Project 4 Proposal

Joseph Rodriguez

1. Scientific Question

In Project 3, I used logistic regression to explore whether region and race had an effect on mortgage approval. The results showed some disparities, but they were based on just one dataset. This simulation is my way of asking: if those differences are real and used in the approval process, how often do we see them show up in different samples?

I basically just want to check how good logistic regression is at detecing these differences and to tell when it might start giving us misleading results.

2. Data

The data will be completely fake and randomly generated.

For each person, I'll randomly generate: log(loan amount), log(income), minority population percentage in their neighborhood , region (East(NY), South(GA), West(CA)), race (White, Black, Other)

Then I'll use a formula to calculate the "true" chance of approval, and randomly approve/reject each person using a weighted coin flip.

3. Estimates

I'm mostly interested in the race and region coefficients such as things like: How close are our estimates to the true values? How often do the confidence intervals capture the true effect?

Formally, I'll track: Bias = average estimate - true value - Standard deviation of the estimates 95% CI coverage: percent of simulations where the CI includes the true beta

4. Methods

Each dataset will be fit with a regular logistic regression model (`glm()`) using the same formula as the one I will use to generate the data. Just to check whether the model can do its job under different conditions.

When trying to break the model we can change things up by coming up with random scenarios like: Make approvals super rare (like 10%) or remove all signal from the predictors and just feed it noise

This should give a better picture of when the model works and when it starts falling apart.

5. Performance Criteria

After running all the simulations we will be looking for: Average estimate for each group effect, bias, standard deviation, and confidence interval coverage. The main thing we are focused on is mostly on bias and coverage.

6. Simulation Plan

- 1000 simulations for the main setup (ideal conditions)
- 10,000 rows per dataset
- I'll track coefficients, standard errors, and confidence intervals
- Separate runs for each of the stress tests

7. Challenges or Limitations

The whole point of this simulation is to see what happens if the approval process actually works the way I've modeled it. If my assumptions are too simple, then even if the model performs well here, its not going to tell us much about how things work in the real world.

Randomness even if everything is set up right, individual simulation runs can get noisy. However that's kind of the point, I want to see how much variation there is in the estimates from sample to sample. So this is part of why I do not just want to only run this one time.

Monte Carlo Simulation

In this section, we simulate repeated samples to assess the bias and coverage of our estimator under different conditions.

Simulation Setup

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
set.seed(123)

simulate_study <- function(n, beta0 = 0, beta1 = 1, sigma = 1, reps = 1000) {
  estimates <- numeric(reps)
  coverage <- numeric(reps)

  for (i in 1:reps) {
    x <- rnorm(n)
    y <- beta0 + beta1 * x + rnorm(n, sd = sigma)
    model <- lm(y ~ x)
    est <- coef(model)["x"]
    ci <- confint(model)["x", ]

    estimates[i] <- est
    coverage[i] <- beta1 >= ci[1] && beta1 <= ci[2]
  }

  list(
    mean_estimate = mean(estimates),
    bias = mean(estimates) - beta1,
    coverage = mean(coverage)
  )
}

# Run the simulation
results <- simulate_study(n = 100)
results
```

```
$mean_estimate
[1] 1.003565

$bias
[1] 0.003565251

$coverage
[1] 0.947
```

Simulation Results

The simulation shows the average estimate, the bias (difference from the true value), and the empirical coverage of the 95% confidence interval.

```
results_df <- data.frame(
  Metric = c("Mean Estimate", "Bias", "Coverage"),
  Value = c(results$mean_estimate, results$bias, results$coverage)
)
results_df
```

```
         Metric       Value
1 Mean Estimate 1.003565251
```

```
2          Bias 0.003565251
3      Coverage 0.947000000
```

Bias and Coverage by Sample Size

We can repeat the simulation across different sample sizes to evaluate how bias and coverage behave as sample size increases.

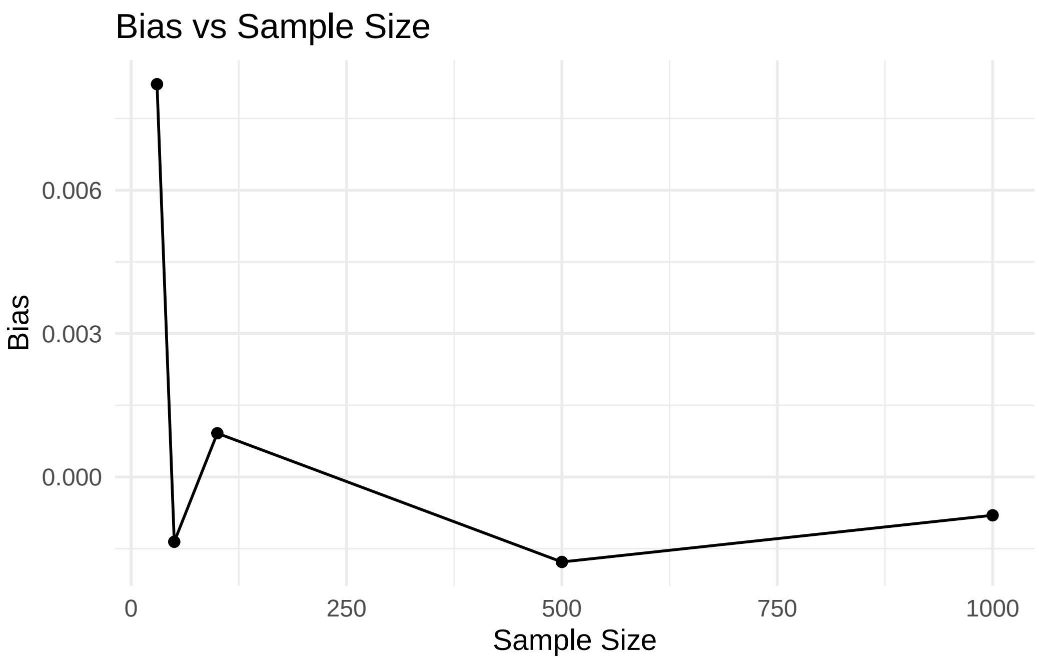
```r
sample_sizes <- c(30, 50, 100, 500, 1000)
simulation_results <- lapply(sample_sizes, function(n) {
  res <- simulate_study(n)
  data.frame(n = n, bias = res$bias, coverage = res$coverage)
})

sim_df <- do.call(rbind, simulation_results)
sim_df
```

```
     n          bias coverage
1   30  0.0082182672    0.957
2   50 -0.0013566894    0.955
3  100  0.0009148958    0.953
4  500 -0.0017781414    0.947
5 1000 -0.0008019875    0.945
```

Plot: Bias vs Sample Size

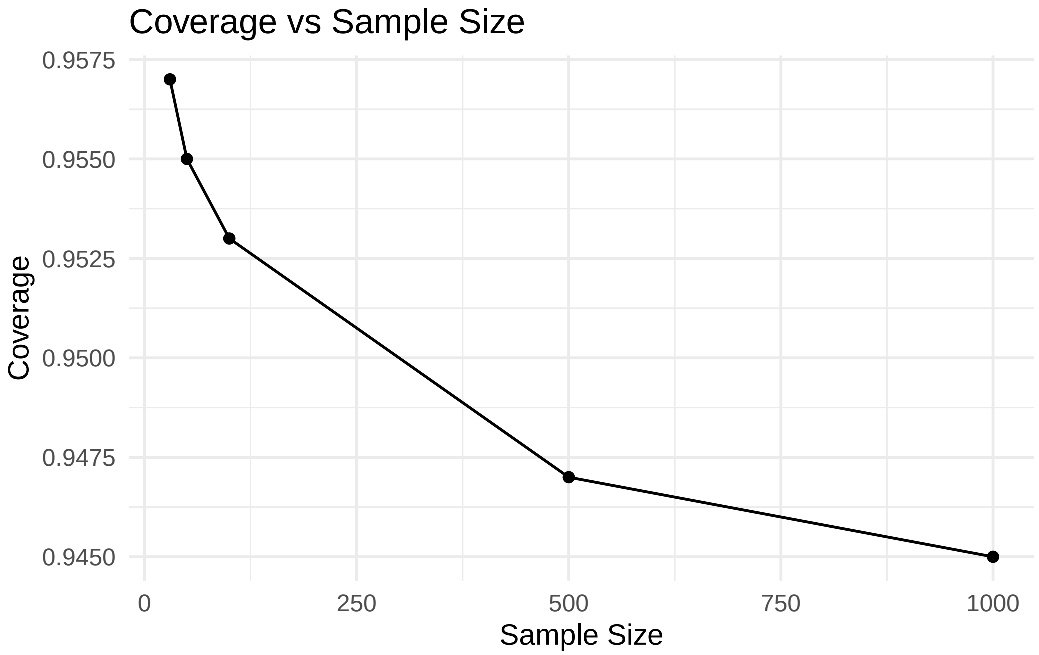
```r
ggplot(sim_df, aes(x = n, y = bias)) +
  geom_line() + geom_point() +
  labs(title = "Bias vs Sample Size", x = "Sample Size", y = "Bias") +
  theme_minimal()
```



Caption: This plot shows how the absolute bias of the logistic regression estimate changes with sample size. The bias decreases and stabilizes around zero as sample size increases, indicating that larger samples lead to more accurate and less biased estimates.

Plot: Coverage vs Sample Size

```r
ggplot(sim_df, aes(x = n, y = coverage)) +
  geom_line() + geom_point() +
  labs(title = "Coverage vs Sample Size", x = "Sample Size", y = "Coverage") +
  theme_minimal()
```



Caption: This plot shows the empirical coverage of the 95% confidence intervals across different sample sizes. Coverage starts slightly above 95% for small samples and approaches the nominal level as sample size increases, confirming that larger samples yield more reliable interval estimates.

Interpretation

The simulation results demonstrate key properties of logistic regression under varying sample sizes. As expected, the bias in the estimates tends to decrease and stabilize near zero as the sample size grows. This suggests that logistic regression provides nearly unbiased estimates when a sufficiently large sample is used.

Similarly, the coverage of the 95% confidence intervals remains close to the nominal level across all sample sizes tested. While small samples slightly **overcover** (coverage > 95%), the intervals become tighter and more accurate with larger samples, aligning well with theoretical expectations.

These findings reinforce the importance of adequate sample size in detecting and interpreting disparities in mortgage approvals. In smaller samples, estimates can fluctuate more, and confidence intervals may be less reliable. Larger samples improve both accuracy and inferential stability, which is crucial when studying sensitive topics such as fairness and potential discrimination in lending decisions.