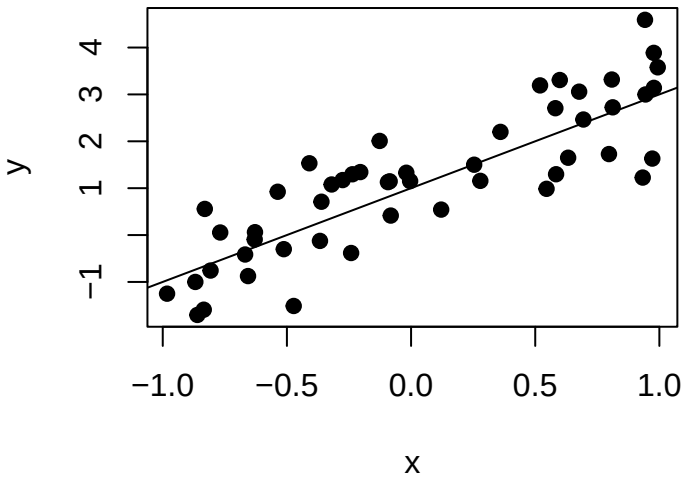
INF511: Modern Regression I

Lecture Materials - Ordinary Least Squares - Part I

Simple linear regression

In the case of simple linear regression, we are seeking to determine if there is a linear association between two variables, an outcome y and a covariate x. If a linear association exists (i.e., "beyond a reasonable doubt"), we seek to quantify that linear relationship and the uncertainty in the parameters that describe that relationship.

Defining the model



Define the parameters:

More on residuals:

Formal assumptions of least squares (and maximum likelihood) regression

By looking at the model definition, we can see some inherent assumptions, but some of the assumptions are more subtle. Violating these assumptions can bias the analysis, meaning that the parameter estimates could be incorrect, or that our interpretation of the analysis (e.g., null hypothesis testing) could be flawed.

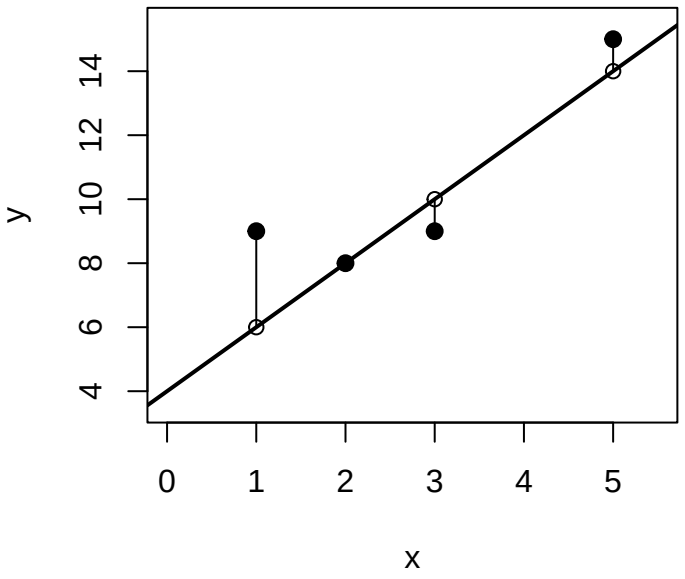
1. *Validity*: Your inputs and outcome variable must be appropriate to your research question. (An issue of research methodology)
2. *Additivity and Linearity*: Your outcome variable must be a *linear* function of the predictors (i.e., no non-linearities). INF512 will deal with non-linear regression analysis.
3. Residuals are I.I.D. (independently and identically distributed). This can be seen in the notation $\epsilon \sim N(0, \sigma^2 I)$, and the online book goes into more detail on this.

4. Equal variances in residuals. Similar to the statement above, the OLS linear regression analysis assumes the same σ^2 across all values of the covariate. This is referred to as homoscedasticity. Heteroscedasticity, which is the violation of this assumption, is one of the more serious issues that can bias linear regression analysis.

5. Normality of residuals. Again, shown with $\epsilon \sim N(0, \sigma^2 I)$. Importantly, this **does not** mean that your outcome data must be normally distributed or that your covariate data must be normally distributed. Additionally, although this assumption gets a lot of attention, current understanding is that this assumption is the least important, compared to the other points mentioned above.

Matrix notation

Example



Ordinary Least Squares (OLS)

Estimating B as $\hat{B}$

In reality, we don't know the model coefficients, stored in column vector B. We need to produce an estimate for these coefficients, $\hat{B}$, which we derive from the data. $\hat{B}$ are therefore known as the regression coefficients or the regression-estimated values of the model coefficients.

There are various methods to estimate the parameters of the linear model, including ordinary least squares (OLS), maximum likelihood, and Bayesian inference. We will start with OLS.

In OLS regression, we **minimize** the "random" component of the model in order to maximize how much variation in the outcome data is explained by the systematic component of the model, which is in this case is the linear association.

The "random" component of the model is best quantified as the **sum of squared errors**.

How do we find a minimum using calculus?

Use the chain rule to minimize $||\epsilon||^2$