A common way to detect outlying points in the Y direction is to look at the residuals.
- The residuals are also fundamental in looking at other assumptions of the model. However, the residuals $\hat{e}$ satisfy $var(\hat{e}) = \sigma^2(I - H)$. In particular, it is important to know that the residuals are correlated and have different variances.
- **Standardized Residuals** For diagnostics, it is useful to look at standardized residuals; defined as

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_i}}.$$

- Under the assumption of normality on $e_1, \ldots, e_n$, we know that the residuals $\hat{e} \sim N(0, \sigma^2(I - H))$. Does the standardized residual r_i have a t-distribution?
- NO! because $\hat{e}_i$ and $\hat{\sigma}$ are not independent.

- Generally, values greater than say 3 are large;

- Generally, values greater than say 3 are large; those greater than two somewhat large (but 5% should be beyond two).

- Generally, values greater than say 3 are large; those greater than two somewhat large (but 5% should be beyond two). Formally, we'd like to say this based on a test, e.g. a t-test.

- Generally, values greater than say 3 are large; those greater than two somewhat large (but 5% should be beyond two). Formally, we'd like to say this based on a test, e.g. a t-test.
- Also in general, when $|e_i|$, $\hat{\sigma}^2$ tends to be large as well. Outliers will inflate $\hat{\sigma}$.

- Generally, values greater than say 3 are large; those greater than two somewhat large (but 5% should be beyond two). Formally, we'd like to say this based on a test, e.g. a t-test.
- Also in general, when $|e_i|$, $\hat{\sigma}^2$ tends to be large as well. Outliers will inflate $\hat{\sigma}$.
- **Predicted Residuals (Leave-one-out analysis)**
  One elegant way to get around the correlation of our error estimate with the residuals is by successively running analyses where we leave out an observation and run a regression without the observation.

- Generally, values greater than say 3 are large; those greater than two somewhat large (but 5% should be beyond two). Formally, we'd like to say this based on a test, e.g. a t-test.
- Also in general, when $|e_i|$, $\hat{\sigma}^2$ tends to be large as well. Outliers will inflate $\hat{\sigma}$.
- **Predicted Residuals (Leave-one-out analysis)**
  One elegant way to get around the correlation of our error estimate with the residuals is by successively running analyses where we leave out an observation and run a regression without the observation.
- These are useful to measure the *influence* of the *i*th subject on the regression line.

- Generally, values greater than say 3 are large; those greater than two somewhat large (but 5% should be beyond two). Formally, we'd like to say this based on a test, e.g. a t-test.
- Also in general, when $|e_i|$, $\hat{\sigma}^2$ tends to be large as well. Outliers will inflate $\hat{\sigma}$.
- **Predicted Residuals (Leave-one-out analysis)**
  One elegant way to get around the correlation of our error estimate with the residuals is by successively running analyses where we leave out an observation and run a regression without the observation.
- These are useful to measure the *influence* of the *i*th subject on the regression line.
- The *i*th predicted residual is defined as follows. First throw away the *i*th subject and fit the linear model.

- Using that linear model, predict the value of y_i based on x_i.

- Using that linear model, predict the value of y_i based on x_i.
- The difference between y_i and this predicted value is called the ith predicted residual.

- Using that linear model, predict the value of y_i based on x_i.
- The difference between y_i and this predicted value is called the ith predicted residual.
- Let $X_{[i]}$ denote the X-matrix with the ith row deleted.

- Using that linear model, predict the value of y_i based on x_i.
- The difference between y_i and this predicted value is called the ith predicted residual.
- Let $X_{[i]}$ denote the X-matrix with the ith row deleted.
- Also, let $Y_{[i]}$ denote the Y-vector with the ith entry deleted.

- Using that linear model, predict the value of y_i based on x_i.
- The difference between y_i and this predicted value is called the ith predicted residual.
- Let $X_{[i]}$ denote the X-matrix with the ith row deleted.
- Also, let $Y_{[i]}$ denote the Y-vector with the ith entry deleted.
- And let x_i^T denote the ith row of the original X matrix.

- Using that linear model, predict the value of y_i based on x_i.
- The difference between y_i and this predicted value is called the ith predicted residual.
- Let $X_{[i]}$ denote the X-matrix with the ith row deleted.
- Also, let $Y_{[i]}$ denote the Y-vector with the ith entry deleted.
- And let x_i^T denote the ith row of the original X matrix.
- The estimate of β after deleting the ith row is:

$$\hat{\beta}_{[i]} = \left(X_{[i]}^T X_{[i]} \right)^{-1} X_{[i]}^T Y_{[i]}.$$

- Using that linear model, predict the value of y_i based on x_i.
- The difference between y_i and this predicted value is called the ith predicted residual.
- Let $X_{[i]}$ denote the X-matrix with the ith row deleted.
- Also, let $Y_{[i]}$ denote the Y-vector with the ith entry deleted.
- And let x_i^T denote the ith row of the original X matrix.
- The estimate of β after deleting the ith row is:

$$\hat{\beta}_{[i]} = \left(X_{[i]}^T X_{[i]} \right)^{-1} X_{[i]}^T Y_{[i]}.$$

- The ith predicted residual is defined as

$$\hat{e}_{[i]} = y_i - x_i^T \hat{\beta}_{[i]}.$$

- Under the assumptions of the linear model (i.e., under $Y = X\beta + e$ with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I$), what are $\mathbb{E}\hat{e}_{[i]}$ and $var(\hat{e}_{[i]})$?

- Under the assumptions of the linear model (i.e., under $Y = X\beta + e$ with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I$), what are $\mathbb{E}\hat{e}_{[i]}$ and $var(\hat{e}_{[i]})$?
- It is easy to check that $\mathbb{E}\hat{e}_{[i]} = 0$.

- Under the assumptions of the linear model (i.e., under $Y = X\beta + e$ with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I$), what are $\mathbb{E}\hat{e}_{[i]}$ and $var(\hat{e}_{[i]})$?
- It is easy to check that $\mathbb{E}\hat{e}_{[i]} = 0$.
- For the variance, note that y_i and $\hat{\beta}_{[i]}$ are uncorrelated.

- Under the assumptions of the linear model (i.e., under $Y = X\beta + e$ with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I$), what are $\mathbb{E}\hat{e}_{[i]}$ and $var(\hat{e}_{[i]})$?
- It is easy to check that $\mathbb{E}\hat{e}_{[i]} = 0$.
- For the variance, note that y_i and $\hat{\beta}_{[i]}$ are uncorrelated.
- Therefore,

- Under the assumptions of the linear model (i.e., under $Y = X\beta + e$ with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I$), what are $\mathbb{E}\hat{e}_{[i]}$ and $var(\hat{e}_{[i]})$?
- It is easy to check that $\mathbb{E}\hat{e}_{[i]} = 0$.
- For the variance, note that y_i and $\hat{\beta}_{[i]}$ are uncorrelated.
- Therefore,

$$var(\hat{e}_{[i]}) = var(y_i) + var(x_i^T \hat{\beta}_{[i]})$$

- Under the assumptions of the linear model (i.e., under $Y = X\beta + e$ with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I$), what are $\mathbb{E}\hat{e}_{[i]}$ and $var(\hat{e}_{[i]})$?
- It is easy to check that $\mathbb{E}\hat{e}_{[i]} = 0$.
- For the variance, note that y_i and $\hat{\beta}_{[i]}$ are uncorrelated.
- Therefore,

$$
\begin{aligned}
var(\hat{e}_{[i]}) &= var(y_i) + var(x_i^T \hat{\beta}_{[i]}) \\
&= \sigma^2 + x_i^T Cov(\hat{\beta}_{[i]}) x_i
\end{aligned}
$$

- Under the assumptions of the linear model (i.e., under $Y = X\beta + e$ with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I$), what are $\mathbb{E}\hat{e}_{[i]}$ and $var(\hat{e}_{[i]})$?
- It is easy to check that $\mathbb{E}\hat{e}_{[i]} = 0$.
- For the variance, note that y_i and $\hat{\beta}_{[i]}$ are uncorrelated.
- Therefore,

$$\begin{aligned} var(\hat{e}_{[i]}) &= var(y_i) + var(x_i^T \hat{\beta}_{[i]}) \\ &= \sigma^2 + x_i^T Cov(\hat{\beta}_{[i]}) x_i \\ &= \sigma^2 + \sigma^2 x_i^T (X_{[i]}^T X_{[i]})^{-1} x_i \end{aligned}$$

- Under the assumptions of the linear model (i.e., under $Y = X\beta + e$ with $\mathbb{E}e = 0$ and $Cov(e) = \sigma^2 I$), what are $\mathbb{E}\hat{e}_{[i]}$ and $var(\hat{e}_{[i]})$?
- It is easy to check that $\mathbb{E}\hat{e}_{[i]} = 0$.
- For the variance, note that y_i and $\hat{\beta}_{[i]}$ are uncorrelated.
- Therefore,

$$\begin{aligned} var(\hat{e}_{[i]}) &= var(y_i) + var(x_i^T \hat{\beta}_{[i]}) \\ &= \sigma^2 + x_i^T Cov(\hat{\beta}_{[i]}) x_i \\ &= \sigma^2 + \sigma^2 x_i^T (X_{[i]}^T X_{[i]})^{-1} x_i \\ &= \sigma^2 \left(1 + x_i^T (X_{[i]}^T X_{[i]})^{-1} x_i\right). \end{aligned}$$

- It might seem that to calculate $\hat{e}_{[i]}$ for different i, one would need to perform many regressions deleting each subject separately.

- It might seem that to calculate $\hat{e}_{[i]}$ for different i, one would need to perform many regressions deleting each subject separately.
- Fortunately, one can calculate these in a simpler way using the following formula from matrix algebra:

- It might seem that to calculate $\hat{e}_{[i]}$ for different i, one would need to perform many regressions deleting each subject separately.
- Fortunately, one can calculate these in a simpler way using the following formula from matrix algebra:

Theorem

(Woodbury matrix identity) *Suppose A is an $n \times n$ matrix and a and b are $n \times m$ matrices, then*

$$(A - ab^T)^{-1} = A^{-1} + A^{-1}a\left(I_m - b^T A^{-1} a\right)^{-1} b^T A^{-1} \tag{2}$$

provided all the inverses above make sense.

- Because $X^T X = X_{[i]}^T X_{[i]} + x_i x_i^T$, we get from (2) that

- Because $X^T X = X_{[i]}^T X_{[i]} + x_i x_i^T$, we get from (2) that

$$(X_{[i]}^T X_{[i]})^{-1} = (X^T X - x_i x_i^T)^{-1}$$

- Because $X^T X = X_{[i]}^T X_{[i]} + x_i x_i^T$, we get from (2) that

$$\begin{aligned} (X_{[i]}^T X_{[i]})^{-1} &= (X^T X - x_i x_i^T)^{-1} \\ &= (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \end{aligned}$$

- Because $X^T X = X_{[i]}^T X_{[i]} + x_i x_i^T$, we get from (2) that

$$\begin{aligned}
(X_{[i]}^T X_{[i]})^{-1} &= (X^T X - x_i x_i^T)^{-1} \\
&= (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \\
&= (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - h_i}.
\end{aligned}$$

- Because $X^TX = X_{[i]}^TX_{[i]} + x_ix_i^T$, we get from (2) that

$$
\begin{aligned}
(X_{[i]}^TX_{[i]})^{-1} &= (X^TX - x_ix_i^T)^{-1} \\
&= (X^TX)^{-1} + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1-x_i^T(X^TX)^{-1}x_i} \\
&= (X^TX)^{-1} + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1-h_i}.
\end{aligned}
$$

- Also check that

$$X_{[i]}^TY_{[i]} = X^TY - y_ix_i.$$

Therefore

$$\hat{\beta}_{[i]} = (X_{[i]}^T X_{[i]})^{-1} X_{[i]}^T Y_{[i]}$$

Therefore

$$\begin{aligned}
\hat{\beta}_{[i]} &= (X_{[i]}^T X_{[i]})^{-1} X_{[i]}^T Y_{[i]} \\
&= \left[(X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i}\right] \left[X^T Y - y_i x_i\right]
\end{aligned}$$

Therefore

$$
\begin{aligned}
\hat{\beta}_{[i]} &= (X_{[i]}^T X_{[i]})^{-1} X_{[i]}^T Y_{[i]} \\
&= \left[(X^T X)^{-1} + \tfrac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i}\right] \left[X^T Y - y_i x_i\right] \\
&= \hat{\beta} + \tfrac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i} X^T Y - y_i (X^T X)^{-1} x_i
\end{aligned}
$$

Therefore

$$\begin{aligned}
\hat{\beta}_{[i]} &= (X_{[i]}^T X_{[i]})^{-1} X_{[i]}^T Y_{[i]} \\
&= \left[(X^T X)^{-1} + \tfrac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i}\right] \left[X^T Y - y_i x_i\right] \\
&= \hat{\beta} + \tfrac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i} X^T Y - y_i (X^T X)^{-1} x_i \\
&\quad - y_i \tfrac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i} x_i
\end{aligned}$$

Therefore

$$
\begin{aligned}
\hat{\beta}_{[i]} &= (X_{[i]}^T X_{[i]})^{-1} X_{[i]}^T Y_{[i]} \\
&= \left[(X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i}\right] \left[X^T Y - y_i x_i\right] \\
&= \hat{\beta} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i} X^T Y - y_i (X^T X)^{-1} x_i \\
&\quad - y_i \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i} x_i \\
&= \hat{\beta} + \frac{(X^T X)^{-1} x_i x_i^T \hat{\beta}}{1-h_i} - y_i (X^T X)^{-1} x_i - y_i \frac{(X^T X)^{-1} x_i h_i}{1-h_i}
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\hat{\beta}_{[i]} &= (X_{[i]}^T X_{[i]})^{-1} X_{[i]}^T Y_{[i]} \\
&= \left[(X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i}\right] \left[X^T Y - y_i x_i\right] \\
&= \hat{\beta} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i} X^T Y - y_i (X^T X)^{-1} x_i \\
&\quad - y_i \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i} x_i \\
&= \hat{\beta} + \frac{(X^T X)^{-1} x_i x_i^T \hat{\beta}}{1-h_i} - y_i (X^T X)^{-1} x_i - y_i \frac{(X^T X)^{-1} x_i h_i}{1-h_i} \\
&= \hat{\beta} + \left[\frac{x_i^T \hat{\beta}}{1-h_i} - \frac{y_i (1-h_i)}{1-h_i} - \frac{y_i h_i}{1-h_i}\right] (X^T X)^{-1} x_i
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\hat{\beta}_{[i]} &= (X_{[i]}^T X_{[i]})^{-1} X_{[i]}^T Y_{[i]} \\
&= \left[(X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i} \right] \left[X^T Y - y_i x_i \right] \\
&= \hat{\beta} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i} X^T Y - y_i (X^T X)^{-1} x_i \\
&\quad - y_i \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1-h_i} x_i \\
&= \hat{\beta} + \frac{(X^T X)^{-1} x_i x_i^T \hat{\beta}}{1-h_i} - y_i (X^T X)^{-1} x_i - y_i \frac{(X^T X)^{-1} x_i h_i}{1-h_i} \\
&= \hat{\beta} + \left[\frac{x_i^T \hat{\beta}}{1-h_i} - \frac{y_i (1-h_i)}{1-h_i} - \frac{y_i h_i}{1-h_i} \right] (X^T X)^{-1} x_i \\
&= \hat{\beta} - \frac{\hat{e}_i}{1-h_i} (X^T X)^{-1} x_i .
\end{aligned}
$$

- As a result

$$\hat{e}_{[i]} = y_i - x_i^T \hat{\beta}_{[i]}$$

- As a result

$$
\begin{aligned}
\hat{e}_{[i]} &= y_i - x_i^T \hat{\beta}_{[i]} \\
&= y_i - x_i^T \hat{\beta} + \frac{\hat{e}_i}{1-h_i} x_i^T (X^T X)^{-1} x_i
\end{aligned}
$$

- As a result

$$\begin{aligned} \hat{e}_{[i]} &= y_i - x_i^T \hat{\beta}_{[i]} \\ &= y_i - x_i^T \hat{\beta} + \tfrac{\hat{e}_i}{1-h_i} x_i^T (X^T X)^{-1} x_i \\ &= \hat{e}_i + \tfrac{h_i}{1-h_i} \hat{e}_i \end{aligned}$$

- As a result

$$
\begin{aligned}
\hat{e}_{[i]} &= y_i - x_i^T \hat{\beta}_{[i]} \\
&= y_i - x_i^T \hat{\beta} + \tfrac{\hat{e}_i}{1-h_i} x_i^T (X^T X)^{-1} x_i \\
&= \hat{e}_i + \tfrac{h_i}{1-h_i} \hat{e}_i \\
&= \tfrac{\hat{e}_i}{1-h_i}.
\end{aligned}
$$

- Therefore, the predicted residual $\hat{e}_{[i]}$ is the usual residual divided by 1 minus the leverage.

- As a result

$$
\begin{aligned}
\hat{e}_{[i]} &= y_i - x_i^T \hat{\beta}_{[i]} \\
&= y_i - x_i^T \hat{\beta} + \tfrac{\hat{e}_i}{1-h_i} x_i^T (X^T X)^{-1} x_i \\
&= \hat{e}_i + \tfrac{h_i}{1-h_i} \hat{e}_i \\
&= \tfrac{\hat{e}_i}{1-h_i}.
\end{aligned}
$$

- Therefore, the predicted residual $\hat{e}_{[i]}$ is the usual residual divided by 1 minus the leverage.
- One can thus see, if the leverage of the *i*th subject, h_i, is very large, then the residual $\hat{e}_i$ will be small, but the predicted residual might be very large.