Health in Numbers: A Logistic Model of Herd Immunity for the DTP and MMR Vaccines in Toronto Schools

Paul Hodgetts, Thomas Rosenthal, and Qi Wen

March 1, 2020

Abstract

In this work, we have performed binomial logistic regression to examine the relative likelihood that schools in the city of Toronto will reach herd immunity levels for Diphtheria, Tetanus, Pertussis ("DTP") and Measles, Mumps, and Rubella ("MMR") vaccines. Herd immunity is an important measure of disease resistance as it protects individuals who are not or cannot be immune to a given disease. Our model found that schools are under-protected, and this is inadequately explained by religious exemption data collection. We have hypothesized that the rise of the vaccine hesitance movement (more colloquially referred to as "anti-vax") has likely contributed to this shortfall. However, conclusions from this analysis have been limited by insufficient data depth and are difficult to make with appropriate veracity. Nonetheless, applying this model to like-structured datasets over time may help confirm the effect of the anti-vax movement on herd immunity in Toronto schools.

Introduction

Schools provide a prime space for the spread of disease. With people interacting in close proximity through shared spaces and items, students and teachers alike are susceptible to falling ill. To ensure a level of protection against a disease, an individual can receive a vaccination for that disease. A population can become protected when enough members of a population are vaccinated against a disease, known as herd immunity. At this threshold, fewer people will become sick as there are fewer germs to spread between the members of the population (Watson, 2018). This provides a layer of protection for those most susceptible to disease. Following global outbreaks of measles (previously declared exterminated in the U.S. in 2000) (CBC, 2019; BBC News, 2019) and the prominence of the vaccine hesitance/anti-vaccination movement, there are concerns as to the current rate of vaccinations and the possibility of maintaining long-standing herd immunity thresholds. Thus, it is increasingly important to emphasize that the rate of vaccinated students reaches these thresholds. This report, using data on immunization rates for the DTP (Diphtheria, Tetanus, Pertussis) and MMR (Measles, Mumps, and Rubella) vaccines for Toronto schools from the 2018-2019 school year, aims to assess the full herd immunity of schools. Our approach is conservative, using the higher end of recommended herd immunity rates for the most contagious diseases covered by the DTP and MMR vaccines—pertussis (a.k.a. "whooping cough") and measles respectively. This ensures that a school has obtained herd immunity for the relevant immunization. This report examines whether a school has obtained full herd immunity, which is defined as being when a school has obtained herd immunity for both DTP and MMR. As designated by Ontario's Immunization of School Pupils Act (ISPA), it is required that, unless a valid exemption is obtained, grade-school aged students be appropriately immunized (Ontario, 2018). Within this dataset, this is strictly limited to religious exemptions. Based upon this, a binomial logistic regression model was used to assess whether a school had achieved full herd immunity, using religious exemption rate as a means of prediction and controlling for unvaccinated students without religious exemption. It was found that after controlling for unvaccinated students with no religious exemption, the odds of a school achieving full herd immunity decreased by 93.8% as religious exemption rate increased. Furthermore, the model achieved 97% accuracy in predicting the probability of a school having full herd immunity using test data. However, there are concerns of overfitting that should be noted.

Research Question

The question posed is: can it be predicted whether a school will or will not have full herd immunity based on the average religious exemption rate and rate of unvaccinated students without religious exemption? In other

words, whether a school will have herd immunity for both the DTP and MMR vaccines based upon the average religious exemption rate and rate of students without religious exemption who do not have either the DTP or MMR vaccines.

Limitations

This dataset was very limited in its possible approaches. It contained no individual data records, but rather vaccination rates aggregated per school. While the dataset was built using data from students aged 7-17, age was not included in the actual dataset preventing a more in-depth analysis by age. Students without vaccines were not quantified, nor were reasons stated beyond religious exemption (see §Ethics). As such, predicting immunization outcomes has been limited, and we have shifted our focus to herd immunity to approach the dataset within its summarized view. Future application is possible with more detailed datasets, or by applying this model to a similarly structured datasets in future years to track the likelihood of herd immunity over time.

Ethics

It should be no surprise that childhood immunizations are a contested topic. This section will avoid conjectures with regards to the vaccine hesitance/anti-vaccination movement, but rather examine missing immunizations even without being opposed to their administration.

While we have focused on herd immunity, it is important to recognize our dataset did not include any exemption reasons aside from religious ones. In this regard, it is a safe assumption that some unvaccinated children are unable to be vaccinated due to medical reasons (med-exempt). Our approach aims to use herd immunity as a mechanism to prevent diseases for these children, but without knowledge of the rate at which these med-exempts occur for a given school or the entire dataset, we are potentially emphasizing gaps in vaccination rates that cannot be closed.

There are some possible reasons for children missing vaccines even without being opposed to their administration:

Firstly, childhood immunization schedules differ across provinces in Canada, across countries, and across health agencies. Some differences may include how many vaccines should be administered at a given time or how frequently a child should receive vaccines (i.e. between birth and one year of age – every 2 months vs. every 3 months). This may have resulted in slight differences when this data was collected that are not knowable in its current format.

Secondly, there are some side-effect potentials that parents may wish to avoid—some serious, such as anaphylactic shock or seizures (these would be noted med-exempt reasons), others fairly benign (crankiness, soreness, mild fever). For adults, this may resonate with arguments of bodily autonomy, where decisions made for a child do not respect the child's own desires or avoidances of administration. For children, this may also include fear of needles or pain.

Finally, there are many potential real-world scenarios that lead to missed vaccines. Children may be missing or behind on immunizations because of busy schedules, migration, poor record keeping (i.e. where records might be missed when changing doctors), or disagreements between parents on whether or not to vaccinate. We cannot presume to know all of these circumstances, and thus it is important to question the veracity of summarized numbers, as mentioned in §Limitations.

Dataset

The dataset used in the following exploratory analysis is the Immunization Coverage for Students, 2018-2019 from Toronto Open Data Portal. This dataset contains data about immunization coverage rates and exemptions for school children (7-17 years old) in Toronto for the 2017/2018 and 2018/2019 school years. Only the 2018/2019 dataset was used for data cleaning and model construction.

Data Cleaning and Structure

An analysis of the 2018/2019 dataset revealed the dataset to be initially clean. Using the skim function from the skimr package, the dataset contained no missing values, nor were any duplicates found. As can be seen in *Table 1*, the mean percentage of students covered by the DTP vaccine was 89.4% (M = 89.4, SD = 7.2), while the mean percentage of students covered by the MMR vaccine was 92.7% (M = 92.7, SD = 5.2). The mean

percentage of students with religious exemption was 2.1% for both DTP and MMR vaccines (M = 2.1, SD = 2.8).

Per our research question, new variables indicating whether a school had or had not met the required coverage for herd immunity for the DTP or MMR vaccines were developed. A school was considered to have achieved herd immunity if it had a coverage rate greater than 94% for DTP or 95% for MMR. These ranges were selected because they represent the higher end of the recommended coverage percentages for herd immunity (Majumder, et al., 2015; Oxford Vaccine Group, 2016; Plans-Rubio, 2012). The distribution of the DTP and MMR vaccines can be seen in *Figure 1.* A new binary variable was created to state whether a school had achieved full herd immunity, defined as having obtained herd immunity for both DTP and MMR. If a school had met both the 94% cut-off for DTP and 95% cut-off for MMR, it was considered to have full herd immunity.

As seen in *Table 1*, religious exemption rates for both MMR and DTP are effectively the same. As such, rather than use both individually and experience possible issues of collinearity, the average religious exemption rate was substituted.

To calculate the rate of unvaccinated students without religious exemption, the number of vaccinated students and the number of students with religious exemption were subtracted from the total enrolled population. This was taken to be the number of unvaccinated students enrolled without religious exemption. Rates for these students for both DTP and MMR were calculated by dividing the results by the enrolled population.

Table 1: Data summary

Name	immunization_load_skim
Number of rows	806
Number of columns	6
Column type frequency:	
character	1
numeric	5
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
School Name	0	1	9	43	0	806	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Enrolled Pop	0	1	323.51	282.17	14.0	144.0	236.0	393.0	1818.0	
DTP coverage %	0	1	89.43	7.21	33.3	86.2	90.5	94.4	100.0	
DTP relg exempt %	0	1	2.12	2.76	0.0	0.6	1.5	2.8	36.2	
MMR coverage %	0	1	92.72	5.16	57.9	90.6	93.9	96.1	100.0	
MMR relg exempt %	0	1	2.11	2.76	0.0	0.6	1.5	2.7	36.2	

Table 1

Figure 1 is two density plots showing the distribution of DTP and MMR coverage rate for schools in Toronto. DTP and MMR coverage rates demonstrated similar patterns (both left skewed). The red dashed line represents the conservative recommendation for herd immunity: 94% for DTP and 95% for MMR respectively. Only a small percentage of schools reached herd immunity levels.

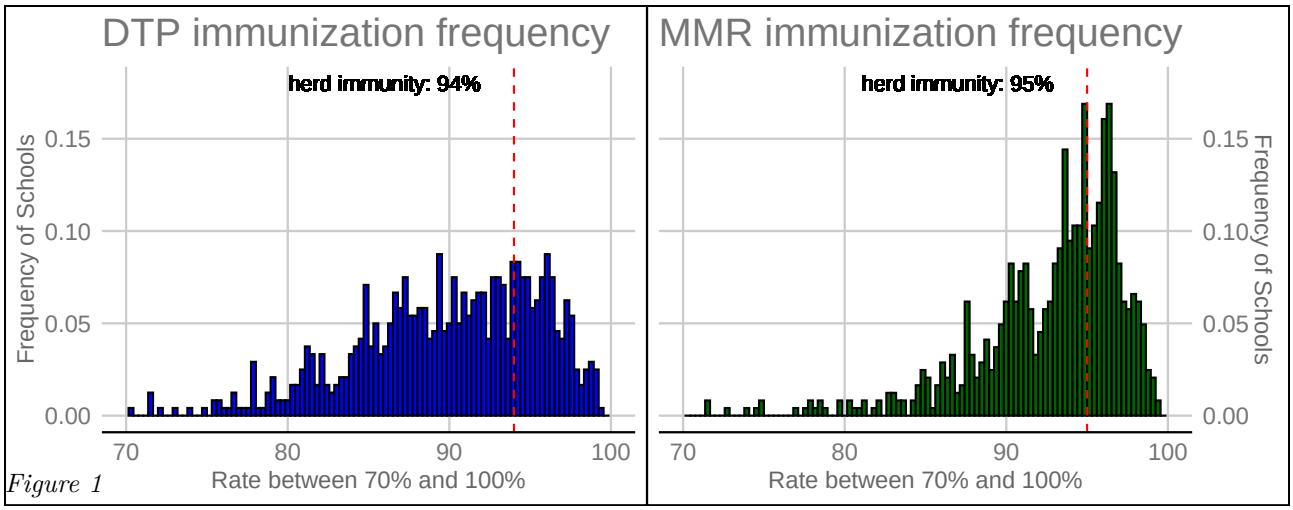


Figure 1

Herd Immunity Binomial Model

To create the model, the dataset was split 70|30 with 70% of the data being used as a training set and 30% as a test set. Splitting the dataset into training and test sets ensured the model was tested on data different than what it had been trained on. This ensured the predictions the model made were not made on data already known to the model. As such, splitting the data prevented that from occurring. A binomial logistic regression was then built using the binary full herd immunity as the dependent variable and the average religious exemption rate as the independent variable controlled by the unvaccinated DTP and MMR students without religious exemption. Logistic regression was chosen as the method of modelling due to the dependent variable, in this case full herd immunity, being binary (i.e. whether a school does or does not have full herd immunity). Furthermore, the independent variables (average religious exemption rate, DTP not-covered rate, and MMR not-covered rate) are all continuous. As such, this is a classification problem that does not lend itself to linear regression. Model results are displayed in *Table 2*. There are some concerns of overfitting with B3 = -166.1 and B4 = -153.8.

```
## # A tibble: 4 x 5
##   term                          estimate std.error statistic  p.value
##   <chr>                            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                      14.4      2.12       6.81 9.93e-12
## 2 AVG_religious_exemption_rate    -2.78     0.421     -6.61 3.88e-11
## 3 DTP_notcovered_rate           -166.      25.9       -6.40 1.52e-10
## 4 MMR_notcovered_rate           -154.      29.8       -5.16 2.47e- 7
```

Table 2

Running the model on the training dataset, it was revealed that the log-odds of a school having full herd immunity were -2.78. Thus, for each increase in the rate of religious exemption, there is an estimated decrease of -2.78 in log-odds of a school having full herd immunity. In other words, the odds of a school having full herd immunity are estimated to decrease by a factor of 0.06 (or 93.8%) for each unit increase in AVG religious exemption rate. Prior to testing, confusion matrix accuracy was 95.7%, as seen in *Table 3*.

```
## $table
##            Reference
## Prediction   0   1
##          0 422  10
##          1  14 118
##
## $overall
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##   9.574468e-01   8.800510e-01   9.373436e-01   9.725482e-01   7.730496e-01
## AccuracyPValue  McnemarPValue
##   1.850723e-34   5.402914e-01
```

Table 3

Test Data comparison

To further test the accuracy of the model, the model was run on the test dataset. Testing results showed the model to have a predictive accuracy of 97.1% (*Table 4*). This difference, 1.4%, is likely due to the difference in size of the training and test datasets. A comparison of the two sets in terms of the predictive capabilities for full herd immunity is demonstrated in *Figure 2* below.

```
## $table
##           Reference
## Prediction   0   1
##          0 191   4
##          1   3  44
##
## $overall
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##   9.710744e-01   9.083234e-01   9.413129e-01   9.882929e-01   8.016529e-01
## AccuracyPValue  McnemarPValue
##   3.306398e-15   1.000000e+00
```

Table 4

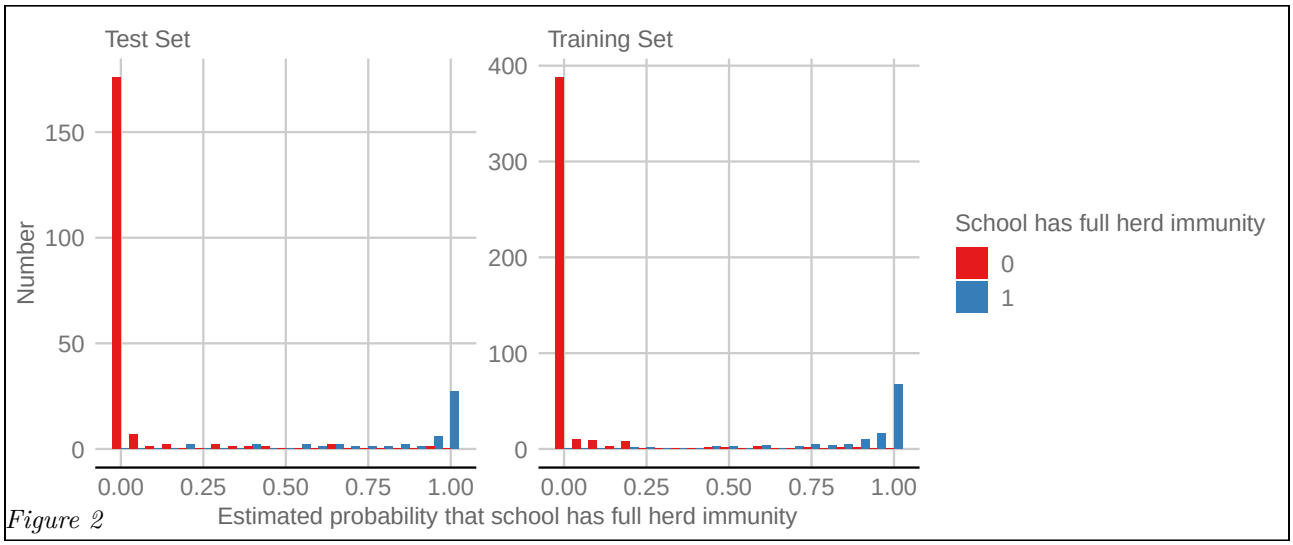


Figure 2

Testing Assumptions

The binomial logistic regression model makes several assumptions about the data. Assumption testing is an essential part of building a valid model. First, the dependent variable is supposed to be binary, or dichotomous. As the dependent variable in our model, full herd immunity was binary (0 and 1), where 0 represented the school did not reach full herd immunity and 1 represented that full herd immunity had been reached. Thus, our model satisfied the first assumption. Second, there should be a linear relationship between the logit of the dependent variable and each predictor variable. The logit function is logit(p) = log(p/(1-p)), where p is the probabilities of the outcome (dependent variable). Scatter plots showed that independent variables (DTP not-covered rate, MMR not-covered rate, AVG religious exemption rate) are all linearly associated with the full herd immunity outcome in logit scale (*Figure 3*). Third, there should not be any influential values in building the model. Influential values are extreme individual data points that can alter the quality of the logistic regression model. A Residuals vs. Leverage plot (*Figure 4* in Appendix B) was created to inspect if there were any influential points. Since no point was beyond Cook's distance (red dashed lines), there were no influential values in our data. As such, the third assumption was satisfied. Finally, there should not be multicollinearity amongst the independent variables. To test for multicollinearity, a VIF (variation inflation factor) test was performed on the model. VIF values below 5 illustrated that little multicollinearity existed. Results illustrated that DTP not-covered rate and MMR not-covered rate satisfied this threshold, while AVG religious exemption rate had a

value slightly larger than 5 (*Table 5*) but was still acceptable. Thus, the fourth assumption was satisfied by our model.

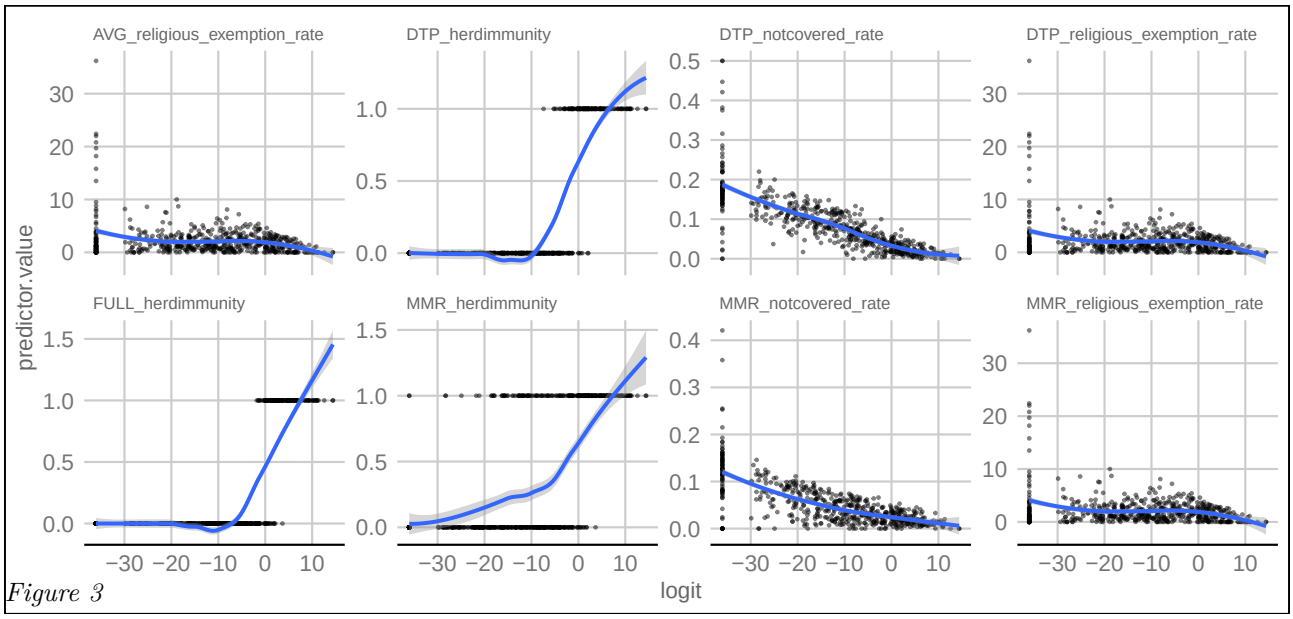


Figure 3

```
## AVG_religious_exemption_rate          DTP_notcovered_rate
##                     5.154746                     2.913906
##          MMR_notcovered_rate
##                     2.604941
```

Table 5

Conclusion

Herd immunization is important to ensure the health of a population. By having more individuals immunized, fewer germs can be spread and fewer people will fall ill. This is especially important for those most susceptible to disease and in places where disease can easily spread, such as schools. This report used the Immunization Coverage for Students, 2018/2019 dataset from Open Data Toronto to analyse whether Toronto schools had met full herd immunity for the DTP and MMR vaccines. Full herd immunity was considered to be when herd immunity was met for both DTP and MMR. There are various reasons why an individual may not receive a vaccine, whether it be medical or religious. As the datasets provided the rate of religious exemptions for vaccinations for schools, the question of whether full herd immunity could be predicted by average religious exemption rate was posed. A binomial logistic regression model was built to answer this question. As the dependent variable, full herd immunity was represented as a binary variable with average religious exemption rate controlled by rate of unvaccinated students without religious exemption acting as the independent variable. The dataset was split into training and test sets to allow for proper prediction modelling. When run on the training and test sets, the prediction model returned 95.7% and 97.1% accuracy respectively. Additionally, it was estimated that average religious exemption, when controlling for unvaccinated students without religious exemption, accounts for a 93.8% decrease in the odds of a school having full herd immunity. However, there were concerns of overfitting and results of the model should be considered carefully. Assumption testing of the model showed it to have met all assumptions. Limitations of the dataset included the fact that it did not lend itself to an in-depth, individual analysis and reasons for being unvaccinated beyond religious exemption were not provided. Ethical concerns were noted related to the unknown reasons for non-vaccination. Thus, it does appear to be possible to predict whether full herd immunity for a school can or cannot be achieved as average religious exemption rate increase or decreases.

References

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

David Robinson and Alex Hayes (2020). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.4. https://CRAN.R-project.org/package=broom

Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2019). skimr: Compact and Flexible Summaries of Data. R package version 2.0.2. https://CRAN.R-project.org/package=skimr

Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.

Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

Hadley Wickham, Romain Fran?ois, Lionel Henry and Kirill M?ller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.4. https://CRAN.R-project.org/package=dplyr

Jeffrey B. Arnold (2019). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 4.2.0. https://CRAN.R-project.org/package=ggthemes

Majumder, M. S., Cohn, E. L., & Mekaru, S. R. (2015). Substandard Vaccination Compliance and the 2015 Measles Outbreak. *JAMA Pediatrics, 169*(5), 494-495.

Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-85. https://CRAN.R-project.org/package=caret

Max Kuhn and Hadley Wickham (2020). tidymodels: Easily Install and Load the 'Tidymodels' Packages. R package version 0.1.0. https://CRAN.R-project.org/package=tidymodels

Ontario. (2018, July 24). *Immunization.* Ontario Ministry of Health, Ministry of Long-term Care. http://www.health.gov.on.ca/en/pro/programs/immunization/ispa.aspx

Oxford Vaccine Group. (2016, April 26). *Herd Immunity: How does it work?.* https://www.ovg.ox.ac.uk/news/herd-immunity-how-does-it-work

Plans-Rubio, P (2012). Evaluation of the establishment of herd immunity in the population by means of serological surveys and vaccination coverage. *Human Vaccines & Immunotherapeutics, 8*(2), 184-188.

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/.

Sharla Gelfand (2019). opendatatoronto: Access the City of Toronto Open Data Portal. R package version 0.1.1. https://CRAN.R-project.org/package=opendatatoronto

Watson, S. (2018, December 3). *What's Herd Immunity, and How Does It Protect Us?.* WebMD. https://www.webmd.com/vaccines/news/20181130/what-herd-immunity-and-how-does-it-protect-us

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Appendices

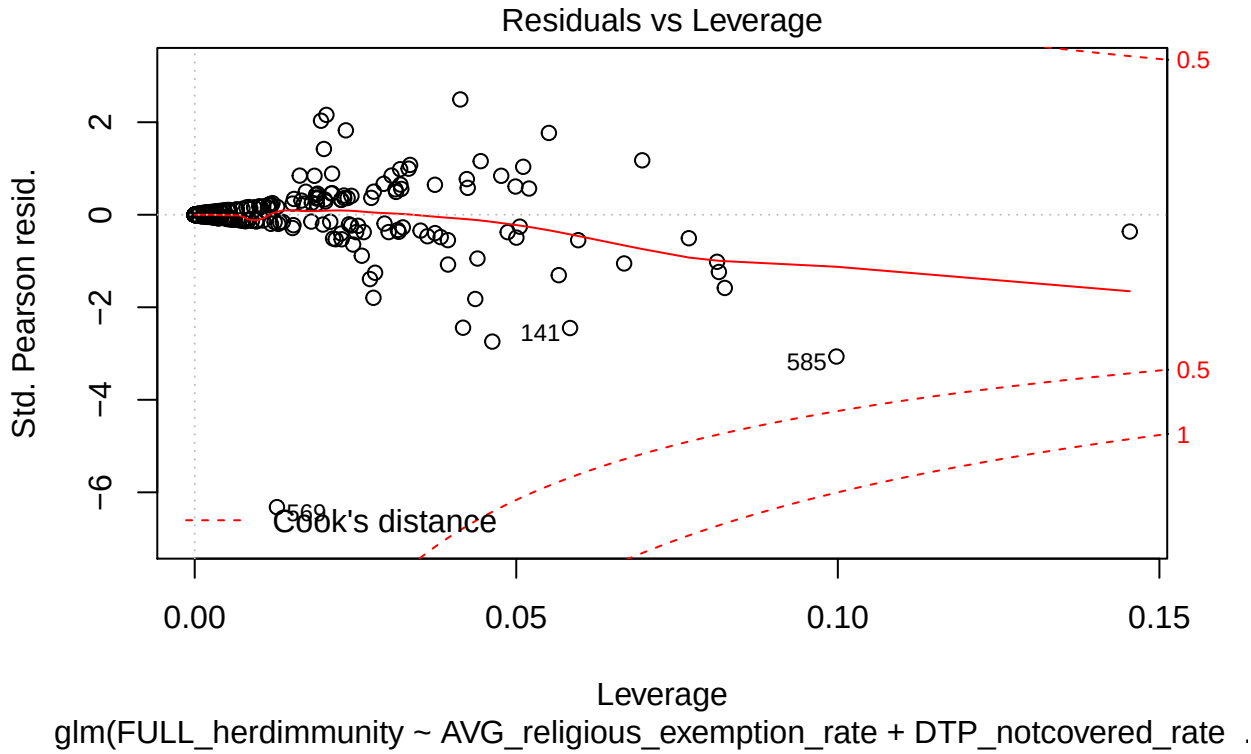
Appendix A

Summary of Model, for additional clarity:

```
##
## Call:
## glm(formula = FULL_herdimmunity ~ AVG_religious_exemption_rate +
##     DTP_notcovered_rate + MMR_notcovered_rate, family = "binomial",
##     data = herd_imm_train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.72012  -0.01583  -0.00006   0.00000   1.96937
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    14.4236     2.1188   6.808 9.93e-12 ***
## AVG_religious_exemption_rate   -2.7834     0.4212  -6.608 3.88e-11 ***
## DTP_notcovered_rate          -166.0645    25.9334  -6.403 1.52e-10 ***
## MMR_notcovered_rate          -153.7818    29.8012  -5.160 2.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 604.12  on 563  degrees of freedom
## Residual deviance: 101.18  on 560  degrees of freedom
## AIC: 109.18
##
## Number of Fisher Scoring iterations: 10
```

Appendix B

Figure 4 demonstrating influential points



Appendix C

```
# knitr::opts_chunk$set(echo = TRUE)
#
# # Load necessary libraries.
# library(opendatatoronto)
# library(dplyr)
# library(tidyverse)
# library(tidymodels)
# library(ggthemes)
# library(plyr)
# library(broom)
# library(skimr)
# library(reshape2)
# library(caret)
#
# #Load data
# immunization_load_new <- search_packages("immunization") %>%
#   list_package_resources() %>%
#   filter(name == "immunization-coverage-2018-2019") %>%
#   get_resource()
#
# immunization_load_new <- immunization_load_new %>%
#   rename(c("DTP coverage rate (%)" = "DTP_coverage_rate",
#   "MMR coverage rate (%)" = "MMR_coverage_rate", "DTP Religious exemption rate (%)" =
#   "DTP_religious_exemption_rate", "MMR Religious exemption rate (%)" =
#   "MMR_religious_exemption_rate"))
#
# immunization_load_new <- immunization_load_new[,-1]
#
# head(immunization_load_new)
#
# #skim data
# immunization_load_skim <- immunization_load_new %>% rename(c(
#   "Enrolled population" = "Enrolled Pop",
#   "DTP_coverage_rate" = "DTP coverage %",
#   "DTP_religious_exemption_rate" = "DTP relg exempt %",
#   "MMR_coverage_rate" =  "MMR coverage %",
#   "MMR_religious_exemption_rate" = "MMR relg exempt %"
# )) %>% select(`School Name`, "Enrolled Pop","DTP coverage %","DTP relg exempt %",
#               "MMR coverage %","MMR relg exempt %")
# skim(immunization_load_skim)
#
# #first ggplot
# p1 <- ggplot(immunization_load_new) +
#   geom_histogram(aes(x = DTP_coverage_rate, y= ..density..), position = 'dodge', bins = 100
#                  , fill = "blue", colour = "black") +
#   xlim(70,100) +
#   ylim(0,.18) +
#   geom_vline(xintercept =94, linetype="dashed", color = "red") +
#   geom_text(aes(x=86, label="herd immunity: 94%", y=.18), colour="black") +
#   theme_gdocs() +
#   labs(title = "DTP immunization frequency",
#        x = "Rate between 70% and 100%",
```

```
#        y = "Frequency of Schools")
#
# p2 <- ggplot(immunization_load_new) +
#   geom_histogram(aes(x = MMR_coverage_rate, y= ..density..), position = 'dodge', bins = 100
#                  , fill = "darkgreen", colour = "black") +
#   xlim(70,100) +
#   scale_y_continuous(position = "right") +
#   geom_vline(xintercept =95, linetype="dashed", color = "red") +
#   geom_text(aes(x=87, label="herd immunity: 95%", y=.18), colour="black") +
#   theme(axis.title.y.right = element_text( angle = 90)) +
#   theme_gdocs() +
#   labs(title = "MMR immunization frequency",
#        x = "Rate between 70% and 100%",
#        y = "Frequency of Schools")
# grid.arrange(p1, p2, ncol=2)
#
# #data model transformation
#
# # conservative recommendation for percentage of the population to be vaccinated
# # for herd immunity.
# dtp_herdimmunity_new = immunization_load_new$DTP_coverage_rate > 94.0
#
# # conservative recommendation for percentage of the population to be vaccinated
# # for herd immunity.#
# mmr_herdimmunity_new = immunization_load_new$MMR_coverage_rate > 95.0

# full herd immunity, if both DTP and MMR percentages are met.
# herdimmunity_combined_new = immunization_load_new$DTP_coverage_rate > 94.0,
# & immunization_load_new$MMR_coverage_rate > 95.0
# dtp_students_new = (immunization_load_new$DTP_coverage_rate/100)*,
# immunization_load_new$`Enrolled population`
# mmr_students_new = (immunization_load_new$MMR_coverage_rate/100)*,
# immunization_load_new$`Enrolled population`
# dtp_relg_students_new = (immunization_load_new$DTP_religious_exemption_rate/100)*,
# immunization_load_new$`Enrolled population`
# mmr_relg_students_new = (immunization_load_new$MMR_religious_exemption_rate/100)*,
# immunization_load_new$`Enrolled population`
# dtp_no_new = immunization_load_new$`Enrolled population` - dtp_students_new,
# - dtp_relg_students_new
# dtp_no_rate_new = dtp_no_new / immunization_load_new$`Enrolled population`
# mmr_no_new = immunization_load_new$`Enrolled population` - mmr_students_new,
# - mmr_relg_students_new
# mmr_no_rate_new = mmr_no_new / immunization_load_new$`Enrolled population`
# avg_relg_rate_new = (immunization_load_new$DTP_religious_exemption_rate,
# + immunization_load_new$MMR_religious_exemption_rate) / 2
#
# immunization_start_new <- immunization_load_new %>%
#   mutate(DTP_students = as.integer(dtp_students_new)) %>%
#   mutate(MMR_students = as.integer(mmr_students_new)) %>%
#   mutate(DTP_relg_students = as.integer(dtp_relg_students_new)) %>%
#   mutate(MMR_relg_students = as.integer(mmr_relg_students_new)) %>%
#   mutate(DTP_religious_exemption_rate) %>%
#   mutate(MMR_religious_exemption_rate) %>%
#   mutate(AVG_religious_exemption_rate = avg_relg_rate_new) %>%
```

```
#   mutate(DTP_no = as.integer(dtp_no_new)) %>%
#   mutate(DTP_notcovered_rate = dtp_no_rate_new) %>%
#   mutate(MMR_no = as.integer(mmr_no_new)) %>%
#   mutate(MMR_notcovered_rate = mmr_no_rate_new) %>%
#   mutate(DTP_herdimmunity = as.integer(dtp_herdimmunity_new)) %>%
#   mutate(MMR_herdimmunity = as.integer(mmr_herdimmunity_new)) %>%
#   mutate(FULL_herdimmunity = as.integer(herdimmunity_combined_new))
#
# dtp_melt_new <- immunization_start_new %>%
#   select('School Name', DTP_herdimmunity, DTP_religious_exemption_rate,
#   DTP_notcovered_rate)
#
# mmr_melt_new<- immunization_start_new %>%
#   select('School Name', MMR_herdimmunity, MMR_religious_exemption_rate, MMR_notcovered_rate,
#   AVG_religious_exemption_rate, FULL_herdimmunity)
#
# immunization_melted_new <- join(dtp_melt_new,mmr_melt_new)
#
# immunization_melted_new <- immunization_melted_new[order,
# (immunization_melted_new$'School Name'),]
#
# immunization_melted_new <- immunization_melted_new %>% distinct()%>%
#   mutate(YEAR = '2018-2019')
#
# #binomial model
# set.seed(999) # Set random number seed for consistency
#
# herd_imm_pred <- sample(nrow(immunization_melted_new),
# floor(nrow(immunization_melted_new)*0.7))  # Create train and test data sets
# herd_imm_train <- immunization_melted_new[herd_imm_pred,]  # Create training dataset
# herd_imm_test <- immunization_melted_new[-herd_imm_pred,] # Create test dataset
#
# model_herd_imm <- glm(FULL_herdimmunity ~ AVG_religious_exemption_rate
#                       + DTP_notcovered_rate + MMR_notcovered_rate,
#                       family = "binomial", data = herd_imm_train)
#
# tidy(model_herd_imm)
#
# #model evaluation
# herd_imm_train_filtered <-
#   augment(model_herd_imm,
#           data = herd_imm_train,
#           type.predict = "response") %>%
#   select(-.hat, -.sigma, -.cooksd, -.std.resid) %>%
#   mutate(predict_full_herd_immunity = if_else(.fitted > 0.5, 1, 0),
#          FULL_herdimmunity_binary = as.factor(FULL_herdimmunity),
#          predict_full_herd_immunity_binary = as.factor(predict_full_herd_immunity))
#
#
# conf1 <- caret::confusionMatrix(data =
#                                 herd_imm_train_filtered$predict_full_herd_immunity_binary,
#                                 reference = herd_imm_train_filtered$FULL_herdimmunity_binary)
# conf1[2:3]
#
```

```
# # Test data predictions
# herd_imm_test_filtered <-
#   augment(model_herd_imm,
#           data = herd_imm_train,
#           newdata = herd_imm_test,
#           type.predict = "response") %>%
#   mutate(predict_full_herd_immunity = if_else(.fitted > 0.5, 1, 0),
#          FULL_herdimmunity_binary = as.factor(FULL_herdimmunity),
#          predict_full_herd_immunity_binary = as.factor(predict_full_herd_immunity))
#
# # Confusion matrix for testing set
# conf2 <- caret::confusionMatrix(data =
#                                 herd_imm_test_filtered$predict_full_herd_immunity_binary,
#                                 reference = herd_imm_test_filtered$FULL_herdimmunity_binary)
# conf2[2:3]
#
# #second ggplot
# training <- herd_imm_train_filtered %>%
#   select(FULL_herdimmunity_binary, .fitted) %>%
#   mutate(type = "Training Set")
#
# test <- herd_imm_test_filtered %>%
#   select(FULL_herdimmunity_binary, .fitted) %>%
#   mutate(type = "Test Set")
#
# both <- rbind(training, test)
# rm(training, test)
#
# both %>%
#   ggplot(aes(x = .fitted,
#              fill = FULL_herdimmunity_binary)) +
#   geom_histogram(binwidth = 0.05, position = "dodge") +
#   theme_gdocs() +
#   labs(x = "Estimated probability that school has full herd immunity",
#        y = "Number",
#        fill = "School has full herd immunity") +
#   scale_fill_brewer(palette = "Set1") +
#   facet_wrap(vars(type),
#              ncol = 2,
#              scales = "free_y")
#
# #probability and logit development
# probabilities <-  predict(model_herd_imm, type="response")
#
# predicted_classes <- ifelse(probabilities > 0.5, "pos", "neg")
#
# immun_lin <-  herd_imm_train %>%
#   select_if(is.numeric)
# predictors <- colnames(immun_lin)
#
# immun_lin <- immun_lin %>%
#   mutate(logit = log(probabilities/(1-probabilities))) %>%
#   gather(key = "predictors", value = "predictor.value", -logit)
#
```

```
# #third ggplot
# ggplot(immun_lin, aes(logit, predictor.value)) +
#   geom_point(size = .5, alpha = .5) +
#   geom_smooth(method = "loess") +
#   theme_gdocs() +
#   facet_wrap(~predictors, nrow = 2, ncol = 4, scale = "free_y", ) +
#   theme(strip.text.x = element_text(size = 9))
#
# #model tests, assumptions, plots, VIF
# plot(model_herd_imm,5)
#
# model_herd_imm_data <- augment(model_herd_imm) %>%
#   mutate(index = 1:nrow(augment(model_herd_imm)))
#
# car::vif(model_herd_imm)
```