

# Final Report

Seo Wook Choi

April 2017

## 1 Introduction

Among the variables in pink salmon data, I am going to use  $R$ ,  $S$ ,  $t$ , stocks and region. My research question is how much of ‘recruits’ returns to their hometown as a ‘spawner’ in various in different regions. Also, I want to find out if there is a difference in returning rate of salmon over the years. I set  $S$  as a response variable and the other variables as predictors. To fit the model to the data, I preprocessed the data to a manageable form. First, I split the data into even year spawning salmon and odd year spawning salmon. And then, I paired and realigned corresponding year of  $S$  to year of  $R$ . The years are adjusted to year of  $S$ . Also, as the original data has a mixture of different stocks, I discard some data which does not have a record of corresponding year’s (two-year gap) data. After, I re-bind to data vertically. The length of data decreased from  $N=1206$  to  $N=1121$ . I append new data set as an appendix. In the data,  $i=1, \dots, 519$  are even year spawning salmons and  $i=520, \dots, 1121$  are odd year spawning salmons. The  $S_i$  and The  $R_i$  are the corresponding pair in a two-year gap. The  $t_i$  are years of spawning,  $stock_i$  is the  $i_{th}$  observation stock and  $region_i$  is region of  $i_{th}$  observation region. So, if  $S_i$  is number of spawner in 1980,  $R_i$  is recruit in 1978. The  $stock_i$  and  $region_i$  are their hometown.

## 2 Model

I am mainly interested in how much recruiters are returning as spawner to their homeland. Also, I am interested in the difference in returning rate for different species. As the response variable  $S$  is count data, I am going to use Poisson regression model. Predictors will be stock, which has a hierarchical structure with the region, a number of recruits in two years before and the year of spawn. Following is a model specification of Poisson regression model.

$$p(S_1, \dots, S_{519} | \lambda_i) = \prod_{i=1}^{512} \frac{\lambda^{y_i} \exp(-\lambda_i)}{y_i}, \quad (1)$$

$$\log(\lambda_i) = \gamma_{stock_i} + \beta_{even, t_i} \log(R_i), \quad (2)$$

Similarly for odd species,

$$p(S_{520}, \dots, S_{1121} | \lambda_i) = \prod_{i=520}^{1121} \frac{\lambda^{y_i} \exp(-\lambda_i)}{y_i}, \quad (3)$$

$$\log(\lambda_i) = \gamma_{stock_i} + \beta_{odd, t_i} \log(R_i), \quad (4)$$

Prior distributions are specified as following.

$$\beta_{(even, t_i)} \sim N(\mu_{even}, \sigma_{even}^2), \quad t_i \in \{2, 4, \dots, 44, 46\}, \quad (5)$$

$$\beta_{(odd, t_i)} \sim N(\mu_{odd}, \sigma_{odd}^2), \quad t_i \in \{3, 5, \dots, 43, 45\}, \quad (6)$$

$$\gamma_i \sim N(\mu_{region_i}, \tau^2), \quad (7)$$

$$\gamma_i \sim N(\mu_{region_i}, \tau^2), \quad (8)$$

$$\mu_{odd} \sim N(0, 10000), \quad (9)$$

$$\mu_{even} \sim N(0, 10000), \quad (10)$$

$$\sigma_{odds}^2 \sim IG(0.001, 0.001), \quad (11)$$

$$\sigma_{even}^2 \sim IG(0.001, 0.001), \quad (12)$$

$$\mu_{region_1, \dots, 14} \sim N(0, 10000) \text{ and} \quad (13)$$

$$\tau_{region_1, \dots, 14}^2 \sim IG(0.001, 0.001). \quad (14)$$

Each stock and its region were modeled with hierarchy. The different region does not share global mean in this model. In the analysis, I will not look for individual stocks, but I will focus on regions (hyper parameter) that show difference from other regions.

I used log link function for the Poisson regression. Also, as  $S_i$  and  $R_i$  has same scale I used  $\log(R_i)$  instead of  $R_i$ . Without this transformation,  $\beta$  comes out with really small values. Also, I put random constant for different stock, but it is same across species type. This is because there might be an effect of different size or environment of stock, but two species are sharing same places, so I did not assign different intercept parameter for each species. Two species has different slopes in different years. The assumption of different slope in the different year is to see that if there is any change of returning rate over the years. Prior distributions for hyper parameters are vague priors using normal distribution (for mean) and inverse-gamma distribution (for variance).

The slope parameters can be interpreted as one unit of  $\log(recruit)$  increases there is change in  $\beta_t$  amount in  $\log(spawner)$ . Slopes are differently assigned to each species and year. The two hyper parameters  $\mu_{even}$  and  $\mu_{odd}$  may show differences of returning rate in even and odds species. Also, I am going to focus on change of  $\beta_t$  over the year.

All the  $\gamma_{stock}$  parameter work as intercept in the Poisson regression model. As the stocks are shared in both species, I assigned the same intercept for same stock regardless of species. The parameter  $\mu_{region}$  will provide information which region contribute to returning of spawners. Variance in different stocks has the same variance  $\tau^2$  which might be an unreasonable assumption. I used fixed variance for two reasons. First, I do not have an interest in variance of each stock. Second, a single parameter will make computational comfort.

### 3 result

First, I am going to talk of about convergence of parameters in the model. There is no problem with convergence in hyper parameters (Figure 1), but it seems there are some issues with lower-level parameters such as  $\gamma$  and  $\beta$

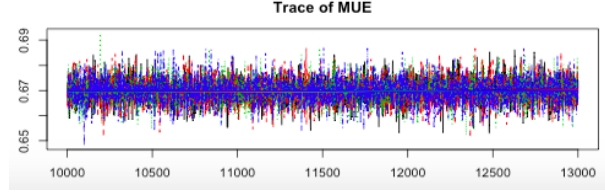


Figure 1: Example of convergence in one of the hyper parameters ( $\mu_{even}$ ).

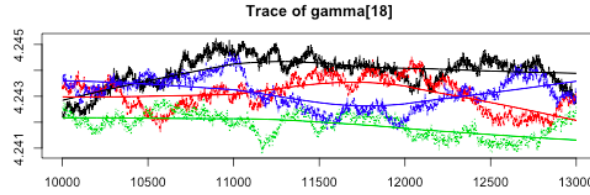


Figure 2: Example of convergence issue in one of the lower level parameters ( $\gamma_{18}$ ).

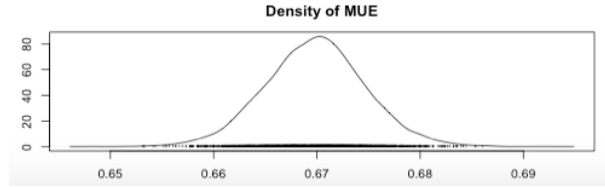


Figure 3: The posterior distribution of  $\mu_{even}$ .

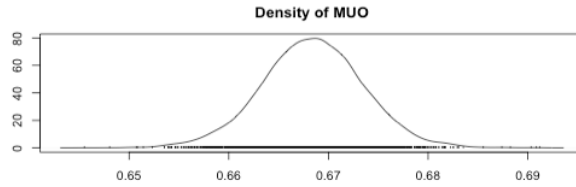


Figure 4: The posterior distribution of  $\mu_{odd}$ .

(Figure 2). However, even though the trace plot does not converge perfectly, it stays in a very small interval. Also, hyper parameters show perfect convergence ( $\hat{R} = 1.00$ ).

The posterior distributions of  $\mu_{even}$  (Five number summary: [0.6602 0.6667

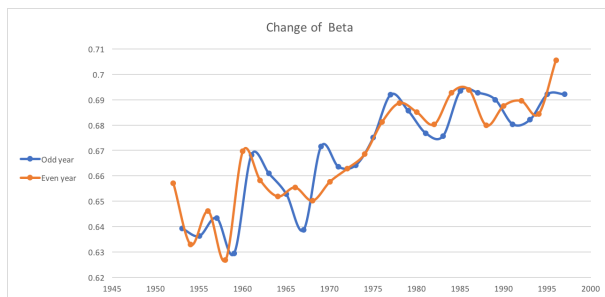


Figure 5: Posterior mean  $\beta$  in each year.

0.6699 0.6729 0.6797]) and  $\mu_{odds}$  (Five number summary: [0.6582 0.6651 0.6684 0.6717 0.6783]) shows that there is no huge difference of returning coefficient in different species of salmon. Also, I was interested in a change of returning rate across years for each species. The dots in Figure 5 are posterior mean of  $\beta_{even,t}$  and  $\beta_{odd,t}$ . Both species seems to have increase in their returning rate over the years.

For the aspect of regions, I found out the best region which contributes to the most return of salmon and the worst. The Southeast Alaska was the best for a return of salmon ( $\mu_{region12}$ , mean= 4.48, sd=0.2501). Prince William Sound was worst for return of salmon ( $\mu_{region10}$ , mean= 3.13, sd=0.2466).

## 4 Discussion

There is a possibility of better models. I tried several models, for example, the slope with the regional difference. However, they did not show better model fit than the model I proposed (DIC= 81290729). Adding intercept for each species had slightly better DIC=81263032), but it does not worthy enough to add more complexity on the model (Difference= 27697.01,sample standard error= 113603.4). One thing that I did not cover is an interaction between region and species. There might be an interesting interaction effect between region and species, but my model has a caveat that it cannot explain interaction effects.