Do Submariners Have a Higher Probability of Fathering Girls?

Nathan Honecker

April 24, 2024

Introduction

The belief that submariners father more girls is prolific among U.S. Navy sailors. This analysis aims to determine if there is evidence to support this belief. We used data collected by survey from 1,000 U.S. sailors. They self-reported if they had a child in the last year and the sex of that child, as well as information pertaining to their duties on the job, the number of years they've been in the service, and if they are currently assigned to a submarine.

There is reasonable concern for reporting bias. Since the data is self-reported by the subjects, there is a possibility of fabricating answers in order to support the myth we are studying. Limitations include our sample size (n=1000), which may appear adequate, but could always stand to be larger. Especially looking at time spent in the service, we received inadequate samples for several individual year values. We also are only looking at active servicemen. It is entirely possible that this myth was true in the past but is no longer, given submarines are much more advanced than they once were and sailors are more protected from dangerous conditions and contaminants.

Exploratory Data Analysis

We are working with five covariates, four are factor variables and one is continuous. Our factor variables include `Sea`, which indicates whether the sailor is currently assigned to a submarine or not, `BM`, which indicates whether they are assigned to a ballistic missile submarine or not, `Engineer`, which indicates whether they are working directly with a nuclear reactor or not, and `Weapons`, which indicates whether they are working directly with nuclear weapons or not. Our sole continuous variables is `Years`, which indicates how many years it has been since they began submarine service.

Proportions of Female Children:

On Shore	At Sea
0.480	0.555

No Ballistic Missile	Yes Ballistic Missile
0.529	0.536

No Nuclear Reactor	Yes Nuclear Reactor
0.535	0.528

No Nuclear Weapons	Yes Nuclear Weapons
0.532	0.524

As we can see from the tables above, whether a sailor is currently assigned to a submarine at sea or not appears to be the most explanatory variable, as the proportion of female children goes from 0.48 for sailors on shore to 0.555 for sailors at sea. The other three factor variables do not have a noticeable correlation with our response variable based on the above metric.

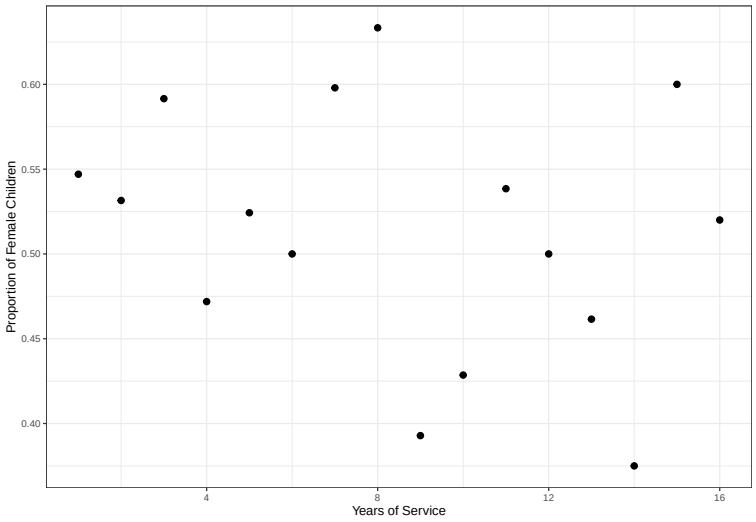

Figure 1: The proportion of female children across years of service

Based on `Figure 1`, the predictor variable `Years` does not appear to have any apparent correlation with with our response variable.

Model and Interpretation

Let Y_i be an indicator of whether the ith sailor has had a female child born in the last 12 months, where $i = 1, 2, 3, ..., 1000$.

Let `Sea = 1` if the ith sailor is currently assigned to a submarine. Let `BM = 1` if the ith sailor is assigned to a ballistic missile submarine. Let `Engineer = 1` if the ith sailor works directly with a nuclear reactor. Let `Weapons = 1` if the ith sailor works directly on nuclear weapons. Let `Years` represent the number of years the ith sailor has been in submarine service.

We will assume each y_i is a realization of a Bernoulli random variable. Thus

$$Y_i \mid \alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5 \sim \text{Bernoulli}(\theta_i)$$

and

$$\log(\frac{\theta_i}{1-\theta_i}) = \alpha + \beta_1 * \text{Sea}_i + \beta_2 * \text{BM}_i + \beta_3 * \text{Engineer}_i + \beta_4 * \text{Weapons}_i + \beta_5 * \text{Years}_i$$

Let

$$\alpha \sim \text{N}(-0.06, .05), \beta_k \sim \text{N}(0, 1), k = 1, 2, 3, 4, 5$$

The predictor variables were all standardized, therefore each β_k has a mean of 0 and a variance of 1, as when variables are standardized, they achieve a mean of 0 and unit variance. Therefore, we started all β_k parameters at 0.

Note that α is our reference level, so α represents the probability of a girl given a sailor has 0 years of experience and no interaction with a submarine. We chose the α starting value (-0.06) because logit(0.485) = -0.06, and 0.485 is the proportion of female births in the general human population. We also assigned this value to the mean of the distribution of α with a relatively high precision (20), since we have confidence that a sailor not on a submarine and with 0 years of experience can be reasonably considered to near the general population probability.

This model runs 5,000 total iterations. We allowed 5,000 iterations for burn-in and 5,000 for finding the best proposal distribution, which leaves 5,000 iterations for posterior estimation.

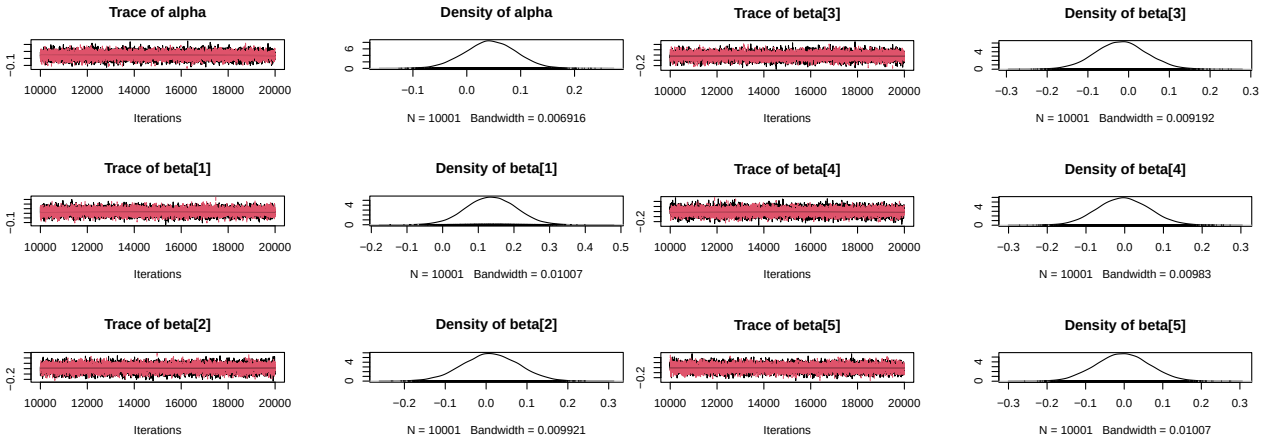

Figure 2: Diagnostic plots

As we can see from `Figure 2`, our density plots appear smooth and our trace plots appear to converge with no burn-in during the final 5,000 iterations for all six parameters. Based on this, we conclude that our algorithm did converge.

We can see, looking at the density plots, that the majority of our predictors have highest density near zero. The exceptions being our reference level, α, and, to a greater extent, β_1, which is the coefficient for `Sea`. It therefor appears that `Sea` has the most explanatory power in regards to our response variable compared to our other predictors.

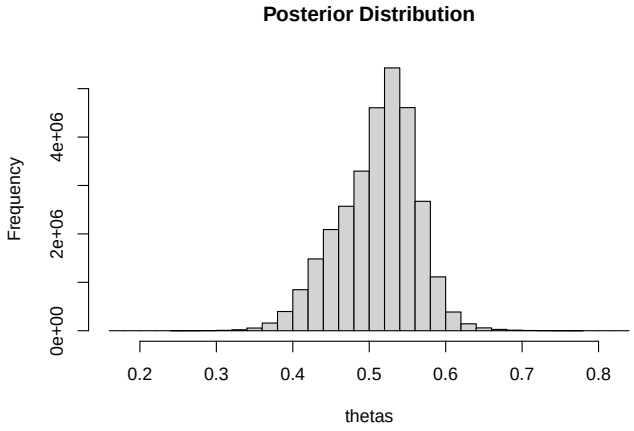

Figure 3: Posterior distribution

Our posterior distribution appears to have a mean significantly above 0.5, and indeed the mean of our posterior samples is 0.511. This places our sample of sailors well above the general population's proportion of having female children.

Model Summary

```
Iterations = 10000:20000
Thinning interval = 1
Number of chains = 2
Sample size per chain = 10001

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

             Mean      SD  Naive SE Time-series SE
alpha    0.043938 0.04756 0.0003363      0.0004372
beta[1]  0.136016 0.06889 0.0004871      0.0006881
beta[2]  0.011042 0.06784 0.0004797      0.0006775
beta[3] -0.018441 0.06337 0.0004481      0.0005748
beta[4] -0.003028 0.06721 0.0004753      0.0006695
beta[5] -0.004056 0.06885 0.0004868      0.0007202

2. Quantiles for each variable:

              2.5%      25%       50%     75%  97.5%
alpha   -0.0502107  0.01246  0.044144 0.07583 0.1366
beta[1]  0.0005741  0.08932  0.135933 0.18309 0.2705
beta[2] -0.1237010 -0.03467  0.010937 0.05738 0.1436
beta[3] -0.1434182 -0.06111 -0.018527 0.02311 0.1066
beta[4] -0.1324115 -0.04867 -0.003553 0.04214 0.1301
beta[5] -0.1379184 -0.05036 -0.003837 0.04244 0.1301
```

In the provided summary, we can interpret our slope coefficients as such: $\hat{\beta}_1 = 0.136$, therefore, assuming other covariates are fixed, the log-odds ratio having a girl increases by 0.136 if a sailor is currently assigned to a submarine.

Given our θ is in the form of log-odds, the following are more intuitive interpretations of our β estimations: $e^{\hat{\beta}_1} = 1.145$, therefore the multiplicative change in odds of having a girl is 1.145 if a sailor is currently assigned to a submarine. The same interpretation can be applied to our other factor variables: $e^{\hat{\beta}_2} = 1.011$, $e^{\hat{\beta}_3} = 0.981$, $e^{\hat{\beta}_4} = 0.997$.

Note that if the multiplicative change in odds is less than 1, then that indicated a decrease in odds.

$e^{\hat{\beta}_5} = 0.996$, meaning that the multiplicative change in odds of having a girl is 0.996 for each 1 year increase in `Years`.

Using our model in full, we can calculate posterior predictive estimations. For example, assume all of our covariates equal 0. That is, given a sailor with 0 years of experience who is not currently assigned to a submarine, and does not work with nuclear reactors, nuclear weapons, or on a sub with ballistic weapons,

the probability that that sailor has a girl is 0.511. Recalling the possible bias during data collection, we believe this value should be closer to 0.485, given this is the reference level with all covariates equal to 0 and the informative nature of our α prior distribution.

Given `Sea = 1` and our other four covariates equal 0, such a sailor's probability of having a girl is 0.545. This is relatively large leap in probability given we only changed the value of one covariate. Looking at $\hat{\beta}_1$, it is by far the largest and it does lead to a significant jump in the probability of a sailor having a girl. The density plot for β_1 stood out as well, showing a distribution departing from 0.

Notice our other four coefficient estimations are much closer to 0, or, when converted to multiplicative change, they are much closer to 1. A coefficient equaling 0 means that our response is completely independent of the corresponding covariate.

Conclusions

Given our summary output, there is not enough evidence to conclude that the probability of a sailor having a girl is dependent in any way on the `BM`, `Weapons`, `Engineer`, or `Years` variables. However, this data and subsequent analysis has provided evidence that the probability of a sailor having a girl is dependent on whether a sailor is currently assigned to a submarine or not, and that a sailor being actively at sea and assigned to a submarine has a positive correlation with the probability of fathering a baby girl. We cannot say if this is a direct causation and there are still questions of bias in data collection method, however given this data there is a positive correlation.

Appendix

```
## setup
set.seed(99)
library(patchwork)
library(jtools)
library(tidyverse)
library(matrixcalc)
library(coda)
library(rjags)
library(readxl)
library(knitr)

## functions
logit <- function(x) log(x/(1 - x))
```

```r
invlogit <- function(x) exp(x)/(1 + exp(x))

## plots bw theme
ggplot2::theme_set(
    ggplot2::theme_bw(base_size = 8)
)

## Importing data + clean-up
subdata <- read_excel("data.xlsx")
colnames(subdata) = c("sea", "bm", "eng", "wpns", "years", "y")

## Standardize
scaledsubdata <- subdata
scaledsubdata[c(1,2,3,4,5)] <-  lapply(subdata[c(1,2,3,4,5)], function(x) c(scale(x)))
scaledsubdata <- cbind(scaledsubdata$sea, scaledsubdata$bm, scaledsubdata$eng, scaledsubdata$wpns,
                       scaledsubdata$years, subdata$y)
scaledsubdata <- as.data.frame(scaledsubdata)
colnames(scaledsubdata) = c("sea", "bm", "eng", "wpns", "years", "y")

## Post-standardize check for N(0,1)
view(subdata)
view(scaledsubdata)
colMeans(scaledsubdata)
apply(scaledsubdata, 2, sd)

## data mean
prop_girls <- mean(subdata$y)

# EDA TABLES
## sea proportion
sea0 <- subdata %>%
  filter(sea == 0)
sea0 <- (sum(sea0$y) / nrow(sea0))

sea1 <- subdata %>%
  filter(sea == 1)
sea1 <- (sum(sea1$y) / nrow(sea1))

seavec <- c(sea0,sea1)

## bm proportion
bm0 <- subdata %>%
  filter(bm == 0)
bm0 <- (sum(bm0$y) / nrow(bm0))

bm1 <- subdata %>%
  filter(bm == 1)
bm1 <- (sum(bm1$y) / nrow(bm1))

bmvec <- c(bm0,bm1)

## eng proportion
eng0 <- subdata %>%
```

```
  filter(eng == 0)
eng0 <- (sum(eng0$y) / nrow(eng0))

eng1 <- subdata %>%
  filter(eng == 1)
eng1 <- (sum(eng1$y) / nrow(eng1))

engvec <- c(eng0,eng1)

## wpns proportion
wpns0 <- subdata %>%
  filter(wpns == 0)
wpns0 <- (sum(wpns0$y) / nrow(wpns0))

wpns1 <- subdata %>%
  filter(wpns == 1)
wpns1 <- (sum(wpns1$y) / nrow(wpns1))

wpnsvec <- c(wpns0,wpns1)

## Scatterplot of Years vs dependent variable in proportion
yearsgroup <- subdata %>%
  group_by(years) %>%
  summarize(n())

years_y <- subdata %>%
  group_by(years) %>%
  summarize(count = sum(y))

years <- merge(yearsgroup, years_y, by="years")

yearsplotdata <- years %>%
  mutate(prop = count /`n()`)

yearsplot <- ggplot(yearsplotdata, aes(years, prop)) + labs(x = "Years of Service", y = "Proportion of
yearsplot

## calculation of alpha starting value
logit(0.485)

## JAGS
mydata <- list(n = nrow(scaledsubdata), sea = scaledsubdata$sea, bm = scaledsubdata$bm, eng =
                 scaledsubdata$eng, wpns = scaledsubdata$wpns, years = scaledsubdata$years, y =
                 scaledsubdata$y)

niter = 15000
nburn = 5000
nadapt = 5000
nchains = 2

myinit=list(alpha = -0.06, beta=rep(0,5))

mymodel = "model{
```

```
#likelihood
for(i in 1:n){
y[i] ~ dbern(theta[i])
logit(theta[i]) <- alpha + beta[1] * sea[i] + beta[2] * bm[i] + beta[3] * eng[i] + beta[4] * wpns[i] + b
}

#prior
for (j in 1:5) {
beta[j] ~ dnorm(0,1)
}
alpha ~ dnorm(-0.06,1/0.005)
}"

## Model Output
fit1 = jags.model(textConnection(mymodel), data=mydata ,inits=myinit, n.chains=nchains, n.adapt=nadapt)
fit.samples = coda.samples(fit1, c("beta", "alpha"), n.iter=niter)
post.sample = coda.samples(fit1, c("theta"), n.iter=niter)
whole_summary <- summary(fit.samples)

## Diagnostic plots
plot(window(fit.samples,start=nburn+nadapt))

## Posterior distribution plot
thetas = as.matrix(post.sample)
hist(thetas, main = "Posterior Distribution")
postmu <- mean(thetas)

## model summary post-burn-in and post-adapt
summary(window(fit.samples,start=nburn+nadapt))

## conversion of beta estimations
## b1
exp(0.135769)
## b2
exp(0.010451)
## b3
exp(-0.019154)
## b4
exp(-0.003360)
## b5
exp(-0.004496)

## Posterior Predictive Estimations
invlogit(0.044302)
invlogit(0.044302 + 0.135769)
```