Lecture 16

October 15, 2018

Predicted Residuals

Predicted Residuals

- Predicted residuals are useful for detecting outliers. How does one detect outliers?

Predicted Residuals

- Predicted residuals are useful for detecting outliers. How does one detect outliers?
- A first answer might be to look for subjects having large residuals.

Predicted Residuals

- Predicted residuals are useful for detecting outliers. How does one detect outliers?
- A first answer might be to look for subjects having large residuals.
- But the problem with this approach is that when the outlier also has a large leverage, then the residual will not be that large.

Predicted Residuals

- Predicted residuals are useful for detecting outliers. How does one detect outliers?
- A first answer might be to look for subjects having large residuals.
- But the problem with this approach is that when the outlier also has a large leverage, then the residual will not be that large.
- Therefore, one needs to look at a combination of leverage and the value of the residual. It turns out that predicted residuals are a natural way of combining the residuals and the leverages.

Predicted Residuals

- Predicted residuals are useful for detecting outliers. How does one detect outliers?
- A first answer might be to look for subjects having large residuals.
- But the problem with this approach is that when the outlier also has a large leverage, then the residual will not be that large.
- Therefore, one needs to look at a combination of leverage and the value of the residual. It turns out that predicted residuals are a natural way of combining the residuals and the leverages.
- The ith predicted residual is defined as follows. First throw away the ith subject and fit the linear model. Using that linear model, predict the value of y_i based on the explanatory variable values of the ith subject. The difference between y_i and this predicted value is called the ith predicted residual.

Predicted Residuals

- 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.

Predicted Residuals

- 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:

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

Predicted Residuals

- 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

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

Predicted Residuals

- 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

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

- It might seem that to calculate $\hat{e}_{[i]}$ for different i, one would need to perform many regressions deleting each subject separately.Fortunately, there is a simple formula for $\hat{e}_{[i]}$ in terms of h_{ii} and $\hat{e}_i$:

$$\hat{e}_{[i]} = \frac{\hat{e}_i}{1 - h_{ii}} \tag{1}$$

Predicted Residuals

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

Predicted Residuals

- 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 ith subject, h_{ii}, is very large, then the residual $\hat{e}_i$ will be small, but the predicted residual might be very large.

Predicted Residuals

- 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 ith subject, h_{ii}, is very large, then the residual $\hat{e}_i$ will be small, but the predicted residual might be very large.
- 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 $var(\hat{e}_{[i]})$?

Predicted Residuals

- 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 ith subject, h_{ii}, is very large, then the residual $\hat{e}_i$ will be small, but the predicted residual might be very large.
- 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 $\mathbb{E}\hat{e}_{[i]} = 0$.

Predicted Residuals

- 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 ith subject, h_{ii}, is very large, then the residual $\hat{e}_i$ will be small, but the predicted residual might be very large.
- 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,

$$var(\hat{e}_{[i]}) = var\left(\frac{\hat{e}_i}{1 - h_{ii}}\right) = (1 - h_{ii})^{-2} var(\hat{e}_i) = \frac{\sigma^2}{1 - h_{ii}}.$$

Predicted Residuals

- The predicted residuals therefore have different variances (if h_{ii} is large, then $\hat{e}_{[i]}$ has high variance).

Predicted Residuals

- The predicted residuals therefore have different variances (if h_{ii} is large, then $\hat{e}_{[i]}$ has high variance). We can thus standardize them and we get what are called Standardized Predicted Residuals. These are described next.

Standardized Predicted Residuals

These are also called Jackknife Residuals.

Standardized Predicted Residuals

These are also called Jackknife Residuals.

- How do we decide if a subject is an outlier?

Standardized Predicted Residuals

These are also called Jackknife Residuals.

- How do we decide if a subject is an outlier? One idea is to calculate its predicted residual $\hat{e}_{[i]}$ and check it is too high.

Standardized Predicted Residuals

These are also called Jackknife Residuals.

- How do we decide if a subject is an outlier? One idea is to calculate its predicted residual $\hat{e}_{[i]}$ and check it is too high. How do we decide if it is too high?

Standardized Predicted Residuals

These are also called Jackknife Residuals.

- How do we decide if a subject is an outlier? One idea is to calculate its predicted residual $\hat{e}_{[i]}$ and check it is too high. How do we decide if it is too high?
- We have just seen that the variance of the predicted residual is given by

$$var(\hat{e}_{[i]}) = \frac{\sigma^2}{1 - h_{ii}}.$$

Standardized Predicted Residuals

These are also called Jackknife Residuals.

- How do we decide if a subject is an outlier? One idea is to calculate its predicted residual $\hat{e}_{[i]}$ and check it is too high. How do we decide if it is too high?
- We have just seen that the variance of the predicted residual is given by

$$var(\hat{e}_{[i]}) = \frac{\sigma^2}{1 - h_{ii}}.$$

- When judging if $\hat{e}_{[i]}$ is high or not, we need to keep this variability in mind. It makes to consider a standardized version of the predicted residual where we divide it by its standard deviation:

$$\frac{\hat{e}_{[i]}\sqrt{1 - h_{ii}}}{\sigma}.$$

Standardized Predicted Residuals

- Because σ is unknown, we cannot compute these.

Standardized Predicted Residuals

- Because σ is unknown, we cannot compute these. So we need to estimate σ.

Standardized Predicted Residuals

- Because σ is unknown, we cannot compute these. So we need to estimate σ. We have two choices:

Standardized Predicted Residuals

- Because σ is unknown, we cannot compute these. So we need to estimate σ. We have two choices:
  1. We can use the Residual Standard Error in the full model: $\sqrt{RSS/(n-p-1)}$.

Standardized Predicted Residuals

- Because σ is unknown, we cannot compute these. So we need to estimate σ. We have two choices:
  1. We can use the Residual Standard Error in the full model: $\sqrt{RSS/(n-p-1)}$. We then get

$$\frac{\hat{e}_{[i]}\sqrt{1-h_{ii}}}{\sqrt{RSS/(n-p-1)}} = \frac{\hat{e}_i}{\sqrt{(1-h_{ii})RSS/(n-p-1)}}.$$

Standardized Predicted Residuals

- Because σ is unknown, we cannot compute these. So we need to estimate σ. We have two choices:
  1. We can use the Residual Standard Error in the full model: $\sqrt{RSS/(n-p-1)}$. We then get

$$\frac{\hat{e}_{[i]}\sqrt{1-h_{ii}}}{\sqrt{RSS/(n-p-1)}} = \frac{\hat{e}_i}{\sqrt{(1-h_{ii})RSS/(n-p-1)}}.$$

     This is nothing but the standardized residual r_i that we defined previously.

Standardized Predicted Residuals

- Because σ is unknown, we cannot compute these. So we need to estimate σ. We have two choices:
  1. We can use the Residual Standard Error in the full model: $\sqrt{RSS/(n-p-1)}$. We then get

$$\frac{\hat{e}_{[i]}\sqrt{1-h_{ii}}}{\sqrt{RSS/(n-p-1)}} = \frac{\hat{e}_i}{\sqrt{(1-h_{ii})RSS/(n-p-1)}}.$$

     This is nothing but the standardized residual r_i that we defined previously. Note that these do NOT have the t-distribution (under normality) because the numerator and denominator are not independent.

Standardized Predicted Residuals

- We can use the Residual Standard Error in the model obtained by dropping the *i*th subject.

Standardized Predicted Residuals

- We can use the Residual Standard Error in the model obtained by dropping the ith subject. We then get

$$t_i = \frac{\hat{e}_{[i]}\sqrt{1-h_{ii}}}{\sqrt{RSS_{[i]}/(n-p-2)}}.$$

Standardized Predicted Residuals

- We can use the Residual Standard Error in the model obtained by dropping the ith subject. We then get

$$t_i = \frac{\hat{e}_{[i]}\sqrt{1 - h_{ii}}}{\sqrt{RSS_{[i]}/(n-p-2)}}.$$

Now it can be checked that the numerator and the denominator are independent (why?) which implies that t_i has the t-distribution with $n - p - 2$ degrees of freedom.

Standardized Predicted Residuals

- We can use the Residual Standard Error in the model obtained by dropping the ith subject. We then get

$$t_i = \frac{\hat{e}_{[i]}\sqrt{1-h_{ii}}}{\sqrt{RSS_{[i]}/(n-p-2)}}.$$

Now it can be checked that the numerator and the denominator are independent (why?) which implies that t_i has the t-distribution with $n-p-2$ degrees of freedom. These quantities $\{t_i\}$ are called standardized predicted residuals or externally studentized residuals.

Standardized Predicted Residuals

- Under the assumption that the linear model is true and normality of the errors, the standardized predicted residual t_i has the $t(n-p-2)$ distribution.

Standardized Predicted Residuals

- Under the assumption that the linear model is true and normality of the errors, the standardized predicted residual t_i has the $t(n-p-2)$ distribution.
- This can therefore be used to assess whether the ith subject is an outlier.

Standardized Predicted Residuals

- Under the assumption that the linear model is true and normality of the errors, the standardized predicted residual t_i has the $t(n-p-2)$ distribution.
- This can therefore be used to assess whether the ith subject is an outlier.
- Indeed, one can conduct a formal test of whether the ith subject is an outlier or not (what are H_0 and H_1 here?) by looking at t_i and rejecting the null (i.e., declaring ith subject as the outlier) if it is larger in absolute than the $\alpha/2$ critical value for the $t(n-p-2)$ distribution.

- One has to be careful with this above test.

- One has to be careful with this above test.
- Suppose that the standardized predicted residuals are $t_1, \ldots, t_n$.

- One has to be careful with this above test.
- Suppose that the standardized predicted residuals are $t_1, \ldots, t_n$.
- Suppose also that the model is true so that $t_1, \ldots, t_n$ have the $t(n-p-2)$ distribution. They are not necessarily independent however. But just for the sake of the discussion, assume that they are roughly independent.

- One has to be careful with this above test.
- Suppose that the standardized predicted residuals are $t_1, \ldots, t_n$.
- Suppose also that the model is true so that $t_1, \ldots, t_n$ have the $t(n-p-2)$ distribution. They are not necessarily independent however. But just for the sake of the discussion, assume that they are roughly independent.
- Then we would expect about $n\alpha$ number of $t_1, \ldots, t_n$ to be larger in absolute value than the $\alpha/2$ critical value of the $t(n-p-2)$. Therefore, we would be tagging $n\alpha$ of the subjects as outliers even when there are no outliers in the data.

- One has to be careful with this above test.
- Suppose that the standardized predicted residuals are $t_1, \dots, t_n$.
- Suppose also that the model is true so that $t_1, \dots, t_n$ have the $t(n-p-2)$ distribution. They are not necessarily independent however. But just for the sake of the discussion, assume that they are roughly independent.
- Then we would expect about $n\alpha$ number of $t_1, \dots, t_n$ to be larger in absolute value than the $\alpha/2$ critical value of the $t(n-p-2)$. Therefore, we would be tagging $n\alpha$ of the subjects as outliers even when there are no outliers in the data.
- To counter this, one takes a value of α much smaller than 0.05 while testing whether the ith observation is an outlier or not.

- One has to be careful with this above test.
- Suppose that the standardized predicted residuals are $t_1, \dots, t_n$.
- Suppose also that the model is true so that $t_1, \dots, t_n$ have the $t(n-p-2)$ distribution. They are not necessarily independent however. But just for the sake of the discussion, assume that they are roughly independent.
- Then we would expect about $n\alpha$ number of $t_1, \dots, t_n$ to be larger in absolute value than the $\alpha/2$ critical value of the $t(n-p-2)$. Therefore, we would be tagging $n\alpha$ of the subjects as outliers even when there are no outliers in the data.
- To counter this, one takes a value of α much smaller than 0.05 while testing whether the ith observation is an outlier or not.
- A particularly conservative value is $\alpha = 0.05/n$. In this case, one can show that the probability that at least one subject is tagged an outlier when in fact there are none is atmost 0.05. This is known as the Bonferroni correction.

Cook's Distance

Cook's Distance

- This measures the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$.

Cook's Distance

- This measures the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$.
- Recall from the last section that $\hat{\beta}_{[i]}$ is the estimate of β with the *i*th subject removed.

Cook's Distance

- This measures the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$.
- Recall from the last section that $\hat{\beta}_{[i]}$ is the estimate of β with the *i*th subject removed.
- A naive way of measuring the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$ is to use the Euclidean distance: $(\hat{\beta} - \hat{\beta}_{[i]})^T(\hat{\beta} - \hat{\beta}_{[i]})$.

Cook's Distance

- This measures the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$.
- Recall from the last section that $\hat{\beta}_{[i]}$ is the estimate of β with the *i*th subject removed.
- A naive way of measuring the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$ is to use the Euclidean distance: $(\hat{\beta} - \hat{\beta}_{[i]})^T(\hat{\beta} - \hat{\beta}_{[i]})$.
- But this ignores the fact that the different elements of $\hat{\beta}$ have different variances.

Cook's Distance

- This measures the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$.
- Recall from the last section that $\hat{\beta}_{[i]}$ is the estimate of β with the *i*th subject removed.
- A naive way of measuring the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$ is to use the Euclidean distance: $(\hat{\beta} - \hat{\beta}_{[i]})^T(\hat{\beta} - \hat{\beta}_{[i]})$.
- But this ignores the fact that the different elements of $\hat{\beta}$ have different variances.
- Because $Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$, it makes sense to use the idea underlying Mahalanobis distance to measure the distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$ by

$$\frac{(\hat{\beta} - \hat{\beta}_{[i]})^T X^T X(\hat{\beta} - \hat{\beta}_{[i]})}{\sigma^2}. \tag{2}$$

Cook's Distance

- Because σ is unknown, we can estimate it by the Residual Standard Error, $\hat{\sigma}$. This gives us the notion of Cook's distance:

$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{[i]})^T X^T X (\hat{\beta} - \hat{\beta}_{[i]})}{(p+1)\hat{\sigma}^2}. \tag{3}$$

Cook's Distance

- Because σ is unknown, we can estimate it by the Residual Standard Error, $\hat{\sigma}$. This gives us the notion of Cook's distance:
$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{[i]})^T X^T X (\hat{\beta} - \hat{\beta}_{[i]})}{(p+1)\hat{\sigma}^2}. \tag{3}$$
- Note that there is a division by $p+1$ above which was not there in (2). This does not really matter (as the division is the same for all i). I have kept the divisor because this is the standard way of defining the Cook's distance.

Cook's Distance

- Because σ is unknown, we can estimate it by the Residual Standard Error, $\hat{\sigma}$. This gives us the notion of Cook's distance:
$$C_i = \frac{(\hat{\beta} - \hat{\beta}_{[i]})^T X^T X (\hat{\beta} - \hat{\beta}_{[i]})}{(p+1)\hat{\sigma}^2}. \tag{3}$$
- Note that there is a division by $p+1$ above which was not there in (2). This does not really matter (as the division is the same for all i). I have kept the divisor because this is the standard way of defining the Cook's distance.
- It is possible to give an alternative expression for C_i that depends only on r_i and h_{ii}:
$$C_i = r_i^2 \frac{h_{ii}}{(1-h_{ii})(p+1)}.$$
Note therefore C_i depends only on r_i and the leverage h_{ii}. If r_i^2 is large and/or h_{ii} is large, then C_i will be large.

Detecting Nonlinearity

Detecting Nonlinearity

- We will continue our discussion of regression diagnostics by looking at detecting nonlinearity.

Detecting Nonlinearity

- We will continue our discussion of regression diagnostics by looking at detecting nonlinearity.
- The standard plots for this are (a) plot the response values y against each explanatory variable values, and (b) plot the residuals $\hat{e}$ against each explanatory variable values.

Detecting Nonlinearity

- We will continue our discussion of regression diagnostics by looking at detecting nonlinearity.
- The standard plots for this are (a) plot the response values y against each explanatory variable values, and (b) plot the residuals $\hat{e}$ against each explanatory variable values.
- The problem with these plots however is that they only look at the marginal effect of the ith explanatory variable on y and ignore the presence of the other explanatory variables. To correct this, one often looks at partial regression plots (also called added variable plots) and partial residual plots.

- Let $Res(y, X^{-i})$ denote the vector of residuals obtained by regressing y on all the explanatory variables *except* the ith one.

- Let $Res(y, X^{-i})$ denote the vector of residuals obtained by regressing y on all the explanatory variables *except* the ith one.
- Also let $Res(x_i, X^{-i})$ denote the vector of residuals obtained by regressing x_i (x_i is the column of the X-matrix corresponding to the ith explanatory variable) on all the explanatory variables *except* the ith one.

- Let $Res(y, X^{-i})$ denote the vector of residuals obtained by regressing y on all the explanatory variables *except* the ith one.
- Also let $Res(x_i, X^{-i})$ denote the vector of residuals obtained by regressing x_i (x_i is the column of the X-matrix corresponding to the ith explanatory variable) on all the explanatory variables *except* the ith one.
- In the ith partial regression plot, one plots $Res(y, X^{-i})$ against $Res(x_i, X^{-i})$.

- Let $Res(y, X^{-i})$ denote the vector of residuals obtained by regressing y on all the explanatory variables *except* the ith one.
- Also let $Res(x_i, X^{-i})$ denote the vector of residuals obtained by regressing x_i (x_i is the column of the X-matrix corresponding to the ith explanatory variable) on all the explanatory variables *except* the ith one.
- In the ith partial regression plot, one plots $Res(y, X^{-i})$ against $Res(x_i, X^{-i})$.
- This plot therefore looks at the relationship between y and x_i in the presence of all the other explanatory variables.

- A remarkable feature of the *i*th partial regression plot is that if one performs a regression of $Res(y, X^{-i})$ against $Res(x_i, X^{-i})$, one gets the intercept estimate to be zero and the slope estimate will exactly equal $\hat{\beta}_i$. This fact can be proved for instance using the block matrix inverse formula (see wikipedia for this formula).

- A remarkable feature of the *i*th partial regression plot is that if one performs a regression of $Res(y, X^{-i})$ against $Res(x_i, X^{-i})$, one gets the intercept estimate to be zero and the slope estimate will exactly equal $\hat{\beta}_i$. This fact can be proved for instance using the block matrix inverse formula (see wikipedia for this formula).
- The partial residual plot is another plot for viewing the relationship between y and x_i in the presence of the other variables.

- A remarkable feature of the ith partial regression plot is that if one performs a regression of $Res(y, X^{-i})$ against $Res(x_i, X^{-i})$, one gets the intercept estimate to be zero and the slope estimate will exactly equal $\hat{\beta}_i$. This fact can be proved for instance using the block matrix inverse formula (see wikipedia for this formula).
- The partial residual plot is another plot for viewing the relationship between y and x_i in the presence of the other variables.
- This simply plots $y - \sum_{j \neq i} \hat{\beta}_j x_j$ against x_i.

- A remarkable feature of the ith partial regression plot is that if one performs a regression of $Res(y, X^{-i})$ against $Res(x_i, X^{-i})$, one gets the intercept estimate to be zero and the slope estimate will exactly equal $\hat{\beta}_i$. This fact can be proved for instance using the block matrix inverse formula (see wikipedia for this formula).
- The partial residual plot is another plot for viewing the relationship between y and x_i in the presence of the other variables.
- This simply plots $y - \sum_{j \neq i} \hat{\beta}_j x_j$ against x_i.
- Because $y = \hat{y} + \hat{e} = \sum_j \hat{\beta}_j x_j + \hat{e}$, one can alternately describe the partial residual plot as plotting $\hat{e} + x_i \hat{\beta}_i$ against x_i.

- A remarkable feature of the ith partial regression plot is that if one performs a regression of $Res(y, X^{-i})$ against $Res(x_i, X^{-i})$, one gets the intercept estimate to be zero and the slope estimate will exactly equal $\hat{\beta}_i$. This fact can be proved for instance using the block matrix inverse formula (see wikipedia for this formula).
- The partial residual plot is another plot for viewing the relationship between y and x_i in the presence of the other variables.
- This simply plots $y - \sum_{j \neq i} \hat{\beta}_j x_j$ against x_i.
- Because $y = \hat{y} + \hat{e} = \sum_j \hat{\beta}_j x_j + \hat{e}$, one can alternately describe the partial residual plot as plotting $\hat{e} + x_i \hat{\beta}_i$ against x_i.
- This also has the property that if one were to do simple regression, the fitted slope with be precisely $\hat{\beta}_i$.