ST 517: Data Analytics I

Model assumptions

Outline

Multiple Regression Assumptions

Multiple Regression Assumptions

The assumptions for inference in multiple linear regression are essentially the same as for simple linear regression.

1. The model is specified correctly: the mean of Y actually is a linear function.
2. The observations (Y_i, $\mathbf{X}_i$) are independent of each other.
3. The variance of each Y_i around its mean $\mu(Y_i \,|\mathbf{X}_i)$ is the same value, σ^2.
4. The distribution of Y_i around its mean $\mu(Y_i \,|\mathbf{X}_i)$ is Normal

In simple linear regression X_i was a single value, the value of the explanatory variable for observation i. Here $\mathbf{X}_i$ is many values, the vector of all the explanatory variable values for the ith observation.

Residual Diagnostics

We verify the assumptions are at least reasonable by examining residual plots, just like in simple linear regression.

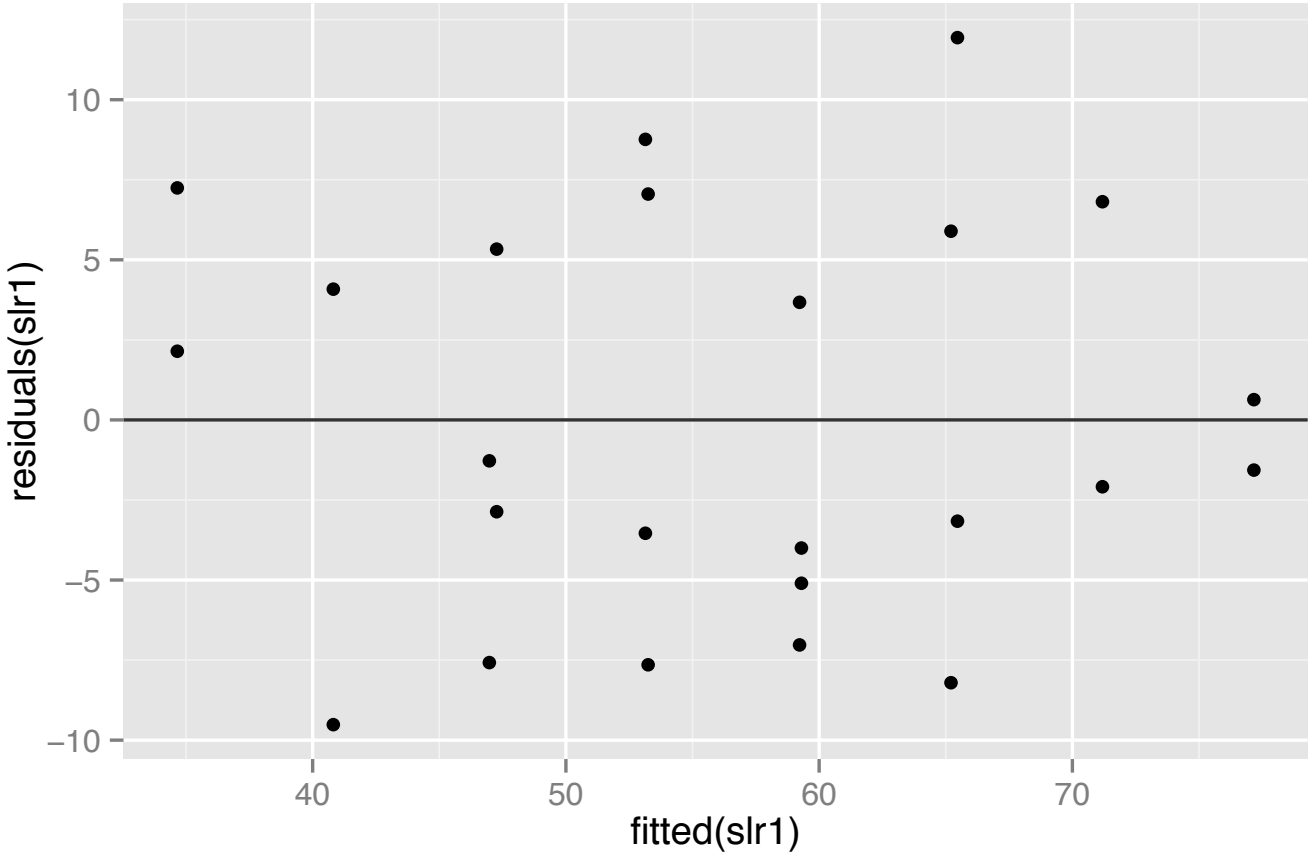
Examine:

- the residuals against the fitted values,
- the residuals against each explanatory variable,
- and particularly if prediction intervals are of interest, a qqplot of the residuals.

In the following examples we'll examine the residuals from the model for the meadowfoam case study:

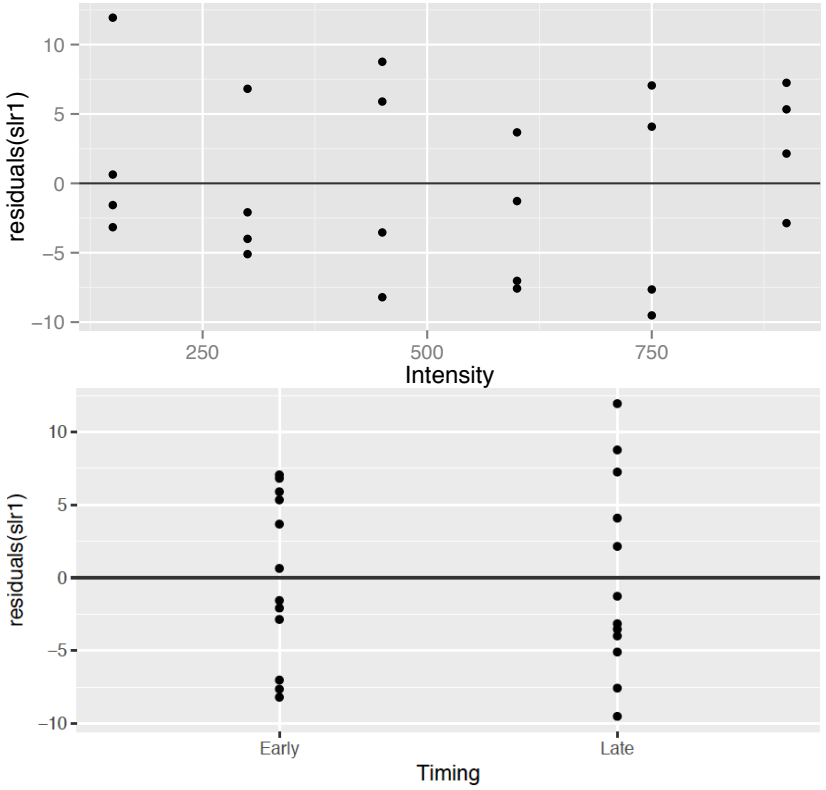
$$\text{Residual}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 Intensity_i + \hat{\beta}_2 late_i + \hat{\beta}_3 (late_i \times Intensity_i))$$

Residuals vs. Fitted: Meadowfoam case study



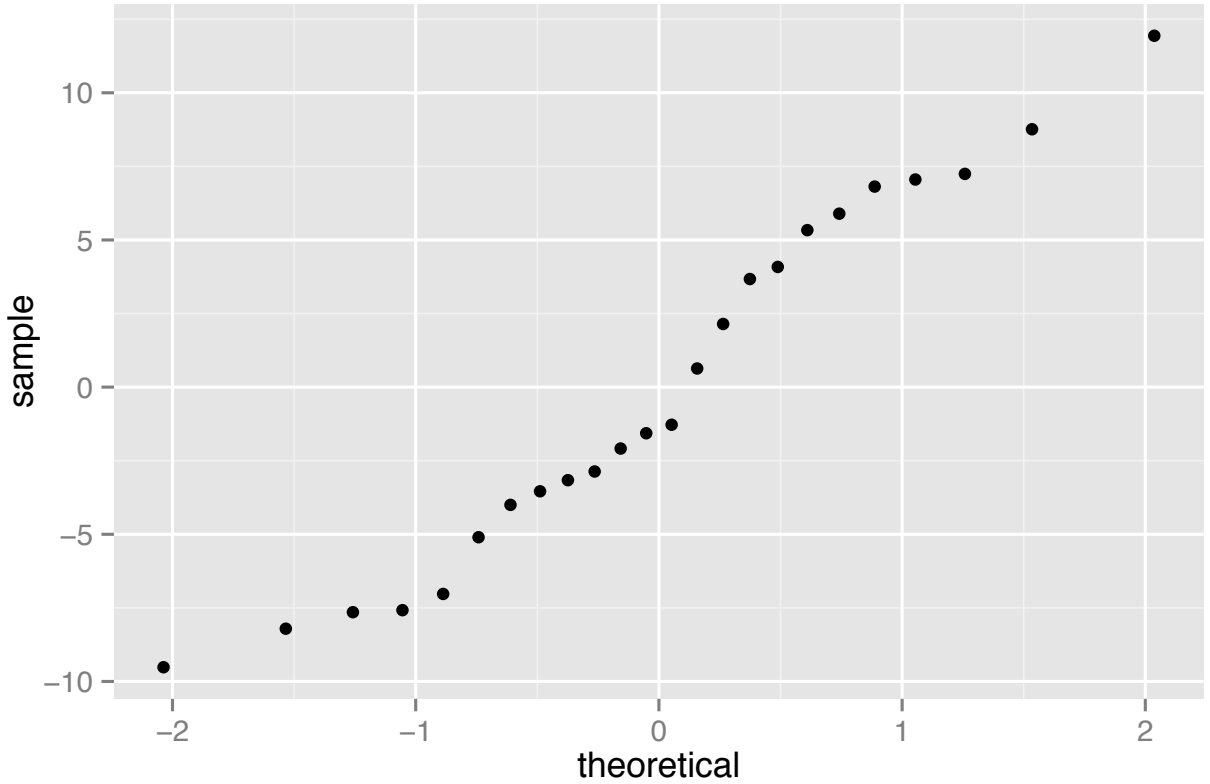
Looks good.

Residuals vs. Explanatory variables: Meadowfoam case study



Very slight curvature in Intensity plot, but don't get too influenced by one or two data points.

Residuals qqplot: Meadowfoam case study



Looks fine, and we are primarily interested in inference on coefficients not prediction.