project4_simulation_Lucas_Neeka

Proposal

---

1. Scientific or Statistical Question

- How can different types of statistical distributions, used to produce simulated data, affect the results of linear and logistic regression models for predicting union membership?

---

2. Data

- The data will be generated through a simulation of the following distributions: gamma, beta, normal, binomial, poisson, exponential, chi-squared, t, F, and uniform.
- Mathematical Notations:
  - linear regression: y = B_0 + B_1*x_1 + B_2*x_2 + … + e
  - logistic regression: ln (p/1-p) = B_0 + B_1*x_1 + B_2*x_2 + … + B_K*x_K
- The assumptions are that whichever statistical distribution simulates data that is closest to the actual data (this is measured by bias of B_1 and B_2, and MSE) has the best affect on results of linear and logistic regression models for predicting union membership. We can assume that all values are scaled and numeric, and that the sample size remains consistent throughout all simulations.

---

3. Estimates

- For each model, the bias of coefficients B__1 and B__2, and MSE will be estimated.

---

4. Methods

- For each distribution, there will be a simulated sample size of n = 100. I will be using linear and logistic regression models, and be evaluating estimates of B__1, B__2, and MSE.
- I hypothesize that distributions with outliers and skewness will produce less accurate estimates for bias of B__1, B__2, and MSE, and produce low performing linear and logistic models.

---

5. Performance Criteria

- Bias of B__1 and B__2: used to evaluate distance between the true values and the estimated coefficients.
- MSE of predicted union membership percentages for linear regression and predicted probabilities (0-1) for logistic regression.
- mean, standard deviation, Type 1 error rate

---

6. Simulation Plan

- Detail how the experiment will be carried out:
  - Number of simulations per distribution: 100
  - Parameter settings:
    - union percent membership (%) = B__0 + 5 categories + e.
      - this formula is derived from the linear regression formula: y = B__0 + B__1*x__1 + B__2*x__2 + ... + e

        * 5 categories = industry, education, sex, sector, year
    - What will be recorded: MSE, bias of B_1 and B_2
    - Any changes from Project III's code: I will add a code to generate simulated data through 10 different distributions, create summary tables of findings from each distribution.

---

7. Anticipated Challenges or Limitations

- Some anticipated challenges or limitations may include: the actual sample of data being relatively small, and since there will be a total of 1,000 simulations (100 simulations per 10 distributions) the run time may be very long.

```r
# Load required libraries
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   3.5.2     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 becon
```

```r
library(broom)

# Setup
n <- 100
num_sim <- 100
```

```r
true_b1 <- 1.5
true_b2 <- -2.0

# Define distributions
distributions <- list(
  gamma = function(n) rgamma(n, 2, 2),
  beta = function(n) rbeta(n, 2, 5),
  normal = function(n) rnorm(n),
  binomial = function(n) rbinom(n, 1, 0.5),
  poisson = function(n) rpois(n, 3),
  exponential = function(n) rexp(n),
  chi_squared = function(n) rchisq(n, 3),
  t = function(n) rt(n, 5),
  F = function(n) rf(n, 5, 2),
  uniform = function(n) runif(n)
)

# Results storage
results <- list()

for (dist_name in names(distributions)) {
  lin_b1 <- c(); lin_b2 <- c(); lin_mse <- c(); lin_type1 <- c()
  log_b1 <- c(); log_b2 <- c(); log_mse <- c(); log_type1 <- c()

  for (i in 1:num_sim) {
    dist_func <- distributions[[dist_name]]

    # Simulate predictors
    x1 <- dist_func(n)
    x2 <- dist_func(n)
    x3 <- dist_func(n)  # noise
    x4 <- dist_func(n)  # noise
    x5 <- dist_func(n)  # noise

    X <- data.frame(x1, x2, x3, x4, x5)

    # Linear outcome
    y_lin <- true_b1 * x1 + true_b2 * x2 + rnorm(n)
    lin_mod <- lm(y_lin ~ ., data = X)
    preds_lin <- predict(lin_mod)

    coefs_lin <- coef(lin_mod)
    lin_b1 <- c(lin_b1, coefs_lin["x1"] - true_b1)
    lin_b2 <- c(lin_b2, coefs_lin["x2"] - true_b2)
    lin_mse <- c(lin_mse, mean((y_lin - preds_lin)^2))

    pvals_lin <- summary(lin_mod)$coefficients[-1, 4]
    lin_type1 <- c(lin_type1, mean(pvals_lin[3:5] < 0.05))

    # Logistic outcome
    logit <- true_b1 * x1 + true_b2 * x2
    p <- plogis(logit)
    y_log <- rbinom(n, 1, p)

    log_mod <- tryCatch(glm(y_log ~ ., data = X, family = binomial), error = function(e) NUL

    if (!is.null(log_mod)) {
      pred_prob <- predict(log_mod, type = "response")
      coefs_log <- coef(log_mod)

      log_b1 <- c(log_b1, coefs_log["x1"] - true_b1)
      log_b2 <- c(log_b2, coefs_log["x2"] - true_b2)
      log_mse <- c(log_mse, mean((p - pred_prob)^2))

      pvals_log <- summary(log_mod)$coefficients[-1, 4]
      log_type1 <- c(log_type1, mean(pvals_log[3:5] < 0.05))
    }
  }

  results[[dist_name]] <- tibble(
    distribution = dist_name,
    lin_bias_b1 = mean(lin_b1),
    lin_bias_b2 = mean(lin_b2),
    lin_mse = mean(lin_mse),
    lin_type1_error = mean(lin_type1),
    log_bias_b1 = mean(log_b1),
    log_bias_b2 = mean(log_b2),
    log_mse = mean(log_mse),
    log_type1_error = mean(log_type1)
  )
}
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: algorithm did not converge

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Warning: glm.fit: algorithm did not converge

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
# Combine and display results
summary_results <- bind_rows(results)
summary_results
```

```
# A tibble: 10 x 9
   distribution lin_bias_b1 lin_bias_b2 lin_mse lin_type1_error log_bias_b1
   <chr>              <dbl>       <dbl>   <dbl>           <dbl>       <dbl>
 1 gamma          -0.0107    -0.00300     0.960          0.0367    1.63e- 1
 2 beta            0.0291    -0.0262      0.953          0.0267    1.28e- 1
 3 normal          0.0120    -0.00600     0.946          0.0733    2.79e- 1
 4 binomial       -0.00414    0.0392      0.957          0.0533    2.48e- 1
 5 poisson         0.00607   -0.00620     0.946          0.0567    3.51e- 1
 6 exponential    -0.0147    -0.0114      0.936          0.0433    2.91e- 1
 7 chi_squared    -0.00320   -0.00274     0.916          0.0433    5.07e- 1
 8 t               0.00804   -0.00748     0.928          0.07      2.09e- 1
 9 F              -0.000371   0.0000530   0.954          0.06      2.15e+12
10 uniform         0.0493    -0.00362     0.927          0.0467    2.66e- 1
# i 3 more variables: log_bias_b2 <dbl>, log_mse <dbl>, log_type1_error <dbl>
```