Homework 4

PubH 7406: Biostatistical Inference II – Jared D. Huling

Due: Tues, April 4, 2023 at the beginning of class

Bootstrap HW - 45 points

Instructions

This assignment has two questions (Exercise A and Exercise B), each with several subparts. The two topics are quite different, but both are investigations into the various utilities of the bootstrap in statistical practice.

Turn in the homework in the form of a PDF. It is fine to use existing functions to answer questions. However, note that simply providing output from statistical software is not sufficient and will not receive full points. Any output/results must be interpreted in the context of the problem and accompanying explanations of the results are necessary. Use clearly-defined and explained statistical notation to accompany results. For example, any statistical tests should be accompanied by a formal statement of the null and alternative hypotheses with additional context explaining what the hypotheses mean in the context of the problem. Please follow the instructions on homeworks in the syllabus in order to receive full credit.

If you have any questions, please ask in the course Q&A on Canvas so that others can see any responses.

The Questions

Exercise A: Thall and Vail dataset

This set of questions will explore Poisson GLMs, one of their common deficiencies in practice, and how the bootstrap can be used as a tool to mitigate that deficiency.

Thall and Vail (1990) presented data from a clinical trial of $n = 59$ epileptic patients who were randomized to take either a new drug $d_i = 1$ or a placebo ($d_i = 0$) in addition to standard chemotherapy. Other baseline data included $a_i = \log(\text{age}_i)$ where age_i is age in years and $b_i = \log(\text{base}_i/4)$, where base_i is number of seizures in preceding 8-week period. The outcome is $Y_i = \sum_{j=1}^{4} Y_{ij}$ the number of seizures within the following 8 weeks.

We can read the data in as follows:

```
library(tidyverse)
fpath <- "https://jaredhuling.org/data/pubh7406/thall_and_vail_1990.dat"
seiz <- read.table(fpath, header=TRUE)

## sum up seizures across visits
seiz_total <- seiz %>%
  group_by(id) %>%
  dplyr::summarize(seiz = sum(seiz),
                   age = age[1],
                   base = base[1],
                   log_base = log(base/4),
                   treat = treat[1])
```

The Poisson assumption and variance estimates

The outcome here is a count, so a first thought would be to fit a regression model via a Poisson GLM. However, one thing to note about Poisson GLMs is that if the Poisson assumption *does not hold* for the outcome, the variance estimates for the coefficients will generally be far under-estimated (it is in fact quite likely for the assumption of a count response being distributed as Poisson *not* to hold, since the Poisson distribution requires the mean and variance of the response to be the same – this almost never happens).

However, contrary to popular belief, this does not mean that Poisson GLMs are not useful. In fact, even if the Poisson assumption does not hold, the *coefficient* estimates can still be good estimates if the model for the *mean* is reasonable (the mean of the response under a Poisson GLM, if you recall, is $\exp\{x_i^T\beta\}$). The issue is often simply that the variance is under-estimated under the Poisson assumption.

Fitting the usual Poisson regression model:

```
summary(f1 <- glm(seiz ~ log(age)+log_base+treat, data = seiz_total,
                  family = poisson()))
```

```
##
## Call:
## glm(formula = seiz ~ log(age) + log_base + treat, family = poisson(),
##     data = seiz_total)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.0834  -2.0602  -0.4096   1.3963   8.1997
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.02151    0.40354  -2.531   0.0114 *
## log(age)     0.58778    0.10992   5.347 8.93e-08 ***
## log_base     1.22522    0.03252  37.672  < 2e-16 ***
## treat       -0.01759    0.04818  -0.365   0.7150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2122.73  on 58  degrees of freedom
## Residual deviance:  556.39  on 55  degrees of freedom
## AIC: 847.66
##
```

```
## Number of Fisher Scoring iterations: 5
```

Quasi-Poisson regression allows for so-called "overdispersion", which accounts for the under-estimation of the variance by applying a correction that makes the variance bigger for each coefficient by a constant (the constant is the "over-dispersion" parameter). The only difference between Poisson and Quasi-Poisson is the variance estimates (the coefficient estimates are the same). A Quasi-Poisson GLM can be fit as:

```
summary(fq1 <- glm(seiz ~ log(age)+log_base+treat, data = seiz_total,
                   family = quasipoisson()))
```

```
##
## Call:
## glm(formula = seiz ~ log(age) + log_base + treat, family = quasipoisson(),
##     data = seiz_total)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.0834  -2.0602  -0.4096   1.3963   8.1997
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.02151    1.34172  -0.761    0.450
## log(age)     0.58778    0.36548   1.608    0.114
## log_base     1.22522    0.10813  11.330 5.27e-16 ***
## treat       -0.01759    0.16019  -0.110    0.913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 11.05488)
##
##     Null deviance: 2122.73  on 58  degrees of freedom
## Residual deviance:  556.39  on 55  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

However, we can also use the bootstrap as another principled means of obtaining uncertainty estimates, which do not require that the correction factor for the variance be the same for every coefficient.

Exercises for Thall and Vail data

1. (2 points) Verify that the variances of the coefficients under the Quasi-Poisson model are inflated from the variances of the standard Poisson regression model by a constant factor.

2. (5 points) Use a non-parametric bootstrap to obtain standard error estimates for all coefficients in the above-fitted Poisson GLM. How do the results compare with the standard Poisson GLM? How do the results compare with the quasi-Poisson model?

3. (5 points) Use the bootstrap to estimate the bias for each coefficient estimate – are the biases small or large?

4. (5 points) Visualize the bootstrap sampling distribution of each of the coefficients – does a normal approximation seem reasonable for all coefficients? Construct bootstrap 95% CI's for each coefficient using the i) normal approximation method and ii) the percentile method and compare the results – if they are substantially different, why do you think that is the case? Hint: there may be more than just 1 reason.

5. (8 points) Repeat questions 2. and 4. above but using the following model that includes all possible interactions of the three predictors: `glm(seiz ~ log(age)*log_base*treat, data = seiz_total, family = poisson())`. Describe any differences in the bootstrap sampling distributions of the coefficients compared with the first model fit.

Exercise B: Fish Dataset

The dataset `fish` contains 40 annual counts of the numbers of spawners `S` and recruits `R` in a salmon population.

```
fish <- read.table('http://jaredhuling.github.io/data/fish.txt',
                   header=TRUE)
```

The units are thousands of fish. Spawners are fish that are laying eggs. Spawners die after laying eggs. Recruits are fish that enter the catchable population.

The Beverton-Holt Model

The classic Beverton-Holt model for the relationship between spawners and recruits is

$$R = \frac{1}{\beta_1 + \beta_2 / S}, \quad \beta_1 \geq 0, \beta_2 \geq 0$$

where R and S are the number of recruits and spawners respectively.

Stable population level

Consider the problem of maintaining a sustainable fishery. The total population abundance will only stabilize if $R = S$. The total population will decline if fewer recruits are produced than the number of spawners who died producing them. If too many recruits are produced, the population will also decline eventually because there is not enough food for them all. Thus, only a balanced level of recruits can be sustained indefinitely in a stable population. This stable population level is the point where the 45° line intersects the curve relating R and S. In other words, it is the N such that

$$N = \frac{1}{\beta_1 + \beta_2/N} \tag{1}$$

Solving for N we see that the stable population level is $N_{stable} = (1 - \beta_2)/\beta_1$.

Goals

In this exercise, you will fit a Beverton-Holt model to the `fish` dataset, by estimating the parameters β_1, β_2. Using this fitted Beverton-Holt model, you will then estimate the corresponding stable population level N by solving equation (1) for N with your estimated β_1, β_2. Finally, you will assess the uncertainty in the estimated stable population level by employing the bootstrap.

Exercises for fish data

1. (4 points) Make a scatterplot of the data and overlay the Beverton-Holt curve for a few different choices of β_1 and β_2. Hint: Write a function to compute R in the Beverton-Holt model and try values near (`2e-3`, `7e-1`):

```
bh <- function(S, beta1, beta2) {
  # TODO: Compute R as a function of S
}
qplot(S, R, data = fish) +
  stat_function(fun = bh, args = list(beta1 = ..., beta2 = ...)) +
  stat_function(fun = bh, args = list(beta1 = ..., beta2 = ...)) +
  stat_function(fun = bh, args = list(beta1 = ..., beta2 = ...))
```

2. (3 points) The Beverton-Holt model can be *linearized* by transforming $R \mapsto (1/R)$ and $S \mapsto (1/S)$:

$$(1/R) = \beta_1 + \beta_2 \times (1/S)$$

This is a linear model with response variable $(1/R)$ and covariate $(1/S)$. Use least squares regression to fit this model to the `fish` dataset:

```
bh_lm <- lm(I(1/R) ~ I(1/S), data = fish)
```

You can retrieve the vector of estimates of β_1, β_2 with the `coef()` function:

```
coef(bh_lm)
```

3. (3 points) Find an estimate for the stable population level, where $R = S$ in the Beverton-Holt model by plugging in your estimated coefficients into the formula for N_{stable}. The resulting estimate is a function of $\hat{\beta}_1$ and $\hat{\beta}_2$.

4. (5 points) Use the bootstrap to obtain the sampling distribution and standard error for the stable population level. You will need to write a function that encapsulates everything you did in the previous two questions. It should take a dataframe (similar to `fish`) as input and return an estimated stable population level based on the input:

```
N_hat <- function(df) {
  estimate <- # TODO: Compute estimate of N_stable based on df
  return(estimate)
}
```

5. (5 points) Read the description of percentile bootstrap confidence intervals from the bootstrap notes and then use the bootstrap to construct a 95% confidence interval for the stable population level.