Tutorial 8

YANG YANG

The Australian National University

Week 9, 2017

Overview

Standardised residuals

- Internally studentised residuals: **rstandard()**
  $r_i = \frac{e_i}{s_\epsilon\sqrt{1-h_{ii}}} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$

- Externally studentised residuals: **rstudent()**
  $t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} = \frac{e_{i,-i}}{s_{-i}/\sqrt{1-h_{ii}}}$

Externally studentised residuals

Externally studentised residuals:

$$t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} = \frac{e_{i,-i}}{s_{-i}/\sqrt{1-h_{ii}}}$$

- Externally studentised residuals are used to construct the **main residual plot** which should look like a random (rectangular) scatter of points.
- Patterns and funnelling indicate potential non-linearity or heteroscedasticity, respectively.
- If the errors are truly normally distributed, then the studentised residual values should generally lie between -2 and 2, regardless of the ordinary residual scale, s.
  $\longrightarrow$ find potential outliers

Outliers

The two most common sources of outliers are:

1. There is a location (mean) shift at the i^{th} data point:
   $E(Y|X = x_i) = \beta x_i + \Delta$ so that $E(\epsilon_i) = \Delta_i \neq 0$

   vs

   $E(Y|X = x_i) = \beta x_i$
2. There is a scale shift at the i^{th} data point, so that $Var(\epsilon_i) > \sigma^2$

Location (mean) shift

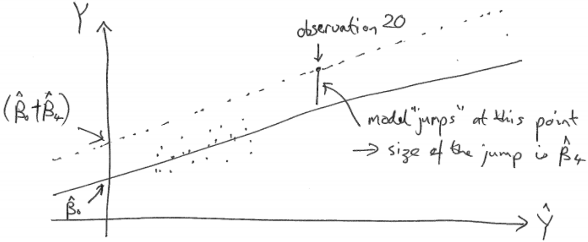
indicator variable $I_{20} = \begin{cases} 0 & \text{if } i = 1, 2, \ldots 19 \\ 1 & \text{if } i = 20 \end{cases}$

fitted model $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 I_{20}$

if $I_{20} = 0 \Rightarrow \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$

if $I_{20} = 1$ (ie observation 20)

model $\Rightarrow \hat{Y} = (\hat{\beta}_0 + \hat{\beta}_4) + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$



Hypothesis test for outliers

Externally studentised residuals:

$$t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} = \frac{e_{i,-i}}{s_{-i}/\sqrt{1-h_{ii}}}$$

- t_i follows a student's t distribution with $n-p-1$ degrees of freedom under assumption that the i^{th} data point does not suffer from a location shift
- $H_0 : \Delta_i = 0$ vs $H_A : \Delta_i \neq 0$
- **qt(0.975, df=error.df-1)**

Q1 (b) (c) and (d)

(b) calculate internally and externally Studentised residuals
- std.res < − data.frame(**cbind(int.stud=rstandard(msleep.lm), ext.stud=rstudent(msleep.lm)),** row.names=row.names(mammalsleep))
- std.res[**order(abs(std.res$int.stud)**, decreasing=T)[1:5],]

(c) cut-off value: **qt(0.975, df=msleep.lm$df-1)**

(d) msleep.loglm < − lm(log(brain) ∼ log(body))

Q2 (b) and (c)

(b) diagnostic plots

1. externally Studentised residuals vs fitted values
2. Normal Q-Q of internally Studentised residuals
3. bar plot of Cook's distances

(c) delete the outlier and refit the model
forbes.lm2 $<-$ lm(log(Pressure)[-12] $\sim$ Boiling.point[-12])