MONASH University

MONASH BUSINESS SCHOOL

ETC3580: Advanced Statistical Modelling

Week 2: Linear model diagnostics

Outline

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

LOO Residuals

Fitted values

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

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

LOO Residuals

Fitted values

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

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the "hat matrix".

Theorem: Leave-one-out residuals

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

LOO Residuals

Fitted values

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

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

Theorem: Leave-one-out residuals

Let $h_1, \dots, h_n$ be the diagonal values of $\boldsymbol{H}$. Then $e_{(i)} = e_i/(1 - h_i)$ is the prediction error that would be obtained if the ith 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 $\boldsymbol{X}_{[i]}$ and $\boldsymbol{Y}_{[i]}$ be similar to $\boldsymbol{X}$ and $\boldsymbol{Y}$ but with the ith row deleted in each case. Let $\boldsymbol{x}_i'$ be the ith row of $\boldsymbol{X}$ and let

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

be the estimate of β without the ith case. Then $e_{[i]} = y_i - \boldsymbol{x}_i'\hat{\beta}_{[i]}$.

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

LOO Residuals

Sherman-Morrison-Woodbury formula

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

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

So by SMW,

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

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

LOO Residuals

Therefore

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

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

$$= \hat{\beta} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i / (1 - h_i)$$

LOO Residuals

Thus

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

LOO Residuals

Cross-validation statistic

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

- 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

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