Class Notes

Statistical Computing & Machine Learning

Class 6

Review

The geometry of fitting

- Data tables: cases and variables.
- A quantitative variable is a vector.
- A categorical variable can be encoded as a set of "dummy" vectors.
- Response variable and explanatory variable
- The linear projection problem: find the point spanned by the explanatory variables that's closest to the response. That linear combination is the best-fitting model.
  - One explanatory and the response
  - Two explanatory on board and the response on the board (perfect, but meaningless fit)
  - Two explanatory in three-space and the response (residual likely)

Precision of the coefficients

$$\text{standard error of B coef.} = |\text{residuals}| \frac{1}{|\text{B}|} \; \frac{1}{\sin(\theta)} \; \frac{1}{\sqrt{n}} \; \sqrt{\frac{n}{n-m}}$$

- m — degrees of freedom in model
- θ — angle between this model vector and the space spanned by the others
- B — this model vector
- residuals — the residual vector

In-class programming activity

Make a function for the histogram.