

ETC3580: Advanced Statistical Modelling

Week 2: Linear model diagnostics

Outline

1 Regression diagnostics

2 Influence

3 R functions

Regression diagnostics

Plots that should be checked

- Residuals vs Fitted: Check heteroskedasticity
- Residuals vs Predictors: Check for non-linearity
- Residuals vs Predictors not in model: Check for missing predictors
- Normal QQ plot: Check for non-normality
- Hat-values and Cooks distances: Check for influential points

Outline

1 Regression diagnostics

2 Influence

3 R functions

LOO Residuals

Fitted values

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

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the “hat matrix”.

LOO Residuals

Fitted values

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

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the “hat matrix”.

Theorem: Leave-one-out residuals

Let h_1, \dots, h_n be the diagonal values of \mathbf{H} . Then $e_{(i)} = e_i / (1 - h_i)$ is the prediction error that would be obtained if the i th observation was omitted.

LOO Residuals

Fitted values

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

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the “hat matrix”.

Theorem: Leave-one-out residuals

Let h_1, \dots, h_n be the diagonal values of \mathbf{H} . Then $e_{(i)} = e_i / (1 - h_i)$ is the prediction error that would be obtained if the i th observation was omitted.

Leverage values

h_i is called the “leverage” of observation i .

In R: `hatvalues(fit)` or `broom::augment(fit)`

LOO Residuals

Let $\mathbf{X}_{[i]}$ and $\mathbf{Y}_{[i]}$ be similar to \mathbf{X} and \mathbf{Y} but with the i th row deleted in each case. Let \mathbf{x}'_i be the i th row of \mathbf{X} and let

$$\hat{\beta}_{[i]} = (\mathbf{X}'_{[i]}\mathbf{X}_{[i]})^{-1}\mathbf{X}'_{[i]}\mathbf{Y}_{[i]}$$

be the estimate of β without the i th case. Then

$$e_{[i]} = y_i - \mathbf{x}'_i\hat{\beta}_{[i]}.$$

Now $\mathbf{X}'_{[i]}\mathbf{X}_{[i]} = (\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}'_i)$ and $\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = h_i$.

LOO Residuals

Sherman-Morrison-Woodbury formula

Suppose \mathbf{A} is a square matrix, and \mathbf{u} and \mathbf{v} are column vectors of the same dimension. Then

$$(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}}.$$

So by SMW,

$$(\mathbf{X}'_{[i]}\mathbf{X}_{[i]})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_i}.$$

Also note that $\mathbf{X}'_{[i]}\mathbf{Y}_{[i]} = \mathbf{X}'\mathbf{Y} - \mathbf{x}_i y_i$.

LOO Residuals

Therefore

$$\hat{\beta}_{[i]} = \left[(\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_i} \right] (\mathbf{X}'\mathbf{Y} - \mathbf{x}_iy_i)$$

$$= \hat{\beta} - \left[\frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}{1 - h_i} \right] [y_i(1 - h_i) - \mathbf{x}_i'\hat{\beta} + h_iy_i]$$

$$= \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_ie_i/(1 - h_i)$$

Thus

$$\begin{aligned}e_{[i]} &= y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{[i]} \\&= y_i - \mathbf{x}'_i \left[\hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i e_i / (1 - h_i) \right] \\&= e_i + h_i e_i / (1 - h_i) \\&= e_i / (1 - h_i),\end{aligned}$$

LOO Residuals

Cross-validation statistic

$$CV = \frac{1}{T} \sum_{i=1}^n [e_i / (1 - h_i)]^2,$$

- Measures MSE of out-of-sample prediction
- Asymptotically equivalent to AIC (up to monotonic transformation)

Cook distances

$$D_i = \frac{e_i^2 h_i}{\hat{\sigma}^2 p(1 - h_i)}$$

- Measures change in fit if observation i dropped.

Outline

1 Regression diagnostics

2 Influence

3 R functions

Residual diagnostics

`broom::augment` computes

- fitted values
- `se(fit)`
- residuals
- hat values
- cooks distance
- standardized residuals

This does not allow for matrix inputs such as `poly(x,2)`

Other functions

- `dfbeta`: $\hat{\beta} - \hat{\beta}_{(i)}$