Overdispersion in Count Modelling with an NHL Draft Application: Measuring Player Success

Sean Farquharson (sfarqu2@uwo.ca)

Supervised by
Dr. Douglas Woolford

A Major Research Project report submitted in partial fulfillment of the requirements for the Master of Statistics degree

Department of Statistical and Actuarial Sciences
University of Western Ontario
London, Ontario, Canada

April 2021

Table of Contents

Abstract

Many hockey players are drafted by National Hockey League (NHL) teams each year. Unfortunately, some drafted players never actually get to play in an NHL game. The objective of our research is to use statistical learning to characterize successful players through an analysis of historical NHL draft data. We first fit count models to predict the number of games played in the NHL from the time a player is drafted. However, the mean-variance relationship in the Poisson model is violated due to excess zeros causing overdispersion. Statistical tests for overdispersion are implemented in R to show why common count models are not appropriate for such a situation. We briefly discuss overdispersion in the case of zero-inflated data and outline modelling solutions to this problem. We investigate this issue using visual and numerical diagnostics and account for this in the model building process. A Hurdle model (two-part model) is fitted to the data to address the issue of our zero-inflated data set. We concluded that a Hurdle model framework is best fitting for our data, as it addresses the issue of overdispersion due to zero-heaviness by combining the power of a Bernoulli and a zero-truncated count distribution. I also discuss the experiential learning received during my internship as a student junior statistician for Statistics Canada in Ottawa, Ontario where I worked in the Economic Statistics and Methods Division (ESMD) on the Special Business projects team. I worked on multiple projects revolving around the Ontario First Nations Point of Sale Exemption Survey and the Survey of Commercial and Institutional Energy Use.

Chapter 1: Overdispersion in Count Modelling with an NHL Draft Application: Measuring Player Success

1.1 Introduction

Before each NHL season, teams are forced to draft players at a very young age with hopes that they will fill a role as an NHL player. Often, we look back at NHL drafts and notice obvious distinctions between the players abilities that were not present at the time of the draft. This seems to be a recurring event as there probably isn’t one draft you can name where you wouldn’t select the players in a different order. This raises questions about the draft: Are players being drafted at too young of an age? Should the NHL be more concerned about the development of these players? We may not be able to answer these questions directly by statistical analysis, but further insight can be gained here and questions like these can spark an interesting debate.

Most of the time, when a player is drafted to an NHL team, they don’t get to play immediately in the season following their draft. As a result, many players can be pushed back into a team’s farm system, where they usually play in either the American Hockey League (AHL) or the East Coast Hockey League (ECHL), ranked as the second and third best North American leagues respectively (Wolski, D., 2021). It can be argued that in some cases, this can have a negative effect on players as it can make them feel as if they are not good enough to play in the NHL, draining their confidence as a result. Although most players that get drafted are sent back to their junior team, their National Collegiate Athletic Association (NCAA) team, or their European league team, this is more of a concern for players who have been drafted higher that have great expectations (Pollard, 2017). The immense amount of pressure on a player at such a young age could have some negative effects on them before they even begin their career (Pollard, 2017). After all, It only took one year from the time Jesse Puljujarvi was drafted for him to be referred to as a “bust” for some time. So, should the NHL allow players to play, for example, until they are 20 years old before entering the NHL draft? Perhaps if they did allow it, they may be competing at an even higher level, working towards their goal of being drafted by an NHL team. It can be argued that the effect it is having now is that they already feel accomplished at the age of 18, being drafted by an NHL team, not understanding the challenges that lie ahead. Not to mention, only a small percentage of players who have been drafted move on to playing in lengthy NHL careers (Odds, 2009).

Instead of filling rosters spots with their draft picks, many NHL teams will move on to signing undrafted players out of NCAA, junior, and leagues across Europe. These players that they are signing are typically anywhere from 22-25 years of age and have developed greatly since the age of 18 (Kirk, 2017). The NHL draft seems more of like a gamble at this point, and although the NHL may want it to remain this way, it can still be argued that there is much room for improvement. I believe that with the help of data analytics, further insight can be gained as to not only how the league can improve the draft, but how NHL teams should be drafting players (or what type of players they should be going for). Since it seems like NHL teams do not decide on most players until around the age of 22-25 but they draft players at the age of 18, this also raises concerns about player development at younger ages, specifically around the age of 15-18. For example, the players who do not get drafted by a major junior team have already felt a sense of failure at the age of 15. However, this can be left for a different discussion.

Once again, the negative impact the draft can have on some players sometimes cannot seem to be reversed. There have been many NHL draft picks referred to as “busts” because they were drafted high and did not meet their expectations (Sapunka, 2017). The result is that it has a negative impact on the player and the team who drafted them, and a hockey fan base may hold this against their team for years. This is the result of drafting players at such a young age and unfortunately it has dismantled some teams like the Edmonton Oilers who have struggled with drafting players.

1.2 Problem and Objective

Many NHL teams are not drafting to the best of their abilities and many great NHL players have been drafted in later rounds, while teams could have drafted them much earlier. Our goal is to fit a model to available draft data in order to attempt to identify those prospects who are most likely going to have successful NHL careers. This could aid the effectiveness of future drafting of players. However, as we will demonstrate, such a model must account for the zero-heaviness and overdispersion that is evident in the data.

1.3 Data

The dataset collected for this project was pulled directly from GitHub and was uploaded as an open source by Schulte (2018). These records contain player data that was scraped from HockeyReference.com. Table 1 summarizes the dataset we are using for this project.

Table 1: Description of Variables in the Dataset

Variable	Variable Description
DraftAge	The age of the player in their draft year.
Country	The country the player is from.
Country_group	The country group the player is from, the groups are Canada, USA, and Europe.
Height	The height of the player.
Weight	The weight of the player.
Position	The position the player plays.
DraftYear	The year the player was drafted in.
Overall	The position at which the player was drafted.
CSS_rank	The draft ranking assigned by NHL central scouting.
rs_GP	The number of regular season games played.
rs_G	The number of regular season goals scored.
rs_A	The number of regular season assists.
rs_P	The number of regular season points.
rs_PIM	The number of regular season penalty minutes.
rs_PlusMinus	The players regular season plus/minus
po_GP	The number of games played in the playoffs.
po_G	The number of goals scored in the playoffs.
po_A	The number of assists in the playoffs.
po_P	The number of points in the playoffs.
po_PIM	The number of penalty minutes in the playoffs.
sum_7yr_GP	The total number of NHL games played 7 years from the time they were drafted.
sum_7yr_TOI	The total time on ice in the NHL 7 years from the time they were drafted.
GP_greater_than_0	An indicator for whether they played an NHL game or not. “0” = no games played and “1” = one or more game(s) played.

For this project, there are two variables that could potentially be our response variable. The first one is total number of NHL games played and the other is the total time on ice in the first seven years of their NHL career. Keep in mind that the data contains only players that have at least seven years played in the NHL. With having at least seven years played, we know that the player has had a well-established NHL career, and these are the players that we are interested in. We have decided to use the total number of NHL games played as the response variable since we believe it is the better measure for player success. While using ice time can be effective, there is usually a big difference in ice time between first line and fourth line players; even though the fourth line players do not get as much ice time, they are still an important part of the team and have successful careers.

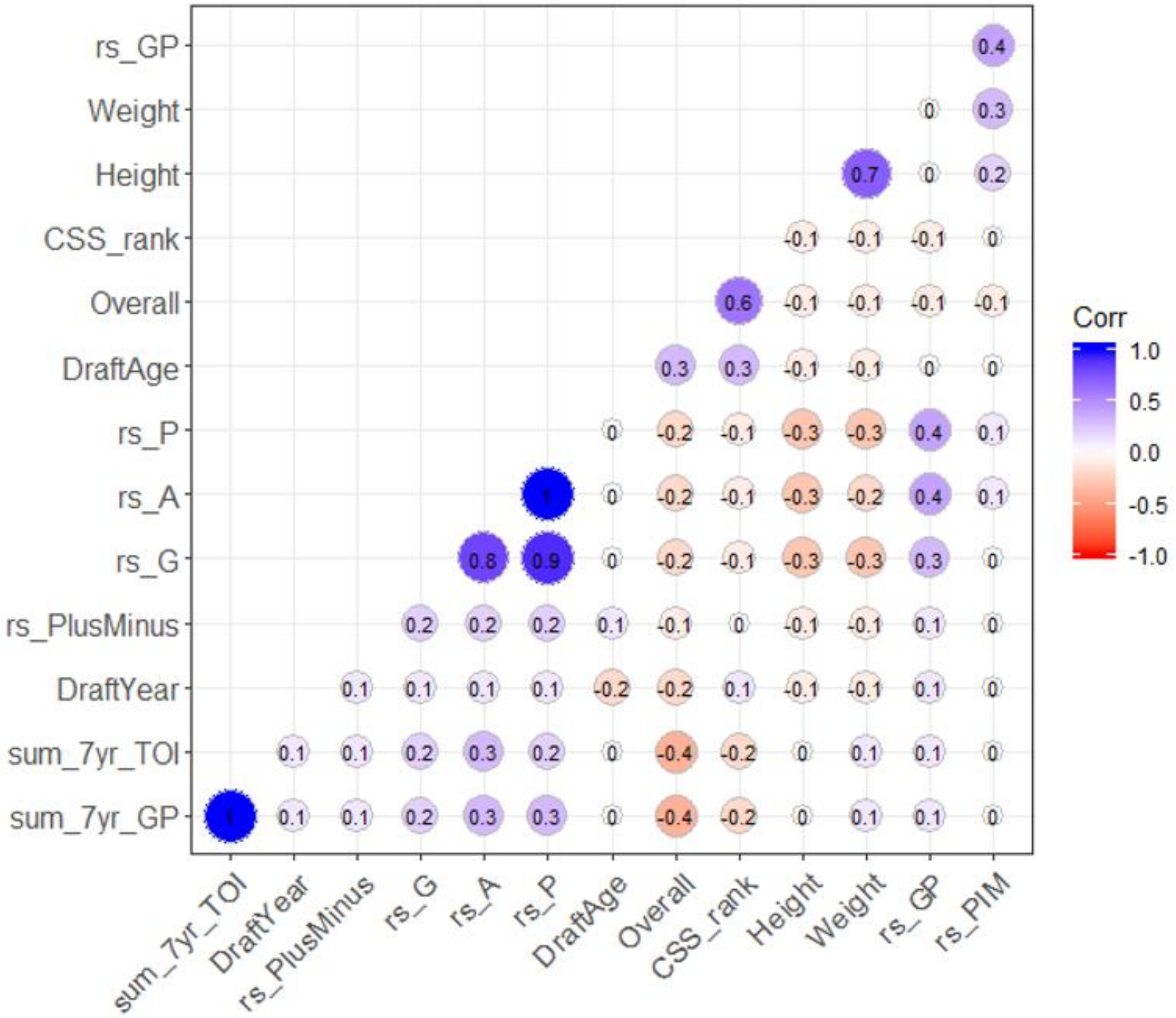

Figure 1: Correlation plot between each variable in the data set, using the "ggcorrplot" library in R. Each correlation is labelled, and colour coded according to the scale on the right-hand side.

As we can see from Figure 1, the correlation between goals, assists, and points are quite high so we will drop points from the data set. We decided to drop points since goals and assists are more indicative of what type of player they are in terms of a goal scorer or a playmaker. This avoids potential issues with multicollinearity as we move forward with the analysis.

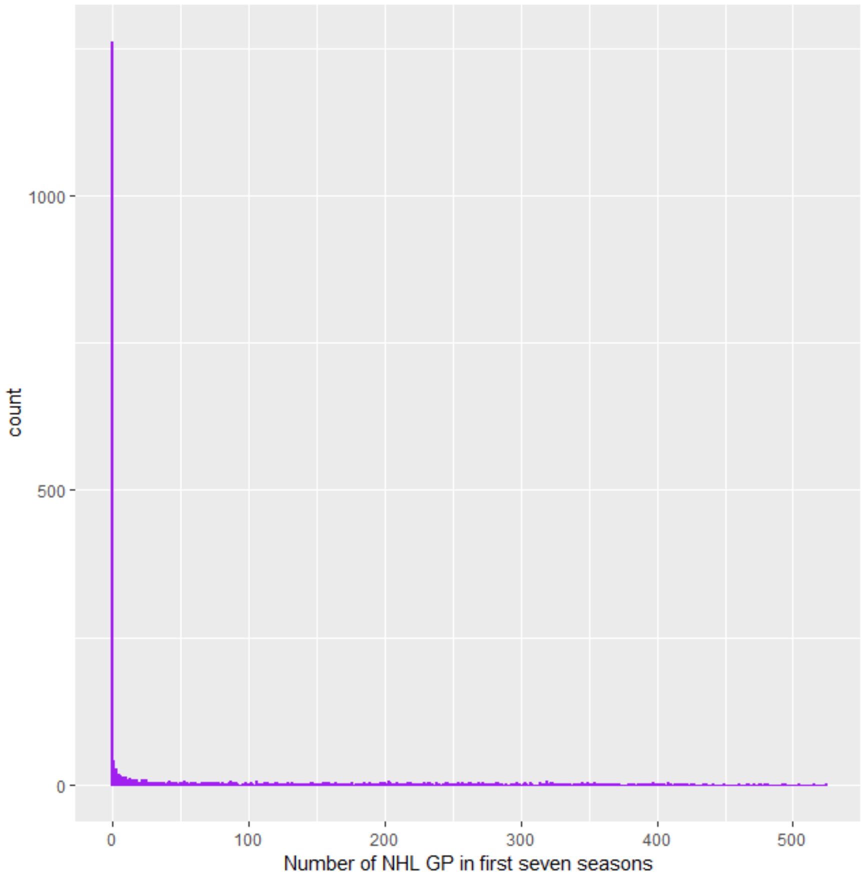

Figure 2: Histogram of the number of NHL games played in a drafted player's first seven seasons.

Figure 2 indicates that there is zero inflation in our response variable. This is also commonly referred to as zero-heaviness. It is important for us to only consider statistical modelling that is suitable for this type of response variable. Figure 3, which is a histogram of the counts with the zeros removed, further illustrates how many drafted players end up playing very few games.

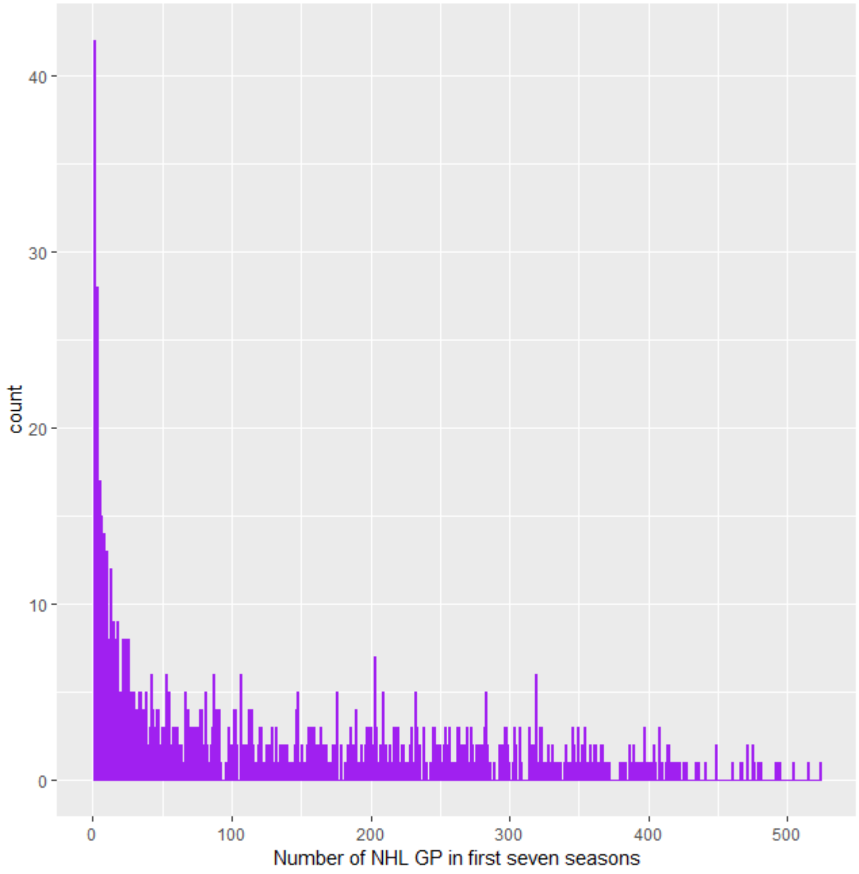

Figure 3: The count distribution of the number of NHL games played in the first seven seasons of their careers with zeros removed. This is done using the "ggplot2" package in R with binwidth equal to 1.

1.4 Methods

Software

The entire statistical analysis involved in this project is done in RStudio, a free integrated development environment for R (R Core Team, 2019), a programming language for statistical computing and analysis (RStudio Team, 2020). R is a user friendly, free to use open-source programming language and is used all over the world by statisticians, data scientists, and large enterprises such as Walmart; it contains many tools and packages that enable us to do any type of statistical analysis or computing along with machine learning techniques (RStudio Team, 2020). R is also very useful for data visualization, as it contains the very popular ggplot2 package that is used for exploratory analysis (Wickham H, 2016). We use the AER package for count modelling (Kleiber C, 2008), we use the dplyr package for data manipulation (Wickham H, 2021), we use the pscl package for zero-inflated and hurdle modelling, and for model diagnostics (Zeileis A, 2008), and we use the countreg package for further model diagnostics

(Kleiber C, 2014). In summary, RStudio provides you with all the tools you need for your statistical analysis at zero cost and it is very user friendly.

Statistical Methods

Count Modelling

The first type of modelling we consider is count modelling since we have a count response variable. However, we note that this is more of an exploratory analysis at a preliminary stage since we expect these models to be affected by the excess zeros we have in the response as illustrated in figure 2. This will create an effect known as overdispersion, meaning that there is greater variability in the data than expected (Dean, C.B., 1998). We will start off by fitting a Poisson and Negative Binomial model to the data. We will use hypothesis testing to test for the presence of overdispersion and other diagnostic tools such as a plot of fitted values versus residuals to evaluate these models.

A brief overview of Poisson and negative binomial regression as part of the Generalized Linear Model Framework follows. Representative references for this material include Dobson and Barnett (2008) or McCullagh and Nelder (1983); consult these for more details.

A Poisson regression model is a Generalized Linear Model (GLM) that is used to model count data. In this model, we need to estimate the parameter λ, the average number of occurrences per unit, also known as the rate parameter.

Suppose that we want λ_i to depend on the set of covariates contained in the vector x_i. Then, we can introduce the simple linear model

$$\lambda_i = x_i^{'}\beta$$

but this model clearly violates the non-negativity of the Poisson mean since it is an expected count, and the linear predictor on the right-hand side can assume any real value. In the GLM framework, we use a log link function to address this problem. The log-linear model is written as

$$\log(\lambda_i) = x_i^{'}\beta.$$

In this model, increasing x_j by one-unit results in an increase in the log of the mean by β_j and we can also write the model as

$$\lambda_i = \exp\{x_i^{'}\beta\}.$$

This is essentially a GLM with Poisson error and a log link function. For parameter estimation, we use maximum likelihood estimation (MLE) where the log-likelihood equation is given by

$$\log L(\beta) = \sum\{y_i log(\lambda_i) - \lambda_i\},$$

where we note that the log is the canonical link for the Poisson distribution and the sum is over all observations. By taking derivatives with respect to β and solving the equation by setting the derivatives to zero, the MLE satisfies the estimating equations given by

$$X'y = X'\hat{\lambda}$$

where X is the model matrix, y is the response vector, and $\hat{\lambda}$ is a vector of fitted values.

In the negative binomial model, we use a very similar framework as the Poisson model. The negative binomial model is more useful for modelling overdispersion in count data. In the negative binomial model, we start from a Poisson regression model but this time we add a multiplicative random effect representing unobserved heterogeneity. This then leads to the negative binomial model.

Zero-Inflated Modelling

Another type of model that we may fit to this data is a zero-inflated model. With the presence of overdispersion and excess zeros, it is a good idea to consider this type of model. The problem that can occur with Poisson and negative binomial models is that we may encounter more zeros than expected under either model. In zero-inflated modelling, we do not have to assume that each observation comes from the same distribution. Assuming that each observation comes from the same distribution in this case may be too restrictive since there is clearly a divide in the response variable. So, with zero-inflated modelling, we work toward identifying subpopulations within the population, and assume that more than one distribution will fit to each subpopulation. For a detailed discussion of zero-inflated modelling see Bilder and Loughlin (2014).

In our case, we consider the zero-inflated Poisson model and the zero-inflated negative binomial model. The zero-inflated Poisson considers two classes, the “always zero” and the “not always zero” classes. Note that “not always zero” implies that zeros may belong to this class. This model combines the power of a logistic regression (logit model) with our Poisson model as outlined in the previous section. The logit model predicts which class an observation belongs to. There are two types of zeros in zero-inflated models, the first being structural zeros from the “always zero” class and then there are random zeros from the “not always zero” class. The zero-inflated negative binomial model works in a very similar way, but this time combining the power of the logit model with a negative binomial model, as we discuss briefly in the previous section. As a result, it is important for us to understand how the logit model is defined in these zero-inflated models.

We assume that the random variable Y_i follows a binomial distribution, denoted by

$$Y_i \sim Bin(n_i, p_i)$$

with $n_i = 1$ for all i and probability p_i. We then assume that the $logit(p_i)$ is a linear function of the predictors such that

$$logit(p_i) = x_i'\beta.$$

This is a GLM with binomial response and a logit link function, where x_i is a vector of covariates and β is a vector of regression coefficients. After exponentiating, we get

$$\frac{p_i}{1-p_i} = exp\{x_i'\beta\}.$$

This defines a multiplicative model for the odds, which may seem more familiar. Furthermore, the model can be rewritten by solving for p_i

$$p_i = \frac{exp\{x_i'\beta\}}{1+exp\{x_i'\beta\}}.$$

Once again, MLE is used for parameter estimation. Here, the log-likelihood function is

$$log\ L(\beta) = \sum\{y_i \log(p_i) + (n_i - y_i)\log(1-p_i)\}.$$

After maximizing the log-likelihood function, we obtain

$$\hat{\beta} = (X'WX)^{-1}X'Wz,$$

where the dependent variable z has elements

$$z_i = \hat{n}_i + \frac{y_i - \hat{u}_i}{\hat{u}_i(n_i - \hat{u}_i)}n_i$$

and W is a diagonal matrix of weights with entries

$$w_{ii} = \frac{\hat{u}_i(n_i - \hat{u}_i)}{n_i}.$$

Hurdle Modelling (Two-part Modelling)

Two-part models, also known as hurdle models, are useful in the case where we are dealing with a count response variable where there is an excess number of zeros and overdispersion is clearly present in the data. In other words, the data consists of zeros and non-zeros. A logistic regression (logit model) is used to distinguish counts of zero from larger counts, and a truncated Poisson model or truncated Negative Binomial model is used to model the counts that are greater than zero. We refer to the two distributions as truncated since they are zero-truncated, meaning they cannot produce any zeros.

In other literatures, a truncated at zero Poisson and a truncated at zero Negative Binomial are also known as zero-altered Poisson (ZAP) and zero-altered Negative Binomial

(ZANB). We call these models two-part since the modelling is done in two stages, and we are using two different models, beginning with the logistic regression. We also call these models Hurdle models since a threshold must be crossed before the greater than zero counts can be modelled using the truncated distribution of choice. For a detailed discussion and comparison of these techniques to zero-inflated methods for count data see Bilder and Loughlin (2014) or Zuur et al (2009).

Although this modelling process is very similar to the zero-inflated modelling approach, the main difference is to understand that zeros are omitted in the second process where we use a truncated distribution. We do not use a truncated distribution in the zero-inflated models.

Testing for Overdispersion

The key assumption of a Poisson distribution is that the variance equals the mean. Analytically, the ratio of the variance to the mean will exceed one in the presence of overdispersion. We have for the Poisson distribution that

$$var(Y) = E(Y) = \mu$$

Now assume we are dealing with count data and the assumption is now that the variance is proportional to the mean, so

$$var(Y) = \varphi E(Y) = \varphi\mu$$

If $\varphi = 1$ we obtain the mean-variance relationship for the Poisson distribution. We will have overdispersion relative to the Poisson distribution when $\varphi > 1$. This is a very simple case for identifying overdispersion, as we are only relying on one assumption, but gives a good approximation for when overdispersion is indeed present.

Rodriguez (2013) noted that applying IRLS (Iteratively Reweighted Least Squares) involves working with the Poisson weights $w^* = \frac{\mu}{\varphi}$. After computing the weighted estimator, φ will cancel out and the weighted estimator reduces to a Poisson MLE. Thus, Poisson estimates are QMLE (Quasi-Maximum Likelihood Estimators) when the variance is proportional to the mean, which is our assumption when dealing with count data.

Now suppose we have an estimate for our regression parameter $\hat{\beta}$. Then

$$var(\hat{\beta}) = \varphi(X'WX)^{-1}$$

where W are the weights $(\mu_1, \dots, \mu_n)$ and we get the Poisson variance when $\varphi = 1$. As a result, the standard way of detecting overdispersion is to compare the ratio of Pearson's test statistic to the corresponding degrees of freedom. Pearson's test statistic is defined as

$$\chi_p^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{var(y_i)} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\varphi \mu_i}$$

and if there is no overdispersion present, then

$$E\left(\chi_p^2\right) \approx n - p$$

Therefore,

$$\hat{\varphi} = \frac{\chi_p^2}{n - p}$$

Score Test Statistic for Overdispersion

`Score tests for overdispersion were presented by Dean (1992) based on mean-variance relationships. A similar approach is taken by (Yang et al., 2009). It begins by introducing the general Poisson (GP) regression modelling framework, an extension of a Poisson regression model. The GP distribution can be extended to GP-1 and GP-2 regression models as a result of two parametrizations.

When overdispersion is present, the GP distribution can be used and is described by the probability mass function

$$f(y_i; \theta_i, \alpha) = \frac{\theta_i (\theta_i + \alpha y_i)^{y_i - 1} e^{-\theta_i - \alpha y_i}}{y_i!}, y_i = 0,1,2, \ldots$$

where $\theta_i > 0$ and $max(-1, -\theta_i/4) < \alpha < 1$. The GP-1 regression model can be established using the log link

$$log \frac{\theta_i}{1 - \alpha} = \sum_{r=1}^{p} x_{ir} \beta_r$$

which results in the model given by

$$P(Y = y_i | x_i, \beta, \alpha) = ((1 - \alpha)\mu_i + \alpha y_i)^{y_i - 1} \frac{(1 - \alpha)\mu_i}{y_i!} \exp(-(1 - \alpha)\mu_i - \alpha y_i),$$
$$y_i = 0,1,2, \ldots$$

$$E(Y_i) = \mu_i = \exp(X'\beta), \ Var(Y_i) = \phi \mu_i.$$

where $\phi = \frac{1}{(1-\alpha)^2}$ is the dispersion factor.

For the Poisson regression versus the GP-1 model, we are interested in testing

$$H_0: \alpha = 0 \text{ vs.}$$

$$H_1 \colon \alpha > 0.$$

The score test statistic is given by (Yang et al., 2007)

$$S(\hat{\beta}) = \frac{1}{\sqrt{2n}} \sum_{i=1}^{n} \left(\frac{(y_i - \hat{\theta}_i)^2 - y_i}{\hat{\theta}_i} \right)$$

where $\hat{\theta}_i$ is the predicted value from the Poisson model and

$$S(\hat{\beta}) \rightarrow N(0,1) \; as \; n \rightarrow \infty.$$

Similarly, the GP-2 regression model is given by

$$P(Y = y_i | x_i, \beta, \varphi) = (\tfrac{\mu_i}{1+\varphi\mu_i})^{y_i} \tfrac{(1+\varphi y_i)^{y_i - 1}}{y_i!} \exp\left(-\tfrac{\mu_i(1+\varphi y_i)}{1+\varphi\mu_i}\right), \qquad y_i = 0,1,2,\ldots$$

$$E(Y_i) = \mu_i = \exp(X'\beta)\,, Var(Y_i) = \mu_i(1+\varphi\mu_i)^2$$

where φ is the dispersion parameter and if $\varphi = 0$ we have the Poisson regression model.

For the Poisson regression versus the GP-2 model, we are interested in testing

$$H_0 \colon \varphi = 0 \text{ vs.}$$

$$H_1 \colon \varphi > 0.$$

Although two score test statistics are presented by (Yang et al., 2009), we will only present one for simplicity. The score test statistic is given by

$$S(\hat{\beta}) = \left(\sqrt{2 \sum_{i=1}^{n} {\hat{\mu}_i}^2} \right)^{-1} \sum_{i=1}^{n} ((y_i - \hat{\mu}_i)^2 - y_i)$$

where $\hat{\mu}_i$ is the predicted value from the Poisson model and

$$S(\hat{\beta}) \rightarrow N(0,1) \; as \; n \rightarrow \infty.$$

1.5 Results

Poisson Model

A Poisson model was fit first. Although we expect this model's fit to be poor because of the excess zeros in our response, we analyze the effects of this zero-inflation on the results. We summarize some results from the modelling in the tables below.

Table 2: Information on Deviance Residuals

Min.	1Q	Median	Mean	3Q	Max.
-28.37	-8.62	-5.41	0.05	0.15	39.05

Table 2 shows that the median of the deviance residuals is equal to -5.412 and the mean is equal to 0.05. This shows that there is extreme skewness, which is the result of having excess zeros in the data.

Table 3: Summary of the Poisson GLM

Variable	Estimate	Std. Error	*Z* value	Pr(>|*z*|)
Intercept	6.7350	0.1252	53.80	≈ 0
DraftAge	0.1796	0.0017	107.83	≈ 0
Country_groupEURO	-0.1623	0.0074	-21.93	≈ 0
Country_groupUSA	-0.0255	0.0081	-3.15	≈ 0
Height	-0.1285	0.0019	-64.55	≈ 0
Weight	-0.0196	0.0003	74.66	≈ 0
PositionD	-0.1295	0.0086	15.06	≈ 0
PositionL	0.0655	0.0084	7.78	≈ 0
PositionR	-0.0134	0.0086	-1.56	0.11901
Overall	-0.0116	0.0001	-189.41	≈ 0
CSS_rank	-0.0001	0.0001	-1.50	0.13378
rs_GP	0.0051	0.0002	20.71	≈ 0
rs_G	0.0027	0.0004	7.62	≈ 0
rs_A	0.0129	0.0003	49.35	≈ 0
rs_PIM	-0.0018	0.0001	-28.44	≈ 0
rs_PlusMinus	-0.0037	0.0003	-14.52	≈ 0

Table 3 shows a summary of the Poisson Regression model. It is important to note that since this is a Poisson GLM, we need to exponentiate the regression coefficient estimates to interpret the effect the covariates have on the number of games played, our response variable. For example, for each one unit increase in rs_G, we get an increase of $exp(0.002684) = 1.003$ in number if games played. We may observe that the estimates seem to be quite low for all of our predictors, which may be a result of the excess zeros and overdispersion present in the data. It is also important to note that one of the main assumptions for the Poisson model is that the variance is equal to the mean, and we can see how the excess zeros may violate this criterion. Consequently, such conclusions aren't valid because assumptions are not met.

Table 4: Information on Deviance from Regression Output

Null Deviance: 345133	2223 d.f.	
Residual Deviance: 232688	2208 d.f.	P-value: 0

Table 4 displays deviance information for the Poisson GLM. The residual deviance is the difference between the deviance of the current model and the maximum deviance of the best model where there is little to no error. Therefore, a large residual deviance suggests that the goodness of fit will be significant, indicating that the model does not fit the data well. The p-value of zero suggests that the Poisson model is a poor fit.

Table 5: Results on Testing for Overdispersion

Z	**p-value**	**alpha**
12.26	≈ 0	163.07

We use the AER package in R to produce the results shown in Table 5. The results show that the testing for overdispersion is highly statistically significant and indicated by the sample estimate of alpha equal to 163.07, as overdispersion corresponds to alpha being greater than zero.

Negative Binomial Model

We now present results of the negative binomial modelling, as this is the next model to consider. We are more confident in the model fit since the negative binomial model has the addition of a dispersion parameter that accounts for overdispersion in the data, but this does not necessarily mean that it will be a good fit, due to the extreme zero-inflation.

At a first glance, Table 6 shows us that many of the predictors are not statistically significant, judging by their respective p-values. However, it is necessary to fit separate models to determine the statistical significance of the categorical variables, since the overdispersion parameter is held constant if we do not. In the second model, we omit the Country_group predictor and compare to the first model.

Table 6: Summary of the Negative Binomial GLM

Variable	Estimate	Std. Error	*Z* value	Pr(>|*z*|)
Intercept	10.5173	2.8473	3.69	≈ 0
DraftAge	0.2331	0.0460	5.07	≈ 0
Country_groupEURO	-0.0516	0.1641	-0.31	0.75
Country_groupUSA	-0.1044	0.1862	-0.56	0.58
Height	-0.2482	0.0435	-5.71	≈ 0
Weight	0.0359	0.0057	6.27	≈ 0
PositionD	0.0209	0.1942	0.11	0.91
PositionL	0.0529	0.1958	0.27	0.79
PositionR	0.2679	0.1994	1.34	0.18
Overall	-0.0105	0.0010	-10.09	≈ 0
CSS_rank	0.0006	0.0011	0.53	0.59
rs_GP	0.0118	0.0050	2.35	0.02
rs_G	0.0010	0.0095	0.11	0.92
rs_A	0.0187	0.0072	2.59	≈ 0
rs_PIM	-0.0019	0.0014	-1.32	0.19
rs_PlusMinus	-0.0058	0.0063	-0.92	0.36

Table 7 shows the results of a likelihood ratio tests comparing the two models, the first with all predictors included and the second with the Country_group predictor excluded. We test the hypothesis that model 1 is a better fit than model 2. The test is not statistically significant as we get a p-value of 0.8529, suggesting that the Country_group variable is not a statistically significant predictor. We perform the same test, but this time omitting the Position predictor.

Table 7: Likelihood Ratio Test of Negative Binomial Models 1 and 2

Model	theta	Resid. d.f.	2x log-lik.	Test	d.f.	LR stat.	Pr(Chi)
2	0.11178	2210	-14412.4				
1	0.11181	2208	-14412.1	1 vs 2	2	0.32	0.85

Similar to the first test, Table 8 shows that the second test suggests that the Position predictor is not statistically significant either.

Table 8: Likelihood Ratio Test of Negative Binomial Models 1 and 3

Model	theta	Resid. d.f.	2x log-lik.	Test	d.f.	LR stat.	Pr(Chi)
3	0.11167	2211	-14413.9				
1	0.11181	2208	-14412.1	1 vs 2	3	1.82	0.61

The negative binomial model surely considers the issue of overdispersion; however, we still question the suitability of the model for our data and that is because of the zero-heaviness in our response variable.

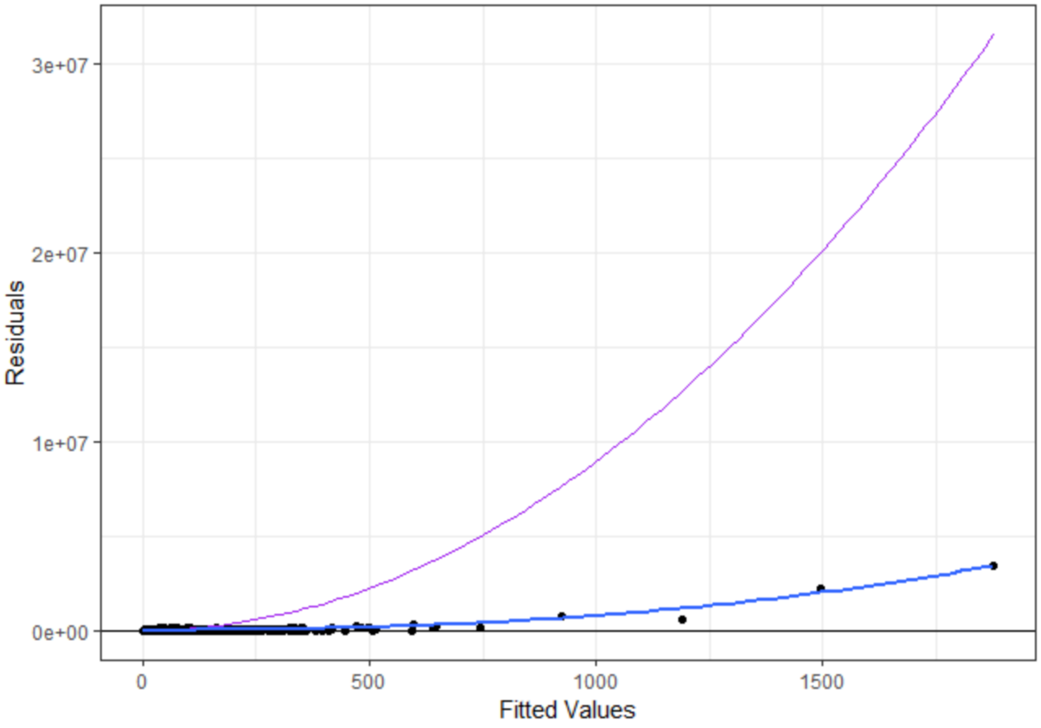

Figure 4: A plot of fitted vs residual values. The purple curve represents the negative binomial model, and the blue curve represents our ideal model fit.

The plots of the residuals versus the fitted values in Figures 4 and 5 gives us further intuition on why the negative binomial model is not suitable for our data because of the zero-heaviness in our response. There is clearly some type of non-linear relationship, however the negative binomial model seems to overestimate this relationship.

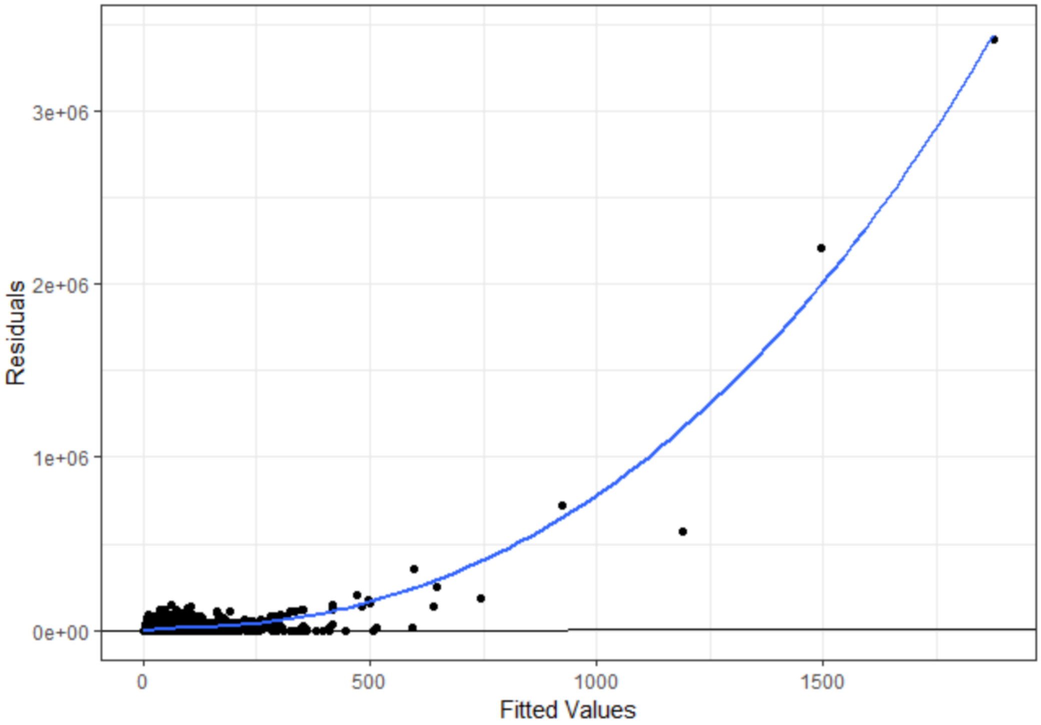

Figure 5: Similar to Figure 4, a plot of residual vs fitted values without the negative binomial curve.

Zero-Inflated Poisson (ZIP) Model

As an attempt to improve upon our ordinary count models, we now fit zero-inflated models beginning with the zero-inflated Poisson model. The summary of the model is presented in Tables 9 and 10, with the former presenting the estimated count model component and the latter presenting the estimated logistic model component of the fitted ZIP model.

Table 9: Summary of the Count Model (Poisson Model)

Variable	Estimate	Std. Error	*Z* value	Pr(>|*z*|)
Intercept	7.1160	0.1273	55.89	≈ 0
DraftAge	0.0542	0.0018	30.89	≈ 0
Country_groupEURO	0.0439	0.0075	5.83	≈ 0
Country_groupUSA	0.0055	0.0081	0.68	0.50
Height	-0.0703	0.0021	-33.95	≈ 0
Weight	0.0102	0.0003	36.84	≈ 0
PositionD	0.0454	0.0087	5.24	≈ 0
PositionL	-0.0765	0.0084	-9.06	≈ 0
PositionR	-0.1241	0.0086	-14.41	≈ 0
Overall	-0.0052	0.0001	-87.67	≈ 0
CSS_rank	0.0002	0.0001	2.66	0.008
rs_GP	0.0005	0.0002	1.83	0.07
rs_G	0.0044	0.0004	12.41	≈ 0
rs_A	0.0070	0.0003	25.79	≈ 0
rs_PIM	-0.0007	0.0001	-11.53	≈ 0
rs_PlusMinus	0.0026	0.0003	8.92	≈ 0

Table 10: Summary of the logit model

Variable	Estimate	Std. Error	*Z* value	Pr(>|*z*|)
Intercept	0.0423	2.2481	0.02	0.99
DraftAge	-0.4246	0.0439	-9.65	≈ 0
Country_groupEURO	0.4729	0.1274	3.71	≈ 0
Country_groupUSA	0.0947	0.1423	0.67	0.51
Height	0.1977	0.0343	5.75	≈ 0
Weight	-0.0349	0.0046	-7.62	≈ 0
PositionD	-0.1294	0.1524	-0.85	0.40
PositionL	-0.3910	0.1538	-2.54	0.01
PositionR	-0.3072	0.1541	-1.99	0.04
Overall	0.0126	0.0009	14.12	≈ 0
CSS_rank	0.0007	0.0009	0.72	0.47
rs_GP	-0.0124	0.0040	-3.08	0.002
rs_G	0.0049	0.0074	0.67	0.50
rs_A	-0.0235	0.0057	-4.16	≈ 0
rs_PIM	0.0036	0.0011	3.26	0.001
rs_PlusMinus	0.0149	0.0049	3.07	0.002

The output from the zero-inflation model is similar to the output from an ordinary least square regression. As mentioned previously, the model considers two classes and the logit model predicts which class an observation belongs to. This is why we have two tables to summarize the results of our model. We can see that there are many predictors in the logit model that are not statistically significant, whereas in the Poisson model most of our predictors are statistically significant.

Our zero-inflated Poisson model fits better than the intercept-only model as indicated by the chi-squared test in Table 11. The degrees of freedom for the test is 12 since we have 12 predictors in the full model. We now test to see if our zero-inflated Poisson model is an improvement over our standard Poisson model.

Table 11: Chi-squared Test on the Difference of Log-Likelihood to compare the ZIP Model to the Null Model

2x log-lik.	P-value	d.f.
23693.77	0	12

We use the Vuong test to compare our zero-inflated Poisson model to our Poisson mode. The results of this appear in Table 12; they show that our zero-inflated model is much better.

Table 12: Summary of the Vuong Test to compare the ZIP model to the Poisson Model

	Vuong z-statistic	H_A	p-value
Raw	21.79	model1 > model2	≈ 0
AIC-corrected	21.79	model1 > model2	≈ 0
BIC-corrected	21.77	model1 > model2	≈ 0

Zero-Inflated Negative Binomial Model

The zero-inflated negative binomial is similar to our zero-inflated Poisson model, but this time using the power of the negative binomial model rather than the Poisson model, which suffers less from overdispersion.

Table 13: Summary of the count model (Negative Binomial Model)

Variable	Estimate	Std. Error	*Z* value	Pr(>|*z*|)
Intercept	7.5465	1.8151	4.16	≈ 0
DraftAge	0.0434	0.0256	1.70	0.09
Country_groupEURO	0.1035	0.1072	0.97	0.33
Country_groupUSA	-0.0127	0.1140	-0.11	0.91
Height	-0.0755	0.0295	-2.56	0.01
Weight	0.0101	0.0040	2.50	0.01
PositionD	0.0749	0.1223	0.61	0.54
PositionL	-0.0650	0.1212	-0.54	0.59
PositionR	-0.0578	0.1251	-0.46	0.64
Overall	-0.0046	0.0007	-6.75	≈ 0
CSS_rank	0.0004	0.0007	0.53	0.59
rs_GP	0.0004	0.0038	0.09	0.92
rs_G	0.0048	0.0054	0.88	0.38
rs_A	0.0069	0.0041	1.68	0.09
rs_PIM	-0.0001	0.0008	-0.17	0.86
rs_PlusMinus	0.0041	0.0047	0.87	0.39
Log(theta)	-0.4139	0.0524	-7.89	≈ 0

Table 14: Summary of the logit model

Variable	Estimate	Std. Error	*Z* value	Pr(>|*z*|)
Intercept	0.5156	2.3927	0.22	0.83
DraftAge	-0.4625	0.0532	-8.70	≈ 0
Country_groupEURO	0.4971	0.1328	3.74	≈ 0
Country_groupUSA	0.1076	0.1481	0.73	0.47
Height	0.2034	0.0361	5.64	≈ 0
Weight	-0.0362	0.0049	-7.47	≈ 0
PositionD	-0.1053	0.1595	-0.66	0.51
PositionL	-0.4033	0.1611	-2.50	0.01
PositionR	-0.3083	0.1611	-1.91	0.06
Overall	0.0128	0.0009	13.75	≈ 0
CSS_rank	0.0007	0.0009	0.76	0.45
rs_GP	-0.0135	0.0042	-3.18	0.001
rs_G	0.0059	0.0078	0.77	0.44
rs_A	-0.0246	0.0060	-4.08	≈ 0
rs_PIM	0.0038	0.0011	3.32	0.001
rs_PlusMinus	0.0159	0.0053	3.04	0.002

Once again, our zero-inflated model considers the “always zero” and the “not always zero” classes, which is why we have a summary of the negative binomial model and the logit model. Unlike the zero-inflated Poisson model, there are many predictors in the count model component that are not statistically significant. In the negative binomial model, we have a parameter estimate for log(theta) as the model accounts for overdispersion.

Our zero-inflated Negative Binomial model is much better than the intercept-only model as we can see from the chi-squared test in Table 15. We now test to see if our zero-inflated Negative Binomial model fits better than our Negative Binomial model.

Table 15: Chi-squared Test on the Difference of Log-Likelihood to compare the ZINB Model to the Null Model

2x log-lik.	P-value	d.f.
710.532	0	12

The Vuong test summarized in Table 16 indicates that our zero-inflated Negative Binomial model fits much better than the Negative Binomial model.

Table 16: Summary of the Vuong Test to compare the ZINB model to the Negative Binomial Model

	Vuong z-statistic	H_A	p-value
Raw	15.2754	model1 > model2	< 2.22e-16
AIC-corrected	14.6842	model1 > model2	< 2.22e-16
BIC-corrected	12.9970	model1 > model2	< 2.22e-16

Hurdle Modelling

The hurdle model is the last model to fit to this data and we expect it to do the best job of handling overdispersion and excess zeros. We present results of the hurdle modelling in this section using a truncated Poisson distribution and a truncated Negative Binomial distribution. Once again, we may refer to Figure 2 to get an idea of the extreme zero-inflation we are dealing with. Furthermore, there are exactly 1260 out of 2224 players that have 0 games played in the NHL, accounting for more than 50% of the population. To get a better idea of why a model like this is needed for the excess zeros, we run a prediction using a typical Poisson count model. Using the “predict” function in R, we predict the expected mean count for each observation and then we use the “sum” function to sum the probabilities of a 0 count, to the predicted number of zeros.

Table 17: Predicted vs. Observed Zeros from Poisson Model

Predicted number of zeros	Observed number of zeros
1	1260

Once again, Table 17 gives us another reason why an ordinary count model isn't suitable for this data, which is why we moved on to zero-inflated modelling and hurdle modelling.

Hurdle Model with a Truncated Poisson

Once again, the regression coefficients in Tables 18 and 19 are interpreted similar to our previous modelling coefficients. We observe that many of the predictors in the truncated Poisson model are statistically significant, but we see that this is not true for the logit model. The logit model in this case models whether a player plays in an NHL game or not. Table 20 illustrates how many zeros are predicted in this case.

Table 18: Summary of the Truncated Poisson Model

Variable	Estimate	Std. Error	*Z* value	Pr(>|*z*|)
Intercept	7.1160	0.1273	55.89	≈ 0
DraftAge	0.0542	0.0018	30.89	≈ 0
Country_groupEURO	0.0439	0.0075	5.83	≈ 0
Country_groupUSA	0.0055	0.0081	0.68	0.50
Height	-0.0703	0.0021	-33.95	≈ 0
Weight	0.0102	0.0003	36.84	≈ 0
PositionD	0.0454	0.0087	5.24	≈ 0
PositionL	-0.0765	0.0084	-9.06	≈ 0
PositionR	-0.1241	0.0086	-14.41	≈ 0
Overall	-0.0052	0.0001	-87.67	≈ 0
CSS_rank	0.0002	0.0001	2.66	0.008
rs_GP	0.0005	0.0003	1.83	0.07
rs_G	0.0044	0.0004	12.41	≈ 0
rs_A	0.0070	0.0003	25.79	≈ 0
rs_PIM	-0.0007	0.0001	-11.53	≈ 0
rs_PlusMinus	0.0026	0.0003	8.92	≈ 0

Table 19: Summary of the Logit Model

Variable	Estimate	Std. Error	*Z* value	Pr(>|*z*|)
Intercept	-0.0396	2.2481	-0.02	0.99
DraftAge	0.4246	0.0440	9.65	≈ 0
Country_groupEURO	-0.4729	0.1274	-3.71	≈ 0
Country_groupUSA	-0.0947	0.1423	-0.67	0.51
Height	-0.1978	0.0344	-5.75	≈ 0
Weight	0.0350	0.0046	7.62	≈ 0
PositionD	0.1294	0.1524	0.85	0.39
PositionL	0.3910	0.1538	2.54	0.01
PositionR	0.3072	0.1541	1.99	0.04
Overall	-0.0126	0.0009	-14.12	≈ 0
CSS_rank	-0.0007	0.0009	-0.72	0.47
rs_GP	0.0124	0.0040	3.08	0.002
rs_G	-0.0049	0.0074	-0.67	0.50
rs_A	0.0235	0.0057	4.16	≈ 0
rs_PIM	-0.0036	0.0011	-3.26	0.001
rs_PlusMinus	-0.0149	0.0049	-3.07	0.002

Table 20: Predicted vs. Observed Zeros from Hurdle Model with Truncated Poisson

Predicted number of zeros	Observed number of zeros
1260	1260

The hurdle model predicts the exact number of zeros as seen in the observed data, which is expected by design. This will be true for any hurdle model, and it is why hurdle models are good for handling excess zeros in modelling. We can visualize the fit of the model using a rootogram from the countreg package in R (Kleiber, 2014). This appears in Figure 6. The horizontal line at zero allows us to analyze whether there is overfitting or underfitting present. At zero, the model fits perfectly by design, similar to why we get a perfect prediction in Table 20. However, we see that there is severe underfitting followed by overfitting for the positive counts. This suggests that the model suffers from overdispersion.

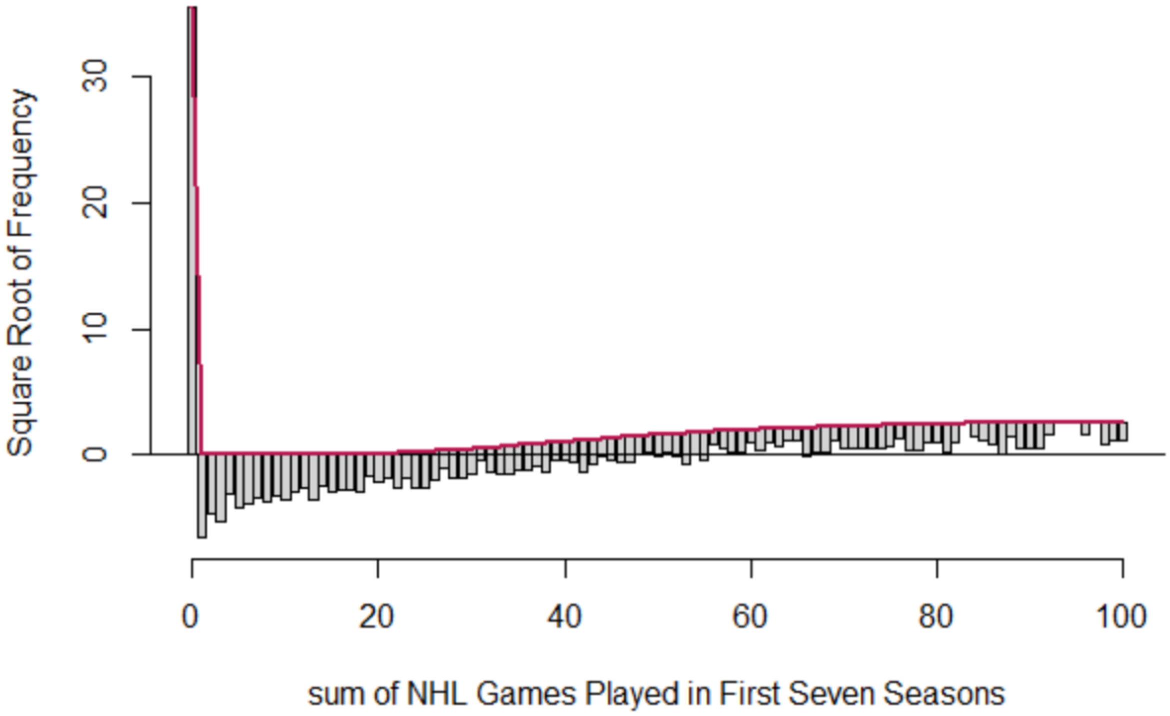

Figure 6: A Rootogram of the Hurdle Model using the Truncated Poisson Model

Hurdle Model with a Truncated Negative Binomial

Unlike our truncated Poisson model, we can see that the truncated negative binomial model gives us a parameter estimate for log(theta) in Table 21. Since we already know the excess zeros are taken care of in the hurdle model by design, we look at another rootogram to see if we addressed the issue of overdispersion.

Table 21: Summary of the Truncated Negative Binomial Model

Variable	Estimate	Std. Error	*Z* value	Pr(>|*z*|)
Intercept	7.4852	1.8142	4.13	≈ 0
DraftAge	0.0461	0.0256	1.80	0.07
Country_groupEURO	0.1041	0.1071	0.97	0.33
Country_groupUSA	-0.0135	0.1141	-0.12	0.91
Height	-0.0758	0.0294	-2.58	0.009
Weight	0.0101	0.0040	2.53	0.01
PositionD	0.0641	0.1223	0.52	0.60
PositionL	-0.0699	0.1214	-0.58	0.56
PositionR	-0.0652	0.1252	-0.52	0.60
Overall	-0.0046	0.0007	-6.66	≈ 0
CSS_rank	0.0004	0.0007	0.53	0.60
rs_GP	0.0007	0.0038	0.18	0.86
rs_G	0.0046	0.0054	0.85	0.40
rs_A	0.0071	0.0041	1.72	0.09
rs_PIM	-0.0002	0.0008	-0.20	0.84
rs_PlusMinus	0.0039	0.0046	0.84	0.40
Log(theta)	-0.4165	0.0530	-7.86	≈ 0

Table 22: Summary of the Logit Model

Variable	Estimate	Std. Error	*Z* value	Pr(>|*z*|)
Intercept	-0.0396	2.2481	-0.02	0.99
DraftAge	0.4246	0.0440	9.65	≈ 0
Country_groupEURO	-0.4729	0.1274	-3.71	≈ 0
Country_groupUSA	-0.0947	0.1423	-0.67	0.51
Height	-0.1978	0.0344	-5.75	≈ 0
Weight	0.0350	0.0046	7.62	≈ 0
PositionD	0.1294	0.1524	0.85	0.40
PositionL	0.3910	0.1538	2.54	0.01
PositionR	0.3072	0.1541	1.99	0.04
Overall	-0.0126	0.0009	-14.12	≈ 0
CSS_rank	-0.0007	0.0009	-0.72	0.47
rs_GP	0.0124	0.0040	3.08	0.002
rs_G	-0.0049	0.0074	-0.67	0.50
rs_A	0.0235	0.0057	4.16	≈ 0
rs_PIM	-0.0036	0.0011	-3.26	0.001
rs_PlusMinus	-0.0149	0.0049	-3.07	0.002

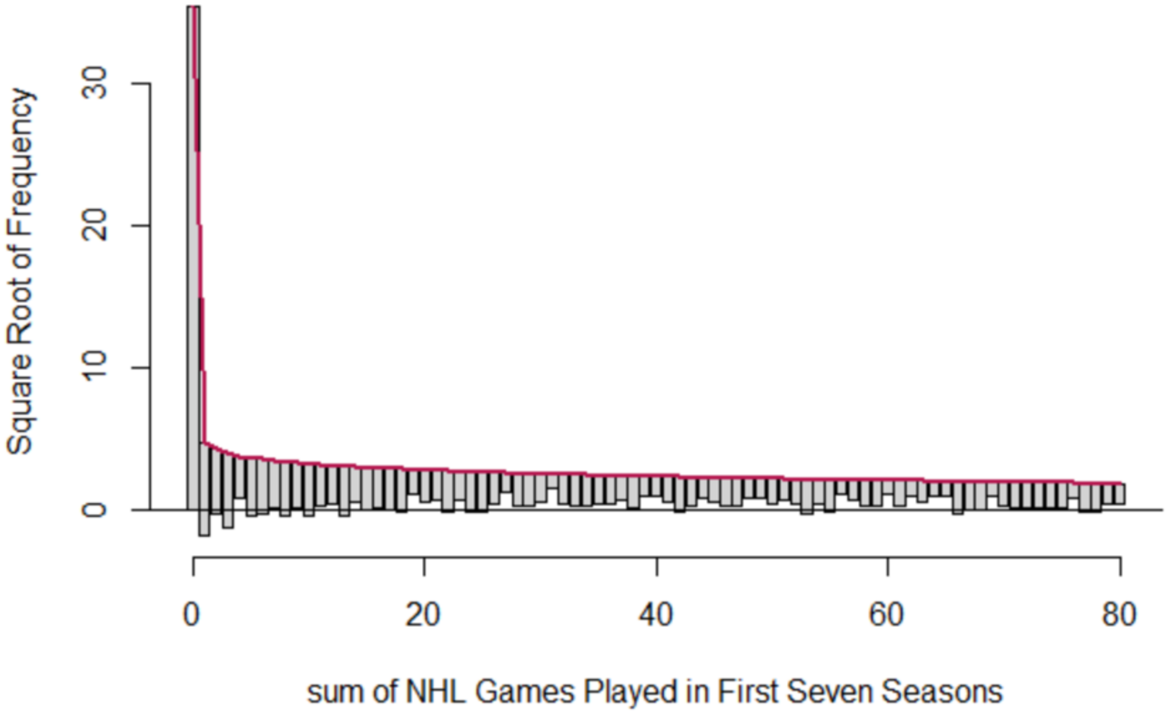

Figure 7: A Rootogram of the Hurdle Model using the Truncated Negative Binomial Model

Although there seems to be slight underfitting just after zero and slight overfitting after that, it seems that this model fits much better and addresses the issue of overdispersion. We also use the traditional AIC metric to compare these models.

Table 23: AIC Model Comparison

	Truncated Poisson Hurdle Model	Truncated Negative Binomial Hurdle Model
AIC	111835.4	13651.42

1.6 Discussion and Conclusions

There are some limitations with our general approach to this project, and how it affects the goal we want to achieve. One such limitation is that we do not include playoff statistics. We decided to not include these statistics since not every team makes the playoffs each year, and there are many great players who do not get to play in the playoffs every year. For this project, we are more focussed on identifying players that have a promising future in the NHL, so we are less worried about measuring team success. However, it may be argued that some sort of playoff statistic could be used as an explanatory variable.

Another limitation is that we do not include any statistics that come from other professional leagues, more notably the Kontinental Hockey League (KHL). But once again, including these statistics would decrease our population quite drastically, similar to playoffs statistics as it effects the completeness of our data. The argument here is that there are many great players that start off their careers in the KHL (mostly Russian players) before signing an NHL contract.

Ordinary count models were fitted to the data since this is the common approach with count data. However, we have shown why this type of model is not suitable in this case because of overdispersion and the extreme zero-inflation. Even when fitting a negative binomial model that is supposed to address the issue of overdispersion, we still see that the zero-inflation causes this model to be an extremely poor fit. Therefore, we cannot make any meaningful conclusions from these count models, but we have demonstrated why common approaches may not work in the case where there are many zeros in the count data.

The second approach was fitting zero-inflated models to our data using the Poisson distribution and the negative binomial distribution. We see much improvement in these models over the count models by using a Vuong test as seen in Table 12 and Table 16. Although the zero-inflated models are much better, we may still improve upon them by using a hurdle model. The hurdle model proves to be the best fit as it addresses the issue of overdispersion and zero-inflation. The main difference between the hurdle models and the zero-inflated models are that the zeros are omitted in the second process of the hurdle model. As a result, it seems that this is much better for handling zero-inflated data. More specifically, a hurdle model with a zero-truncated negative binomial distribution proved to be the best fit, as seen in Figure 7 using a rootogram and using the AIC metric as seen in Table 23. As a result, we conclude that from Table 22 that players drafted as left or right wingers, and at a higher age are more likely to have success in playing in at least one NHL game, and that European players are the least likely to have success in playing in at least one NHL game. From Table 21 we may conclude that players drafted at a higher age, European players, defenceman, and assists contribute most toward success in playing multiple NHL games. On the other hand, American born players and left or right wingers contribute negatively toward having success in playing in multiple NHL games.

Chapter 2: Internship Experience at Statistics Canada

2.1 About the Organization

Statistics Canada is Canada's national statistical agency, headquartered in Ottawa, Ontario. The agency provides the Canadian public with information on Canada's economy, society, and environment that are essential for decision making. Statistics Canada is lead by the Chief Statistician of Canada, Anil Arora. Statistics Canada collects, compiles, analyzes, and publishes information relating to the commercial, industrial, financial, social, economic, general activities and condition of the people of Canada. Further details about the agency can be found on their website (Government of Canada, 2021).

2.2 My Experience

Overview

During my time at Statistics Canada, I worked on the special business projects team and the agriculture team as a student junior statistician for a total of 11 months. Most of this time was spent working for the special business projects team. Both of these teams are part of the Economic Statistics and Methods Division (ESMD) at Statistics Canada. Most of my time was spent working at the headquarters in Ottawa Ontario, before having to work remotely from home due to the Covid-19 pandemic. During my work experience, I worked on several projects revolving around the Ontario First Nations Point of Sale Exemption Survey (OFNPSES) and the Survey of Commercial and Institutional Energy Use (SCIEU).

Projects and My Role

The Ontario First Nations Point of Sale Exemption Survey (OFNPSES)

The objective of this survey is to collect information on the exemptions offered to First Nations people by enterprises in Ontario. This information is used by the Ontario Ministry of Finance and Finance Canada to determine the allocation of Ontario HST revenue between the provincial and federal governments. My role here was to fit a suitable model to historical data that would help identify enterprises in the future that will report a zero-rebate value, and possibly estimate rebate values for those not being surveyed.

The issue here was that many respondents from previous iterations of the survey were reporting a value of zero for their total rebate to first nations people. This is problematic since these respondents do not contribute to the final estimates or the overall goal of this survey, making it a waste of time and money. As a result, it is necessary to define a "take-none" portion of the population and build a model on this portion of the respondent data, so that these rebates can be imputed for future iterations of the survey. Since many of these respondents report a zero-rebate value, our data for this take-none portion becomes zero-inflated or zero-heavy. Therefore, it is necessary to consider specific models for this type of data, including zero-inflated models and hurdle models. Hurdle models are often used in the context of count data, but the same idea applies here with a continuous response.

I carried out exploratory data analyses and fitted some preliminary models to gain more of an understanding of the structures in these data before moving on to final models. The most suitable model identified was a gamma-hurdle model, since it addressed the issues of overdispersion and zero-inflation, and that we had a continuous response variable. Although a model was identified for best representation of this take-none portion, another problem was that the take-none portion can vary quite significantly from year to year. This could be a problem since there is potential for a zero-contributor to become a large contributor in the following year. Therefore, there is no true definitive take-none portion from year to year. This is something that could be worked on/resolved in the future. One idea is to assign probabilities to businesses/ enterprises that determines the likelihood of offering rebates to first nations people. After establishing a threshold for lower probabilities, a take-none could be defined.

Survey of Commercial and Institutional Energy Use (SCIEU)

The Survey of Commercial and Institutional Energy Use (SCIEU) was last conducted in 2014 as a two-stage survey. Establishment was the sampling unit in the first stage and the building was the sampling unit in the second stage. The business register (BR) was used as the frame for the first stage. The sampling strategy has been redesigned for SCIEU 2019, where it is just a single-phase survey with building unit as the sampling unit. SCIEU 2019 will publish energy use, floor area and energy intensity. This is an ongoing survey that is not in collection, so changes have been made over time which had an affect on the work that I have done. The frame for the survey is the Statistical Building Register (SBgR) and newer versions have been released from time to time. I have worked mostly with SBgR Delta 2.2 and Delta 2.3. Our target population for this survey is those buildings that have more than 50% of their floor space occupied by commercial or institutional activity types. There are 24 activity types, which are all either a commercial or institutional activity type. The following is a list of all activity types for SCIEU 2019:

1. Office space (non-medical)
2. Office space (medical)
3. Bank branch
4. Courthouse
5. Police station

6. Fire station

7. Assisted daily care facilities and residential care facilities

8. Hotel, motel, or lodge

9. Preschool or daycare

10. Primary and secondary school

11. Restaurant

12. Food and beverage store

13. Retail store (non-food)

14. Shopping Centre (enclosed mall)

15. Recreation centre

16. Ice rink

17. Performing arts

18. Cinema

19. Place of worship

20. Museum and gallery

21. Library and archives

22. Warehouse

23. Vehicle repair, storage, or dealership

24. Other

The Statistical Building Register (SBgR)

The SBgR contains a list of all buildings and their building units across Canada. It is still a work in progress; it currently consists of the data frame for SCIEU 2019. It is linked to the Business Register and the Address Register at the building unit level and the activity type is coded to the building unit level. For SCIEU 2019, we are sampling and will be publishing at the building unit level. Aside from the information we get from the SBgR, the count of building units is the only auxiliary data available. For stratifying, we need to define a dominant activity type at the building unit level since the building is our sampling unit, which leads to the first task that I completed for SCIEU 2019.

Identifying In Scope Buildings for SCIEU 2019

I wrote a program in SAS to identify the "dominate" building usage of all buildings across Canada, using the Statistical Building Register (SBgR). Although we did consider using some weighted variables to derive the dominant usage (floor area or number of employees), this was solely based on the "building unit count" variable created to calculate proportions of building usage, where the highest proportion was the "dominant usage" for each building. Using this information, we could decide whether a building was in scope for this survey. This was preliminary work that was done for establishing a dominant usage or activity type. The definition of a dominant usage constantly evolved over time and became more complicated.

I also spent some time investigating unknown building usage on the SBgR using auxiliary data, coming from the Business Register (BR). For example, we could look at a variable like Business status code (on the BR) which can tell us if the business is alive or not. The goal of this was to identify and common characteristics among these building units. In the end, we determined that most unknown activity types did not have a link to the BR. For those that did have a link, there were different explanations as to why the activity type was not coded (i.e. no NAICS assigned or it was a dead business).

Frequencies of Building Groups

Using SAS, we created frequencies of groups within the SBgR to investigate how we could stratify when doing our sampling. The SBgR was divided into single-unit and multiple-unit buildings and created frequencies for several groups such as:

- Single usage buildings (buildings with one unit or buildings with multiple units but all have the same usage). Either in scope, out of scope, or unknown.
- Multiple usage buildings (buildings with multiple units, and not all usages are the same). Some have a clear dominant usage, but others are "well mixed".

The goal of this task was to provide us with an idea of how we could stratify, based on how well-populated these groups are.

Identifying Over Coverage on the SBgR

As mentioned previously, we are using a frame for this survey (SBgR) that is still a work in progress with new versions being released every few months or so. Because of this, we have buildings that are missing and buildings that are duplicates. This is a problem known as under coverage and over coverage. As you all know, both are important as under coverage can lead to those units not accounted for as part of the population, which may lead to biased results or a misrepresented sample. Over coverage can lead to one address receiving the same survey more than once which is a waste of time and money. It also adds to response burden and increases sampling variance.

Using SAS, I created address keys for identifying duplicates. This worked by identifying all address variables on the SBgR and considering all possible combinations that made sense for identifying duplicates. I also Incorporated block face into the address keys by using a concordance table. After comparing results of potential duplicates produced by these address keys, we decided that results produced by two of these address keys were most informative. One of these address keys concatenates civic number, civic number suffix, and block face ID. This was the key used previously in an analysis on Delta 2.1 and was identified as the best address key for identifying duplicates. The other address key was one that I found to be informative and seemed like it produced reliable results. It concatenates civic number, civic number suffix, street name, postal code, and municipality name. It is important to note that postal code and municipality are at the building unit level, not at the building level. Building units within a building do not necessarily have the same postal code. Sometimes this represents an error on the SBgR and sometimes it is correct.

SCIEU Modelling in R (Using RStudio)

We modeled energy consumption on the building unit level using respondent data from SCIEU 2014. The goal of this is to predict energy consumption on our new SCIEU frame, where we can use this variable for allocating the sample. We considered using province, climate region, NAICS group, Number of employees, and building unit count for covariates (aka explanatory variable's or independent variables). The final variables used in the model were climate region, NAICS group, and Number of employees. Unweighted and weighted modelling were carried out, using the final sample weight (FWEIGHT) for the weighted model. We used Stepwise Regression, since it takes all linear models into account and identifies the best one based on the Akaike information criterion (AIC). We considered time series modelling, but there was no time component or variable. Therefore, this type of model would not be suitable.

2.3 Reflection

My overall experience at Statistics Canada was extremely valuable toward my education of the field of statistics and will be carried forward in my future endeavours. It allowed me to better understand how my knowledge of statistics can be applied in the field and what an office work environment is like. It also allowed me to improve on how I interact with people as a professional, building upon my communication skills. On a typical work week at Statistics Canada, we would meet with the team for approximately one hour to discuss our progress in our projects, and any updates that have an effect on our work. This would also be a good time to ask for help or to make sure everyone was on the right track in their work progress. To me, those meetings were viewed as a work deadline to assure I was meeting expectations. Such meetings were great for improving on my presentation and collaboration skills. Most of the time, I would be working at my desk on my projects. This involved using software such as SAS, R, and Excel to do my analysis. I would also work on reporting using Microsoft Word. I was located on the same floor as my supervisor, so I would often communicate with them to make sure I was on the right track with my work progress. Most weeks, there would often be a seminar or learning session that I would attend. This was a good opportunity to learn about other projects going on at Statistics Canada that I wasn't directly involved in.

The MSc. statistics program at Western prepared me well for this internship experience. I was able to bring in my skills of statistical analysis that I acquired through the program to make a meaningful impact at Statistics Canada. Also, it was not as big of an adjustment for me to have a supervisor at Statistics Canada, as I have learnt about how this type of professional relationship works at school. This prepared me in a way that I understood what goes into meeting deadlines, and how to interact in a professional manner while staying on track with my work progress.

Acknowledgements

Throughout my time spent on this project, I have received a great deal of support and assistance. I would like to thank my supervisor, Dr. Douglas Woolford, whose expertise was very helpful in the completion of this project. Your expertise was crucial to the methodology and has allowed me to expand my toolset in how to approach a data analysis. Your feedback pushed me to produce a higher-level analysis and to think more in detail, and this will certainly help me in my future endeavours. I would also like to acknowledge my colleagues from my internship at Statistics Canada, who provided me with an invaluable work experience. I was able to learn how to navigate in an office environment, while having the ability to apply my knowledge and learn from others.

References

Bilder, C. R., & Loughin, T. M. (2014). *Analysis of Categorical Data with R*. CRC Press.

Dean, C.B. and Lundy, E.R (2016), Overdispersion, Encyclopedia of Clinical Trials, Wiley, N. Balakrishnan, P. Brandimarte, B. Everitt, G. Molenberghs, W. Piegorsch and F. Ruggeri (eds.), Wiley StatsRef: Statistics Reference Online. 19, DOI: 10.1002/9781118445112.stat06788.pub2

Dean, C.B. (1998), Overdispersion, Encyclopedia of Biostatistics, Wiley, P. Armitage, and T. Colton (eds.), 467-472.

Dean, C.B. (1992), Testing for overdispersion in Poisson and binomial regression models, J. Amer. Statist. Assoc. 87, 451-457.

Dean, C. and Lawless, J.F. (1989), Tests for detecting overdispersion in Poisson regression models, J. Amer. Statist. Assoc. 84, 467-472.

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models* (3rd ed.). Boca Raton: CRC Press

Government of Canada, S. (2021, April 15). Statistics Canada: Canada's national statistical agency. Retrieved April 15, 2021, from https://www.statcan.gc.ca/eng/start

Kirk, R. (2017, October 03). NHL draft: The 50 Best undrafted players in NHL History. Retrieved March 08, 2021, from https://bleacherreport.com/articles/1233971-nhl-draft-the-50-best-undrafted-players-in-nhl-history

Kleiber C, Zeileis A (2008). Applied Econometrics with R. Springer-Verlag, New York. ISBN 978-0-387-77316-2, https://CRAN.R-project.org/package=AER.

Kleiber, Christian & Zeileis, Achim, 2014. "Visualizing Count Data Regressions Using Rootograms," Working papers 2014/13, Faculty of Business and Economics - University of Basel.

McCullagh, P. (Peter), & Nelder, J. A. (1983). *Generalized linear models.* London: Chapman and Hall.

Odds not great for NHL draft picks. (2009). Retrieved March 08, 2021, from https://guelphstorm.com/odds-not-great-for-nhl-draft-picks#:~:text=63%25%20of%20first%20round%20picks,picks%20ever%20become%20impact%20players.

Pollard, D. (2017, November 14). How mindfulness can bring success on the ice. Retrieved March 08, 2021, from https://www.omha.net/news_article/show/856378

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rodriguez, G. (2013). Models for Count Data With Overdispersion. Retrieved June 25, 2019.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Sapunka, J. (2017, October 03). NHL draft: The 50 Biggest busts in NHL draft history. Retrieved March 08, 2021, from https://bleacherreport.com/articles/1225109-nhl-draft-the-50-biggest-busts-in-nhl-draft-history

Schulte, O., Liu, Y., & Li, C. (2018, February 27). Model Trees for Identifying Exceptional Players.

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Wickham H, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.5. https://CRAN.R-project.org/package=dplyr

Wolski, D. (2021). League Ranking. Retrieved March 08, 2021, from https://2112hockeyagency.com/league-rankings/

Yang, Zhao & Hardin, James & Addy, Cheryl. (2009). A score test for overdispersion in Poisson regression based on the generalized Poisson-2 model. Journal of Statistical Planning and Inference. 139. 1514-1521. 10.1016/j.jspi.2008.08.018.

Zeileis A, Kleiber C, Jackman S (2008). "Regression Models for Count Data in R." Journal of Statistical Software, 27(8). http://www.jstatsoft.org/v27/i08/.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.