---
title: "Biostat 561: Homework 9"
author: "Instructor: Amy Willis, Biostatistics, UW"
date: "31 May, 2019"
output:
pdf_document:
extra_dependencies: ["bbm", "hyperref"]
linkcolor: 'red'

Homework 9 due June 5, 3:15pm

No office hours Mon 3 June due to travel; special office hours Tuesday 4 June, 2:30pm

Link to Homework 9 submission: https://classroom.github.com/a/iAetmO52

In this homework, we will continue to explore the coverage of confidence intervals for β_1 under the model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, contrasting confidence intervals created using model-based standard errors ($var(\hat{\beta}) = \hat{\sigma}^2(X^TX)^{-1}$) and confidence intervals created using White's heteroskedasticity-robust "sandwich" standard errors ($var(\hat{\beta}) = (X^TX)^{-1}X^T diag(e_i^2)X(X^TX)^{-1}$.)

We will simulate from the model

$$\begin{aligned} Y_i \mid X_i, u_i &= \beta_1 X_i + \epsilon_i, \text{ where} \\ X_i &\overset{i.i.d.}{\sim} N(0,1), \\ u_i &\overset{i.i.d.}{\sim} N(0,1), \text{ independent of the } X_i, \\ \epsilon_i &= f(X_i) \times u_i. \end{aligned}$$

Question 1: R scripts to assess coverage from the command line

Adapt the R scripts from class to compare the coverage of 95% confidence intervals for β using both model-based and heteroskedasticity-consistent standard errors. Your script should take arguments n (the sample size), β (the true value of β to simulate under), and `option` (which has value either `1` or `2` for the form of $f(x)$, where $f_1(x) = 1 + x^2$ and $f_2(x) = 2 - x^2$), `seed` (the starting seed) and `reps` (the number of replicates to use). Use the package `sandwich` to compute the heteroskedasticity-consistent standard errors.

Show the output of

```
Rscript qtn1_response.R n=100 beta=1 option=1 seed=123 reps=1000
```

and

```
Rscript qtn1_response.R n=100 beta=1 option=2 seed=123 reps=1000
```

What coverage do you find based on each method?

How long did your scripts take to run?

Question 2: Running R scripts on the clusters

Time to use the cluster! Use the examples from lecture to guide your script-writing in this question.

We want to investigate the effect of n on the coverage of the two procedures for the two data-generating mechanisms $f_1(x)$ and $f_2(x)$. We want to run 5000 simulations per `n` and per `option`, and look at $n \in \{50, 100, 150, \ldots, 500\}$.

Adapt your response to question 1 to write an R script that runs `reps` replicates for option 1 and 2, and all $n \in \{50, 100, 150, \ldots, 500\}$. Confirm it works for small `reps`. Call this script `qtn2_response.R`.

Write a shell script called `call_qtn2.sh` that calls `qtn2_response.R`. `qtn2.sh` should be similar to `call_sim_robust_se.sh`, seen in class, and have arguments `sim-name`, `nreps-total` and `nreps-per-job`.

Adapt `submit_sim_robust_se.sh` to create `run_qtn2.sh` to run your simulation study on the cluster. Check it works locally for a small simulation.

Connect to `bayes.biostat.washington.edu` using your favourite `ssh` client (see Homework 8). Run `run_qtn2.sh` on the bayes cluster to run 5000 reps per option and n. Split each simulation into either 5 or 10 jobs (i.e., not 1 job and not 5000 jobs). Perform this batch submission either using a loop in your shell script, or by using a job array.

Present your results graphically and/or tabularly in a way that best illustrates your findings. Comment on what you see.

Optional but recommended for 533 students: Explain why you get the results you see.

Upload your script(s) to your github repository along with a pdf containing a write up of your results.