BDA - Project

Jan Nyberg, Carl-Victor Schauman

4/12/2022

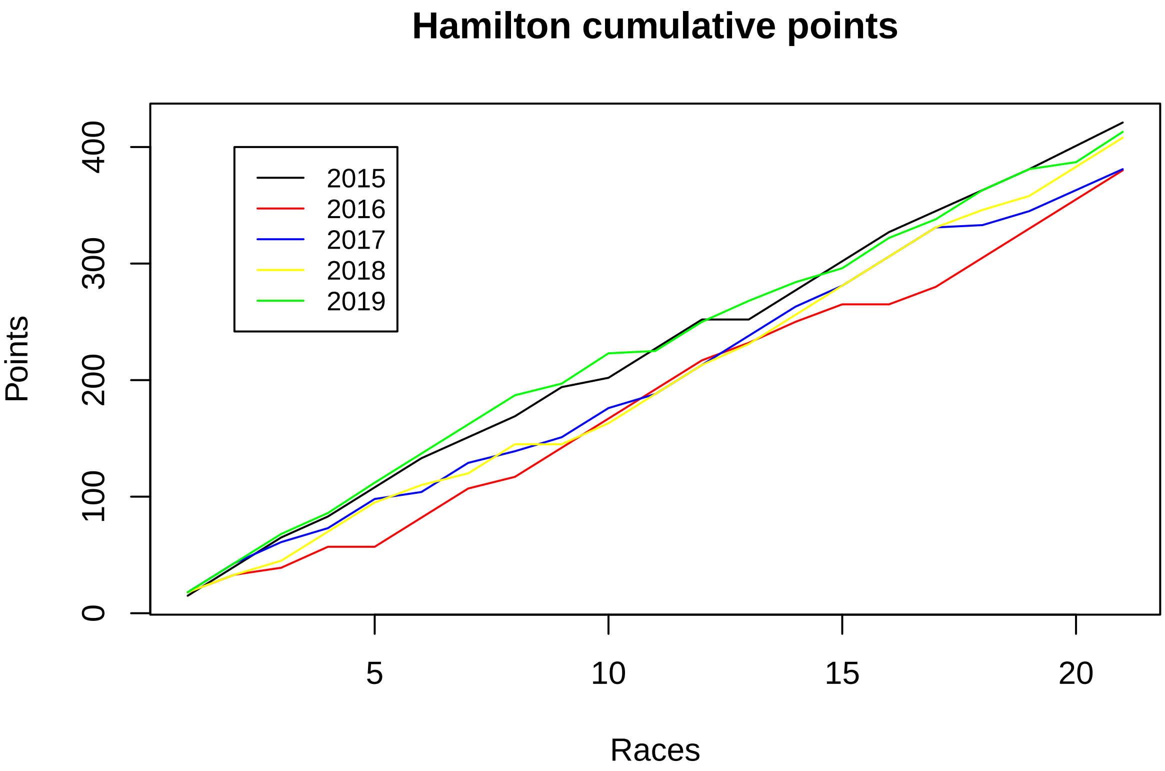
Contents

Introduction

For this project, we are trying to predict Lewis Hamilton's average score in a season based on previous years scores. This is mostly just due to our curiosity if we are able to use this to somehow predict the score. We are modeling his scores from five years, and trying to build a model using it. We want to see what kind of distribution the answer will be and how well it is able to estimate the following year. There are many factors we don't take into consideration, but we hope to see relatively good results and predictions.

Description of the data

The data we use is from Kaggle and can be found here. We took Lewis Hamilton out of the data and chose the years 2015-2019. We selected his scores from all the races from those years. One thing to note with the data is that every year doesn't have an equal amount of races. To account for this we chose to fill in the missing races with the median for the year. This makes the data a bit inaccurate, however, it shouldn't have too big an effect on the data.



Description of models

Priors used

Rstan code

Below is the code for the hierarchical model for Hamilton's points.

```
data {
  int<lower=0> N;
  int<lower=0> J;
  real U;
  vector[J] y[N];
  real<lower=0> mu_s;
  real<lower=0> sigma_prior;
}

parameters {
  real<lower=0> mu;
  real<lower=0> sigma;
  real<lower=0> tau;
  vector[J] mus;
}

model {
  mu ~ normal(0, mu_s);
  tau ~ inv_chi_square(sigma_prior);
  sigma ~ inv_chi_square(tau);
  mus ~ normal(mu, tau);

  for (j in 1:J)
    y[,j] ~ normal(mus[j], sigma);
}

generated quantities {
  vector[J] log_lik[N];
  real ypred;
  real ypred_6;
  ypred = normal_rng(mus[5], sigma);
  ypred_6 = normal_rng(mu, sigma);
  for (j in 1:J){
    for (n in 1:N){
      log_lik[n,j] = normal_lpdf(y[n,j] | mus[j], sigma);
    }
  }
}
```

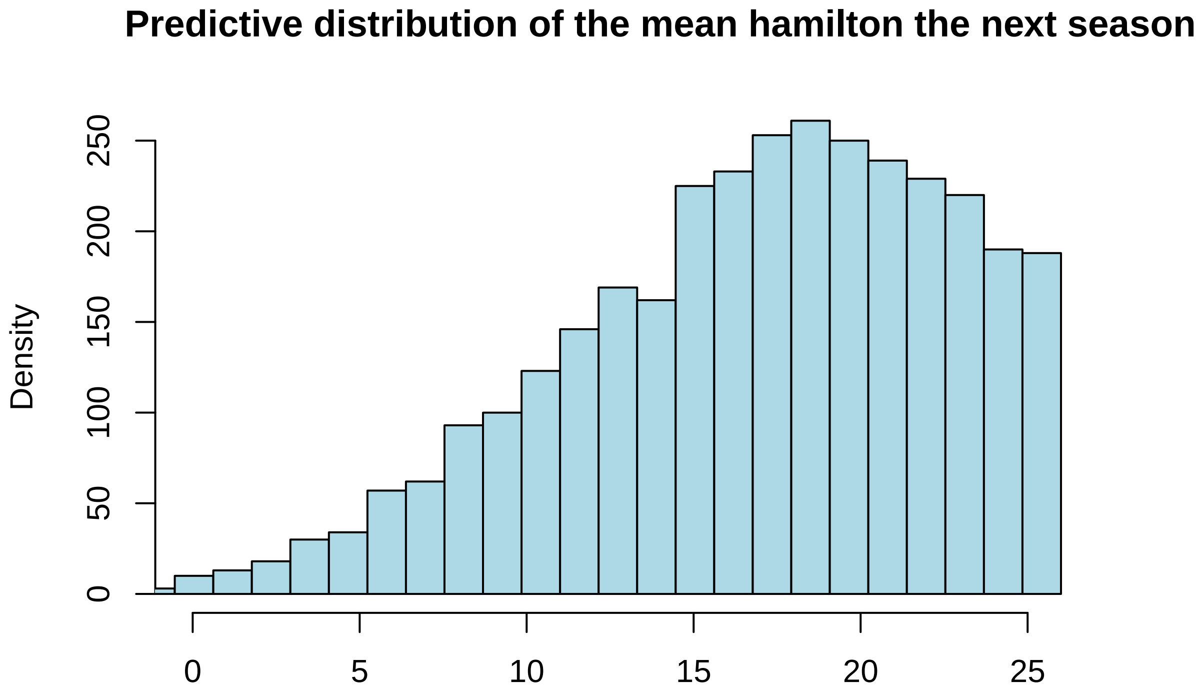
Running of stan model

Below is the hierarchical model with the corresponding histogram with the data.

```
hier_data = list(
  y = ham_data,
  N = nrow(ham_data),
  J = ncol(ham_data),
  U = 26,
  mu_s = 20,
  sigma_prior = 7
)

hier_fit = sampling(
  hier_ham,
  data = hier_data,
  chains = 4,
  iter = 2000,
  warmup = 1000,
  refresh = 0
)
```

```
## Warning: There were 134 divergent transitions after warmup. See
## https://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.

## Warning: Examine the pairs() plot to diagnose sampling problems

## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## https://mc-stan.org/misc/warnings.html#bulk-ess

## Warning: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quant
## Running the chains for more iterations may help. See
## https://mc-stan.org/misc/warnings.html#tail-ess
```

```
ham_extracted = extract(hier_fit)$ypred
ham_extracted26 = ham_extracted[ham_extracted < 26]

hist(ham_extracted26, breaks = seq(min(ham_extracted26), max(ham_extracted26), length.out = 30),
     xlab="", ylab="Density",main="Predictive distribution of the mean hamilton the next season",
     col="lightblue", xlim=c(0,28))
```



Convergence diagnostics

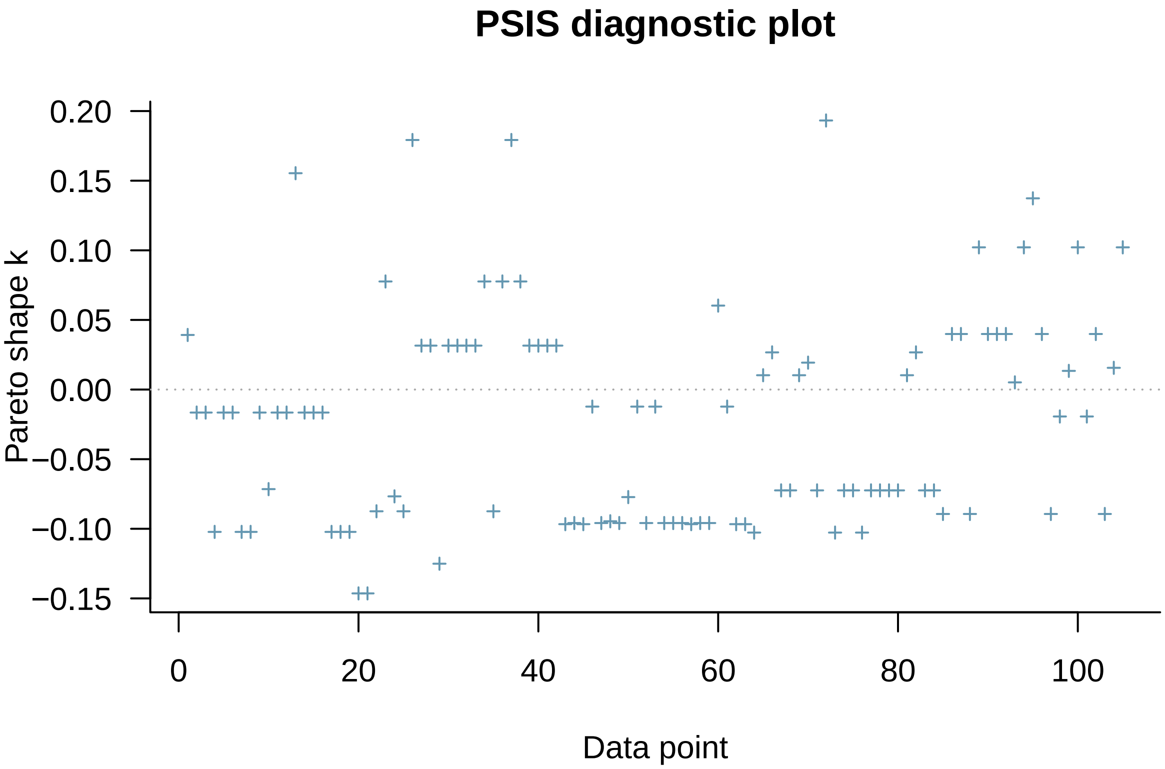
```
hier_extracted = extract_log_lik(hier_fit, merge_chains = FALSE)
r_eff = relative_eff(exp(hier_extracted))
hier_loo = loo(hier_extracted, r_eff = r_eff)
hier_elpd = hier_loo$estimates["elpd_loo",1]
hier_loo
```

```
##
## Computed from 4000 by 105 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo   -357.6  7.9
## p_loo         2.0  0.5
## looic       715.2 15.9
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
paste("Hierarchical model PSIS-LOO elpd value: ", round(hier_elpd,1))
```

```
## [1] "Hierarchical model PSIS-LOO elpd value:  -357.6"
```

```
plot(hier_loo)
```



Posterior predictive checks

Predictive performance assessment

Sensitivity analysis

Model comparison

Discussion of issues and potential improvements

Conclusion what was learned from the data analysis

Self-reflection