STAT6038 week 5 lecture 14

Rui Qiu

2017-03-23

Assessing the underlying (model-specific) assumptions.

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

1. iid = independent and identically distributed
2. N = normally distributed errors
3. mean of distribution is 0 (guaranteed by the least squares estimation -¿ not really an assumption)
4. constant variance σ^2 (homoscedasticity or homoskedasticity)

We assess these assumptions using the residuals (observed errors)

$$e_i = Y_i - \hat{Y}_i, i = 1, 2, \ldots, n$$

and we do this assessment using residual plots.

Key assumptions (in order of importance)

1. errors are independent (no obvious problem)
2. errors are identically distributed with constant variance σ^2 (homoscedastic errors)
3. errors are normally distributed

Use resident plots:

1 and 2 are best assessed using a **plot** of the (standardized) residuals vs. fitted values aka residual plot.

3 is test assessed using a normal quantile plot (qq plot)

Other plots may be useful in diagnosing (getting more details on) problems observed in the main residual plot (and occasionally in normal qq plot).

If residual plot has a "curvature" – a definite pattern $\Longrightarrow$ indicating dependence in the errors $\Longrightarrow$ errors are not independent $\Longrightarrow$ model is probably not appropriate.

If residual plot shows a "heteroscedasticity" $\Longrightarrow$ non-constant variance.

If outliers... outliers...

- lack of independence
- nor constant variance
- potential outlier