Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 19, part 3: Residual Diagnostics

Jinko Graham

2018-11-12

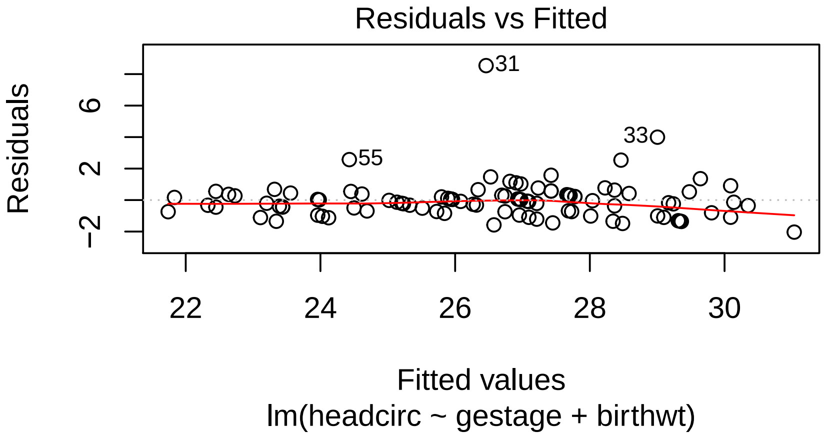
Residual Diagnostics

- Residuals are the main tool for checking model assumptions and identifying outliers
- Recall the model assumptions:
  1. The linear predictor is correctly specified.
  2. The random errors have constant SD.
  3. The random errors are normally distributed.
- The residuals are observed minus fitted values: $y_i - \hat{y}_i$,
- In Chapter 18, we plotted residuals *vs.* fitted values to check assumptions 1 and 2, and also to informally identify outliers.
- To check the normal errors assumption and detect outliers more formally, we'll define the *Q-Q plot* and *standardized residuals*.
- But first let's check assumptions 1 and 2 for the MLR model that we fit to low-birthweight-babies data, by plotting the residuals *vs.* fitted values.

Residuals versus Fitted Values

- Load the data, fit the MLR model and do the plot ...

```
uu <- url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_17/Data/lbwt.csv")
lbwt <- read.csv(uu)
lfit2 <- lm(headcirc ~ gestage + birthwt,data=lbwt)
plot(lfit2,which=1)
```



Comments

- There are no obvious missed trends. As far as we can tell, the linear predictor looks properly specified.
- There is no obvious funnel pattern in the residuals that might suggest that the error terms have non-constant SD.
- The 3 most extreme (farthest from zero) residuals are labelled by their case number. Case 31 in particular stands out.
- Note: Residual diagnostics can be subjective.
  - Whether or not a plot suggests that an assumption is violated can depend on the person looking at it.
  - My concern is that you understand which plots check which assumptions and that you can form an opinion about the assumptions.
  - Different people may have differing opinions.

Software Notes

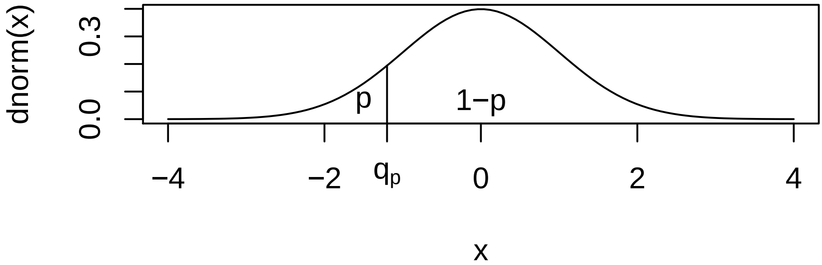
- Recall that R's `plot()` function can do six different diagnostic plots, specified by the `which` argument.
  - The first plot (`which=1`) is the residual *vs* fitted values.
  - The second plot (`which=2`) is the Q-Q plot which we haven't seen yet but which we will discuss next.
  - In this course, we won't be interested in the others.

Q-Q Plots

- A quantile-quantile (Q-Q) plot is a plot of the *quantiles* of one distribution *vs.* another.
  - If the two distributions have similar shape, the points on such a plot should fall roughly on a straight line.
  - We will define quantiles on the next slide.
- Our interest is in using Q-Q plots to compare the distribution of residuals to the distribution they *should have* under the model assumption of *normal random errors*.

Quantiles

- The pth quantile, q_p, of a distribution is the cutpoint such that the proportion p of the distribution is less than or equal to the cutpoint.



- Examples:
  1. The median is the 0.5 quantile, or $q_{.5}$, cutting the distribution into bottom and top halves
  2. The first quartile is the 0.25 quantile, or $q_{.25}$, cutting the distribution into the bottom quarter and the top three quarters

Distribution of Residuals

- We may *standardize* the residuals to have a common distribution. We'll skip the details.
- Under the model assumptions, the standardized residuals have a t distribution with $n - q - 1$ df.
- **Rule of thumb**: Standardized residuals less than -3 or greater than 3 are considered to be obvious outliers.

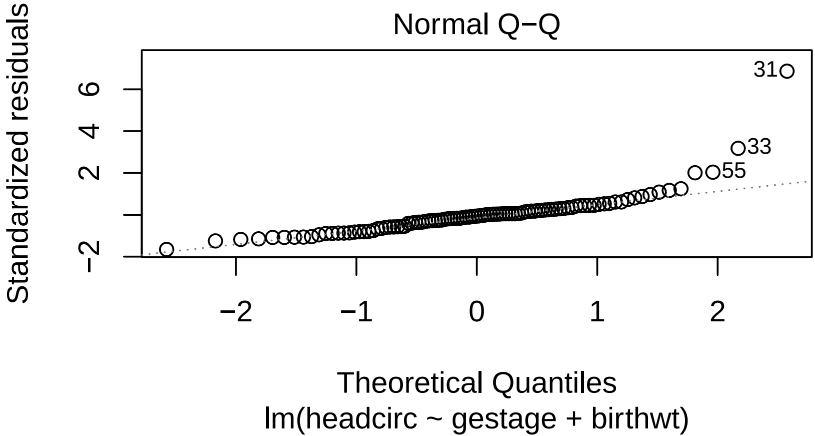
Q-Q Plot of Standardized Residuals

- Idea: Plot the quantiles of the empirical distribution of the standardized residuals against the quantiles of the t distribution with $n - q - 1$ df.
  - Should get a straight line of slope 1 that cuts through the origin.
  - If not, this suggests a violation of the assumption that the error terms are normally distributed with mean 0 and constant SD.
- When $n - q - 1$ is of size 20 or more, the t distribution is similar to the standard normal distribution in shape.
  - Therefore, most software, such as R's `plot()` function, plots the quantiles of the standardized residuals against the quantiles of the standard normal distribution:

Example Q-Q Plot

- For the low-birthweight babies, $n = 100$ babies and we fit $q = 2$ explanatory variables, `gestage` and `birthwt`.
  - So $n - q - 1 = 97$ which is large enough to approximate the t distribution by a standard normal distribution.

```
plot(lfit2,which=2)
```



Comments

- Mostly, the points on the Q-Q plot fall along the straight line that cuts through the origin with slope 1.
- The exceptions are in the upper tail of the distribution of residuals, and labelled as cases 31, 33 and 55.
    - More on outliers next.

Identifying Outliers

- Standardized residuals less than -3 or greater than 3 are considered to be obvious outliers.
- Extract the values of the standardized residuals with the `rstandard()` function;
- E.G., `rstandard(lfit2)` gives the standardized residuals from the `lm()` object `lfit2` that fits the MLR model of `headcirc` as the response variable and `gestage` and `birthwt` as explanatory variables.

- From the resulting output, we see that cases 31 and 33 are outliers. Their standardized residuals r_{31} and r_{33} are greater than 3. As all other r_i's have $|r_i| < 3$, there are no other obvious outliers.

Summary

- We've covered residual diagnostics including:
  1. A plot of residuals *vs.* fitted values to check the assumptions that
     - the linear predictor is correctly specified and
     - the error SD is constant
  2. A Q-Q plot of the standardized residuals *vs.* the quantiles of the standard normal to check the assumption of normal errors
  3. A printout of the sorted list of standardized residuals (the head and tail ends are usually enough) to identify obvious outliers with extreme standardized residuals such that $r_i < -3$ or $r_i > 3$.