Biostat 561: Final Homework

Amy Willis, Biostatistics, UW

05 June, 2019

Homework due Wednesday 12 June, 5 p.m. *Strictly no extensions.*

Office hours Tuesday 11 June, 12:30-1 p.m.

Link to Final Homework submission: https://classroom.github.com/a/svjCDpWB

Be sure to upload a `.cpp` file, a `.py` file, and a `.R` file along with a `.pdf` file showing screenshots of your output along with your commentary.

Question 1: Building familiarity with Rcpp

Write a C++ function that takes arguments `beta` (a vector of dimension `p=3`) and `n` (a scalar), generates data from the following model

$$\begin{aligned}
X \in R^{n\times p}, y \in R^n&, \epsilon \in R^n \\
y &= X\beta + \epsilon \\
X_{i1} &= 1 \\
X_{i2} &\sim Bernoulli(0.7) \\
X_{i3} &\sim Uniform(-1,1) \\
\epsilon_i &\sim N(0,1),
\end{aligned}$$

writes X and y to output files, and returns the following estimate of `beta`:

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Choose a vector `beta` of dimension 3, and confirm the output of your function using R's native matrix multiplication function.

You should be using Rcpp to interface the script with R.

Question 2: Building familiarity with Python

Repeat Question 1, this time writing a Python script to perform the same task.

Question 3:

`reports.csv` is available via Canvas. Use them to answer the following questions:

- Which districts were sampled? How many times was each district sampled?
- Which district and month/year had the most adult dealths?
- Which district had the most client visits for TB?