Title

April 5, 2024

Leo Kraushaar & Ali Hay
STAT 413

1 Introduction

For this project, we are working with a data set containing the number of protests each month from January 2022 until November 2023 in every Canadian province and territory. Our goal in this project is to model the number of protests using our data. Next, we use our model to create confidence intervals, with different bootstrapping methods to determine the most significant parameters in our model. Finally, we use Monte Carlo simulation to create prediction intervals for the monthly number of protests in each province and territory based on projections for retail sales that we created for 2030. With our dataset, we were also given the population size of each province in that month, since it is likely that a province containing more people would have more protests, making it likely that population might be a useful predictor in our model. We also thought it might be interesting to consider the state of the economy in each month over our timeframe, since this could affect how happy individuals are, which could perhaps influence how likely they are to participate in protests. For this reason, we decided to collect some extra data related to economic factors: oil imports in cubic meters, total power generated in megawatt hours, and total retail sales in thousands of dollars. With our dataset, we decided to fit a Negative Binomial Model, as this is an effective model for count data. We decided to combine the months into four seasons, as this improved our model. After fitting our initial model, we found season, retail, and province to be significant, causing us to keep these as the predictors in our final model. Next, we use resampling bootstrapping, parametric bootstrapping, smooth bootstrapping and error-sampling bootstrapping to create 95% confidence intervals, telling us which of the model parameters were most significant. Lastly, we used Monte Caro methods to create 95% prediction intervals for the median number of protests in each Canadian province and territory based on our projected retail sales in the year 2030. We created these retail sales projections using more data on the total retail sales as discussed in later sections of this report. Our main algorithms and results are given, summarized, and analyzed throughout this report.

2 Model Building

For this project, we decided to use a Negative Binomial Generalized Linear Regression Model, as the number of protests appeared to follow a Negative Binomial distribution, and this model works well to model count data. We initially fit our model using the predictors: year, season, province, population, retail, oil, food, and power. Using a stepwise selection method in R, attempting to minimize the AIC of the model, we created our final model, which contained the significant predictors of season, province, and retail.

2.1 Fitting Process

2.2 Model Overview

3 Bootstrapping for Feature Significance

3.1 Resampling Bootstrap

Procedure

Using $B = 10000$, an algorithm was written to perform a resampling bootstrap on the dataset. On each iteration, a sample of size n (where n was the number of rows in the dataset) was taken with replacement. Using this sample, a new model was fit. The parameter estimates were recorded. Relevant code is shown below.

```r
resampBoot <- function(df, B) {
    # Get sample size
    n <- nrow(df)
    # Initialize empty dataframe
    params <- c()
    # Initialize progress bar
    bar <- txtProgressBar(min=0, max=B, style=1)
    # Perform B iterations
    for (b in 1:B) {
        # Select a sample of size n
        indices <- sample(1:n, replace = TRUE)
        samp <- df[indices, ]
        # Fit the model with the sample
        boot_model <- glm.nb(protests ~., data=samp, init.theta = 10)
        boot_params <- coef(boot_model)
        params <- rbind(params, boot_params)
        setTxtProgressBar(bar, b)
    }
    close(bar)
    return(params)
}
```

Results

Using the resampling bootstrap techniques, when testing to see whether there was a significant difference in the mean number of protests in each of the different seasons, our baseline season was fall. The 95% confidence interval for the spring coefficient did contain 0, but the 95% confidence intervals for each of the summer and the winter coefficients did not contain 0. This tells us that the mean number of protests in the spring is not significantly different in the spring than the fall, but the mean number of protests in both the summer and the winter is significantly different than in the fall. Since each of the coefficients for summer and winter are negative, that tells us that there are less protests in both the summer and the winter than there are in the fall. When using this resampling bootstrap technique to test for a significant difference in the mean number of protests in each of the different provinces, our baseline province was Alberta. Here, the 95% confidence intervals for each province and territory showed us that the mean number of protests were significantly different from that of Alberta. British Columbia, Ontario, and Quebec each had a larger mean number of protests than Alberta, whereas the other provinces and territories all had a lower mean number of protests than Alberta. Using the resampling bootstrapping techniques, the 95% confidence interval for the coefficient on retail showed us that retail did not have a significant effect on the mean number of protests.

	mean	sd	2.5%	50%	97.5%	sig
intercept	3.66	0.53	2.70	3.64	4.77	TRUE
Spr	-0.06	0.08	-0.23	-0.06	0.09	FALSE
Sum	-0.55	0.08	-0.72	-0.55	-0.38	TRUE
Win	-0.24	0.10	-0.42	-0.23	-0.05	TRUE
BC	0.82	0.14	0.55	0.82	1.09	TRUE
M	-1.89	0.93	-3.85	-1.85	-0.18	TRUE
NB	-2.60	1.06	-4.84	-2.55	-0.66	TRUE
NL	-3.04	1.13	-5.43	-2.98	-0.99	TRUE
NT	-5.40	1.37	-8.20	-5.34	-2.94	TRUE
NS	-2.38	1.00	-4.53	-2.34	-0.54	TRUE
N	-4.98	1.28	-7.68	-4.92	-2.61	TRUE
O	5.54	2.37	1.18	5.40	10.58	TRUE
PEI	-4.10	1.23	-6.73	-4.04	-1.82	TRUE
Q	2.17	0.88	0.55	2.12	4.04	TRUE
S	-2.56	0.96	-4.57	-2.52	-0.77	TRUE
Y	-4.11	1.26	-6.77	-4.06	-1.81	TRUE
retail	-1.83	1.04	-4.05	-1.77	0.08	FALSE

3.2 Parametric Bootstrap

Procedure

Using $B = 10000$, an algorithm was written to perform a parametric bootstrap on the dataset.

Using the estimated dispersion parameter $\theta \approx 8.36$, each iteration sampled a random vector from a negative binomial distribution. The distribution had dispersion parameter θ and mean $\hat{y}$, where $\hat{y}$ was the predicted mean value for the corresponding input values.

From the new estimates, a model was fit on each iteration and the parameter estimates were recorded. Relevant code is shown below.

```r
conditionalNegBinom <- function(theta, mu) {
    nb_sample <- rnbinom(size=theta, mu=mu, n=1)
    return(nb_sample)
}

paramBoot <- function(B, X, yhat, theta, func) {

    # Initialize empty vector
    params <- c()
    # Iterate B times
    for (b in 1:B) {
        # Simulate NB given means
        sim_y <- sapply(yhat, function(y) func(theta, y))
        # Add to the dataframe
        sim_data <- cbind(X, protests=sim_y)
        # Fit the model to the simulated data
        sim_model <- glm.nb(protests ~., data=sim_data, init.theta =
            theta)
        # Access the coefficients and store
        parameters <- coef(sim_model)
        params <- rbind(params, parameters)
    }
    return(params)
}
```

Results

When using parametric bootstrapping techniques, we came to exactly the same conclusions as we did when using the resampling bootstrapping method.

	mean	sd	2.5%	50%	97.5%	sig
intercept	3.68	0.51	2.68	3.68	4.69	TRUE
Spr	-0.07	0.08	-0.23	-0.07	0.10	FALSE
Sum	-0.55	0.09	-0.72	-0.55	-0.38	TRUE
Win	-0.23	0.09	-0.41	-0.23	-0.05	TRUE
BC	0.82	0.16	0.51	0.82	1.14	TRUE
M	-1.92	0.92	-3.75	-1.91	-0.11	TRUE
NB	-2.63	1.05	-4.72	-2.63	-0.57	TRUE
NL	-3.08	1.11	-5.31	-3.06	-0.89	TRUE
NT	-5.41	1.27	-7.92	-5.39	-2.91	TRUE
NS	-2.42	1.00	-4.39	-2.41	-0.46	TRUE
N	-5.01	1.27	-7.52	-5.01	-2.53	TRUE
O	5.61	2.45	0.80	5.59	10.51	TRUE
PEI	-4.13	1.22	-6.54	-4.12	-1.73	TRUE
Q	2.20	0.93	0.36	2.19	4.07	TRUE
S	-2.59	0.94	-4.46	-2.57	-0.75	TRUE
Y	-4.15	1.25	-6.61	-4.14	-1.71	TRUE
retail	-1.86	1.07	-3.98	-1.85	0.23	FALSE

3.3 Smooth Bootstrap

The smooth bootstrap is not an "ideal" method for the given dataset, as only one predictor **(retail)** was continuous and real-valued. However, results were consistent with other methods, as discussed later.

Again using $B = 10000$, an algorithm was written to perform a smooth bootstrap. The **retail** column was found to have a sample variance of 1, due to the fact that it was standardized prior to model building. A reasonable value for the noise term was chosen, that is, $\frac{1}{\sqrt{n}} \approx 0.05783$. On each iteration, some ε_i was added to each row i, where $\varepsilon \sim N(0, 0.05783)$. Using this "new" dataset, a model was fit and the paramter estimates were recorded. Relevant code is shown below.

Procedure

```
addNoise <- function(X) {

    cols <- colnames(X)
    new_X <- X
    for (col in cols) {
        Xi <- X[, col]
        if (class(data[, col]) != "factor") {
            n <- length(Xi)
            S_sq <- var(Xi)
            noise_var <- S_sq / n
            new_X[, col] <- Xi + rnorm(n=n, mean=0, sd=sqrt(noise_var))
        } else {
            new_X[, col] <- Xi
        }
    }
    return(new_X)
}

smoothBoot <- function(X, y, B, noisefunc) {

    # Get sample size
    n <- nrow(X)
    # Initialize empty vector
    params <- c()

    # Initialize progress bar
    pb <- txtProgressBar(min = 0, max = B, style = 3)

    # Perform B iterations
    for (b in 1:B) {
        # Update progress bar
        setTxtProgressBar(pb, b)

        # Get new dataset
        new_X <- noisefunc(X)
        new_data <- data.frame(protests=y, new_X)

        # Fit the model with the simulated data
        smoothboot_model <- glm.nb(protests ~., data=new_data, init.theta = 5)
        boot_params <- coef(smoothboot_model)
        params <- rbind(params, boot_params)
    }

    # Close progress bar
    close(pb)

    return(params)
}
```

Results

When using smooth bootstrapping techniques, the results of our 95% confidence intervals yielded similar conclusions to that of the first two methods, but there were some differences. This method showed us that the mean number of protests was significantly lower in each of the other three seasons than in the fall. Additionally, using smooth bootstrapping techniques did also lead us to slightly different conclusions when comparing the mean number of protests in each of the provinces and territories. Here we found that each province, except for Manitoba, had a significantly different mean number of protests than Alberta. Again, British Columbia, Ontario and Quebec had more mean protests than Alberta, whereas each of the others had less. This method also showed us that retail did not have a significant effect on the mean number of protests.

3.4 Error-Sampling Bootstrap

Another bootstrap method was implemented, in which the error terms from the fitted model were randomly sampled with replacement, and added to the fitted values. Notably, some resulting simulated counts were rounded up to zero in the case where a negative value was produced. This was required both logically; as protests counts cannot be negative, and mathematically; as the negative binomial glm cannot be fit with negative training outputs. As a result, the integrity of the simulated datasets was not assumed to be completely intact. That being said, the results were quite consistent with the previous methods.

Procedure

```
epsilonBoot <- function(X, model, B, errors) {

    # Get sample size
```

	mean	sd	2.5%	50%	97.5%	sig
intercept	3.00	0.21	2.58	3.01	3.42	TRUE
Spr	-0.04	0.01	-0.07	-0.04	-0.03	TRUE
Sum	-0.55	0.01	-0.56	-0.55	-0.54	TRUE
Win	-0.20	0.01	-0.23	-0.20	-0.18	TRUE
BC	0.67	0.05	0.57	0.67	0.76	TRUE
M	-0.67	0.39	-1.43	-0.68	0.10	FALSE
NB	-1.21	0.44	-2.08	-1.22	-0.33	TRUE
NL	-1.56	0.47	-2.49	-1.57	-0.62	TRUE
NT	-3.68	0.52	-4.70	-3.68	-2.63	TRUE
NS	-1.07	0.42	-1.89	-1.08	-0.23	TRUE
N	-3.30	0.53	-4.33	-3.30	-2.24	TRUE
O	2.24	1.05	0.14	2.25	4.29	TRUE
PEI	-2.47	0.51	-3.48	-2.48	-1.45	TRUE
Q	0.92	0.40	0.13	0.92	1.69	TRUE
S	-1.31	0.40	-2.09	-1.32	-0.52	TRUE
Y	-2.46	0.52	-3.49	-2.47	-1.41	TRUE
retail	-0.39	0.45	-1.29	-0.40	0.52	FALSE

```r
    n <- nrow(X)
    # Initialize empty vector
    params <- c()
    # Perform B iterations
    for (b in 1:B) {
        # Get errors
        errs <- sample(errors, size=n, replace=TRUE)
        # Get fitted values
        yhat <- fitted(model)
        # Get simulated y
        ystar <- yhat + errs
        # round up negative values
        ystar <- pmax(rep(0, n), ystar)
        # Turn into DataFrame
        sim_data <- data.frame(protests=ystar, X)
        # Fit the model with the simulated data
        paramboot_model <- glm.nb(protests ~., data=sim_data, init.theta = 5)
        boot_params <- coef(paramboot_model)
        params <- rbind(params, boot_params)
    }
    return(params)
```

```
}
```

Results

Using the error-sampling bootstrap techniques also produced 95% confidence intervals that lead to slightly different conclusions than the other methods. As with the smooth bootstrapping technique, this method showed us that the mean number of protests was significantly lower in each of the other three seasons than in the fall. When comparing the mean number of protests in each of the Canadian provinces and territories, we came to the same conclusions using this method as both the resampling and parametric bootstrapping techniques. Unlike the other three bootstrapping methods however, our 95% confidence interval using error-sampling bootstrapping showed us that the retail sales did have a significantly negative effect on the mean number of protests.

	mean	sd	2.5%	50%	97.5%	sig
intercept	3.66	0.05	3.56	3.66	3.76	TRUE
seasonSpring	-0.07	0.01	-0.09	-0.07	-0.04	TRUE
seasonSummer	-0.56	0.02	-0.59	-0.56	-0.53	TRUE
seasonWinter	-0.24	0.01	-0.27	-0.24	-0.21	TRUE
BC	0.82	0.02	0.78	0.82	0.86	TRUE
M	-1.89	0.09	-2.07	-1.89	-1.71	TRUE
NB	-2.61	0.11	-2.83	-2.61	-2.40	TRUE
NL	-3.07	0.12	-3.30	-3.07	-2.83	TRUE
NT	-5.20	0.28	-5.80	-5.18	-4.70	TRUE
NS	-2.40	0.10	-2.60	-2.40	-2.20	TRUE
N	-4.96	0.25	-5.50	-4.95	-4.51	TRUE
O	5.54	0.24	5.07	5.54	6.02	TRUE
PEI	-4.16	0.18	-4.51	-4.15	-3.83	TRUE
Q	2.17	0.09	1.99	2.17	2.35	TRUE
S	-2.58	0.10	-2.78	-2.57	-2.38	TRUE
Y	-4.18	0.17	-4.54	-4.18	-3.85	TRUE
retail	-1.82	0.10	-2.03	-1.82	-1.62	TRUE

3.5 Method Comparison

Using the different Bootstrapping techniques did yield similar conclusions, however there were a few differences, as discussed in each of the results sections above. We can see from the Z-scores based on the resampling bootstrapping that parameters with the highest absolute value of Z-score were summer, followed by British Columbia, then the Northwest Territories, telling us that these were the most significant parameters in our model. The parametric bootstrapping method yielded the same top three most significant parameters as the resampling method. With the smooth bootstrapping method, summer was again the most significant parameter, however it was followed by winter, and then British Columbia. Finally, using error-sampling bootstrapping, British Columbia was the most significant parameter, followed by Summer, and then Saskatchewan. Using the error-sampling technique, our Z-scores were much higher in absolute value than the other methods, although smooth bootstrapping did produce some high z-scores. Each of our bootstrapping methods did give similar results, with the resampling method and the parametric being the most similar, while the error-sampling method gave the most unique results.

4 Monte Carlo Estimation

Methods

Using data from Statistics Canada of the total retail sales in thousands of dollars in each Canadian province and territory each month from January 2017 until November 2023, we fit a linear model predicting the total retail sales using a stepwise selection method. We used this model to create 95% prediction intervals for the mean retail sales in each province for every month in the year 2030. Uniformly across these intervals we randomly sampled values to use as the projected retail sales. Sampling uniformly across this interval may be questionable, but we will use this technique for the purposes of this project. For each of these projected retail sales, we sampled from our negative binomial model to predict each province's mean number of protests each season. Combining the seasons, we predicted the median number of protests in each Canadian province and territory based on our projected retail sales with 95% prediction intervals.

Results

Here we compare 95% Monte Carlo prediction intervals with the observed number of protests in each province during the year 2023. Each of the confidence intervals, except for Quebec and Ontario are entirely below the observed number of protests in 2023. This is to be expected as total retail sales have a negative association with number of protests, which we can see from the negative estimate of the parameter, meaning that as the total retail sales increase, the number of protests is expected to decrease. In 2030, it is expected that the total retail sales will increase, so this increase in retail sales explains the typical predicted decrease in protests. However, Ontario and Quebec each have a positive association with the number of protests, so the increase in total retail sales is not enough to lower the predicted number of protests in our model. Note that British Columbia also has a positive association with the number of protests, but it is not as strong as that of Ontario and Quebec, which is shown by its lower parameter estimate. This explains why the positive association between British Columbia and the number of protests is not enough to increase the predicted number of protests based on the increased projection in the total retail sales in 2030 using our model.

5 Conclusion

As discussed in previous sections, for this project we fit a Negative Binomial Generalized Linear Regression Model which in its final form contained the significant predictors of season, retail, and province. We used this model to perform resampling bootstrapping, parametric bootstrapping, smooth bootstrapping and error-sampling bootstrapping to create 95% confidence intervals. In all four cases, summer and British Columbia were in the top three most significant parameters. Finally, we created projections for retail sales in 2030 and used these to create 95% Monte Carlo prediction intervals for the median number of protests in each Canadian province and territory. These projected increases in sales lead to decreased predictions for the number of protests in each province, except for the two in our model with the largest positive parameter estimates, Ontario and Quebec. Overall, using methods learned in STAT 413, we were able to come to interesting conclusions regarding Canadian protests.