Andrew Gard - equitable.equations@gmail.com

Multiple linear regression in *R*

Multiple Linear Regression

- A model of the form $y \sim \beta_0 + \sum_{k=1}^{p} \beta_k x_k + \epsilon_i$
- Assumes the response variable y has a linear relationship to the explanatory variables $x_1, x_2, \ldots, x_p$.
- Assumes the irreducible error (encoded by ϵ_i) is normally distributed.
- Assumes *homoscedacisity*, that is, that the spread of the residuals isn't dependent on the values of the explanatory variables.
- In this model, the variables x_i do not interact. The influence of one does not depend on the value of any of the others.
- Always watch out for outliers!

A few important questions

- What if the relationship between y and $x_1, \ldots, x_p$ isn't linear?
- What if some of the explanatory variables are categorical?
- What if the variability of the output depends on one or more of the explanatory variables? That is, what if the data is heteroscedastic?
- What if the explanatory variables are correlated with one another?
- What if the random component doesn't have a normal distribution?