lab 16

2025-06-25

```r
set.seed(42)

# Number of parks, teams
n_parks <- 30
n_teams <- 90

# True coefficients
beta_0 <- 0.4
beta_park <- rnorm(n_parks, mean = 0.04, sd = sqrt(0.065))
beta_off <- rnorm(n_teams, mean = 0.02, sd = sqrt(0.045))
beta_def <- rnorm(n_teams, mean = 0.03, sd = sqrt(0.07))
```

```r
n_innings <- 3000

park_ids <- sample(1:n_parks, n_innings, replace = TRUE)
off_ids  <- sample(1:n_teams, n_innings, replace = TRUE)
def_ids  <- sample(1:n_teams, n_innings, replace = TRUE)

# Construct the design matrix
X <- model.matrix(~ factor(park_ids) + factor(off_ids) + factor(def_ids))

library(truncnorm)
```

```
## Warning: package 'truncnorm' was built under R version 4.3.3
```

```r
# Store y for all 100 datasets
y_list <- vector("list", 100)

for (m in 1:100) {
  mu <- beta_0 + beta_park[park_ids] + beta_off[off_ids] + beta_def[def_ids]
  y <- round(rtruncnorm(n_innings, a = 0, mean = mu, sd = 1))
  y_list[[m]] <- y
}
```

```r
park_effects_ols <- matrix(NA, nrow = 100, ncol = n_parks - 1)
park_effects_ridge <- matrix(NA, nrow = 100, ncol = n_parks - 1)

lambdas = 10^seq(-3, 3, by = 0.2)

for (m in 1:100) {
  y <- y_list[[m]]

  # OLS
  ols_fit <- lm(y ~ factor(park_ids) + factor(off_ids) + factor(def_ids))
  park_effects_ols[m, ] <- coef(ols_fit)[2:n_parks]
```

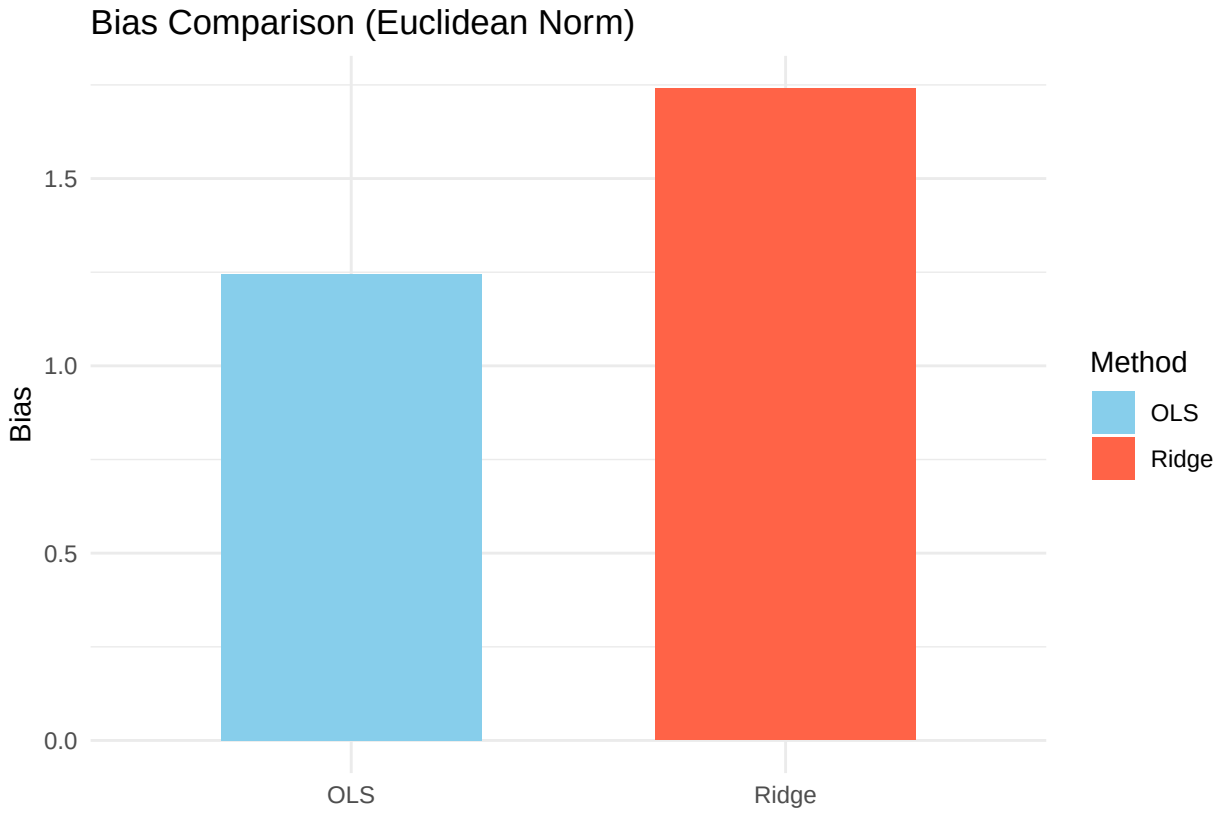
```r
  # Ridge
  ridge_fit <- cv.glmnet(X, y, alpha = 0, lambda = lambdas,
                         nfolds = 5, standardize = FALSE, family = "gaussian")
  beta_hat <- coef(ridge_fit, s = "lambda.min")
  park_effects_ridge[m, ] <- as.numeric(beta_hat[2:n_parks])
}
```

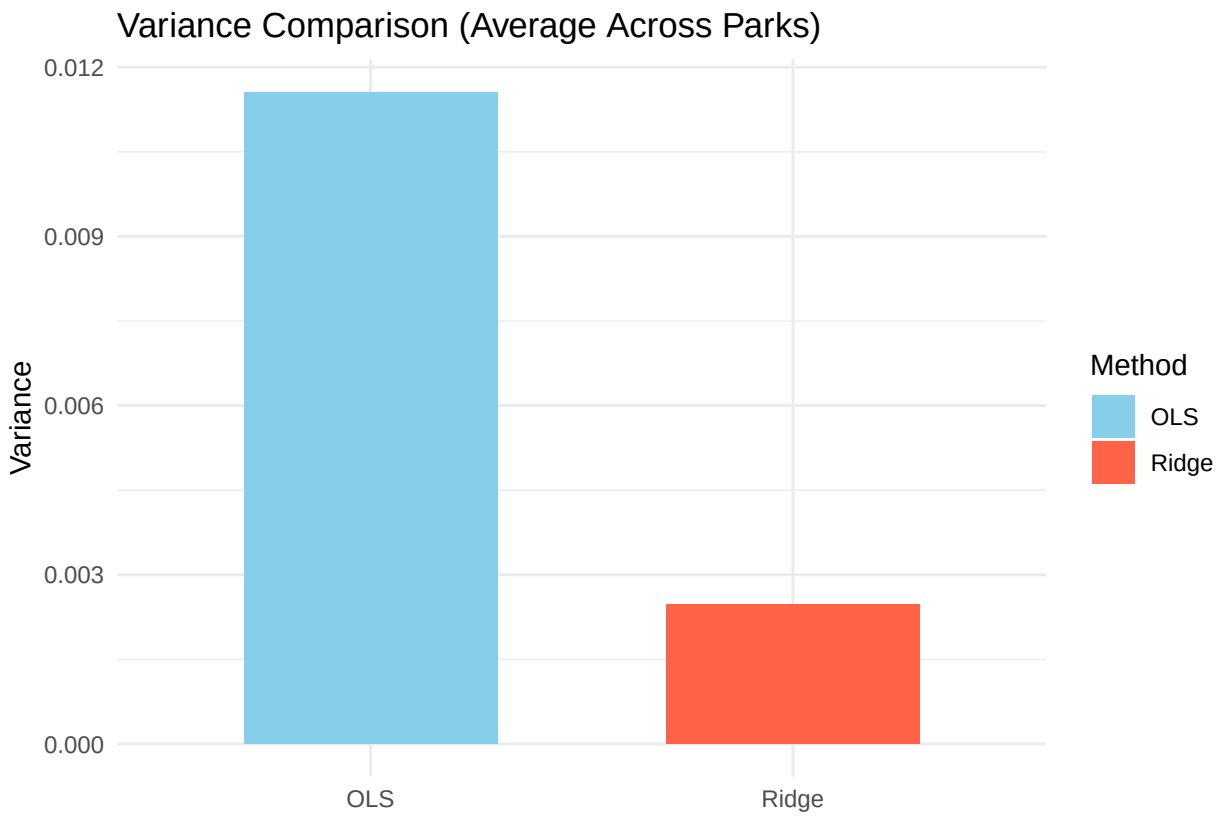
```r
true_park <- beta_park - mean(beta_park)  # since model.matrix drops reference level

# Mean estimates
ols_mean <- colMeans(park_effects_ols)
ridge_mean <- colMeans(park_effects_ridge)

# Bias
bias_ols <- sqrt(sum((ols_mean - true_park[-1])^2))
bias_ridge <- sqrt(sum((ridge_mean - true_park[-1])^2))

# Variance
var_ols <- mean(apply(park_effects_ols, 2, var))
var_ridge <- mean(apply(park_effects_ridge, 2, var))
```




Variance Comparison (Average Across Parks)
Variance
0.012
0.009
0.006
0.003
0.000
OLS
Ridge
Method
OLS
Ridge