Regression Diagnostics

STAT 401A - Statistical Methods for Research Workers

Jarad Niemi

Iowa State University

October 6, 2013

Regression

The simpler linear regression model is

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

this can be rewritten as

$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad e_i \overset{ind}{\sim} N(0, \sigma^2)$$

where we estimate the errors via the residuals

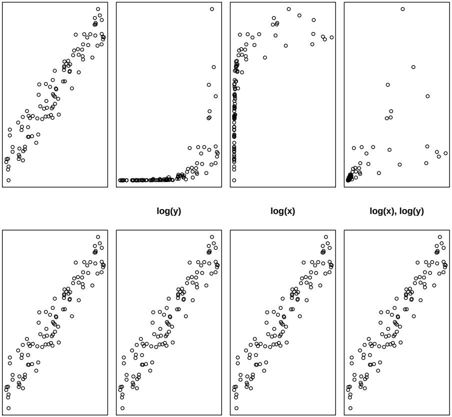
$$r_i = \hat{e}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

Key assumptions are:

- Linearity between mean response and explanatory variable
- Normality of the errors
- Constant variance of the errors
- Independence between observations

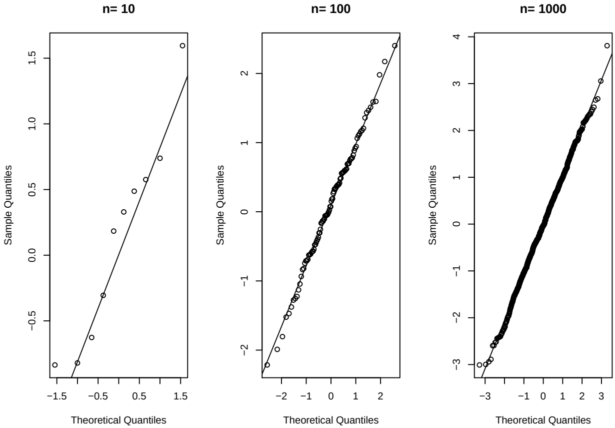
Linearity

Assess using scatterplots of transformed response vs transformed explanatory variable:



Normality

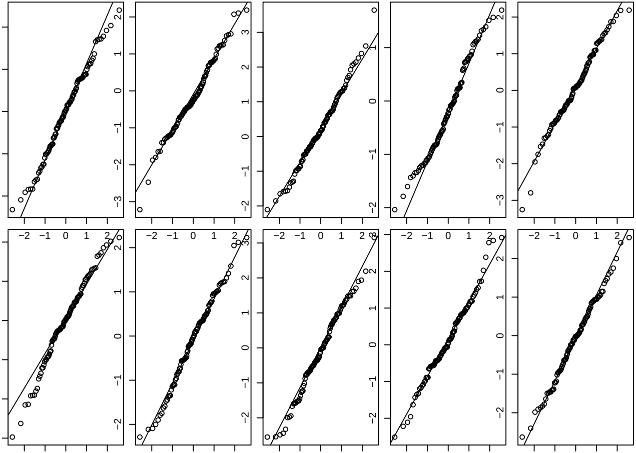
These are normal.



SAS swaps the x and y axes

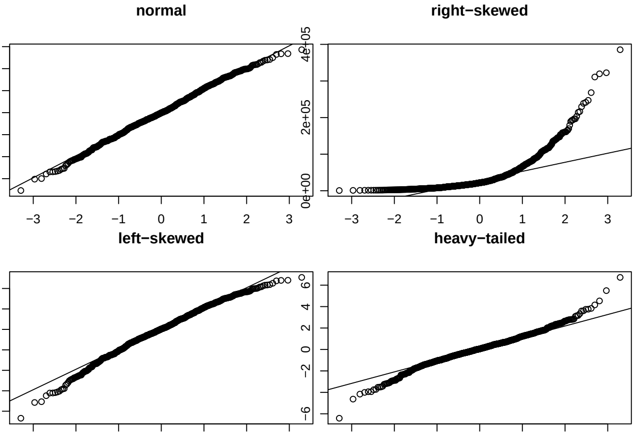
Normality

These are normal.



SAS swaps the x and y axes

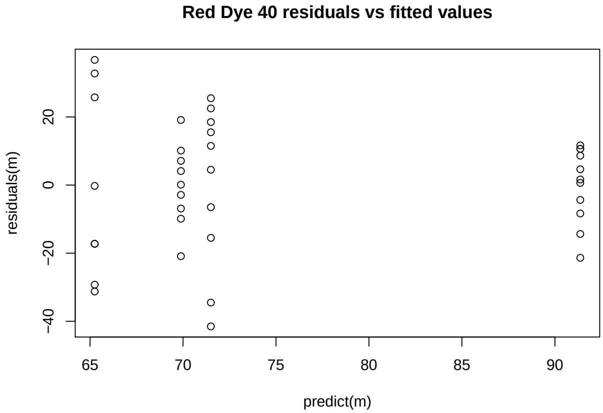
Normality



SAS swaps the x and y axes

Constant variance

Most common non-constant variance is when the variance increases with the mean



Independence

Lack of independence includes

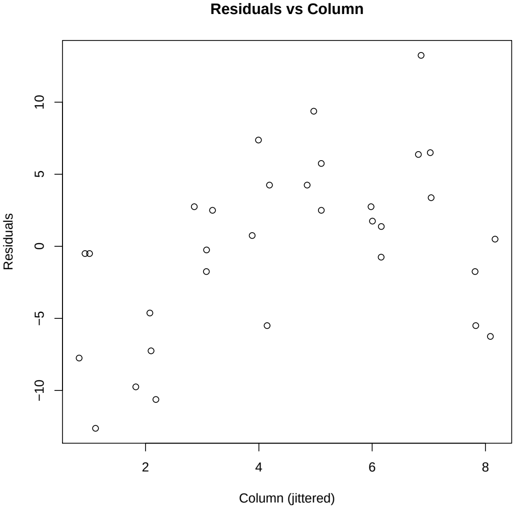
- Cluster effect
- Serial correlation
- Spatial association

Make plots of residuals vs relevant explanatory variables and look for patterns, e.g.

- Residuals vs groups (prefer blocking)
- Residuals vs time (or observation number)
- Residuals vs spatial variable

Spatial association: residuals by spatial coordinates

Potato scab experiment with observations on a 4x8 grid



Summary

Often the best strategy is graphical exploration of the data, here are some relevant graphs:

- transformed response vs transformed explanatory
- transformed response vs transformed explanatory
- qqplot of residuals
- residual vs fitted value
- residual vs explanatory
- residual vs observation number
- residual vs any other variable