Data Analysis Exam: Using Wealth and Health Status Measurements to Predict the Number of Visits to a Doctor/Health Professional

DA-003

January 2022

1 Executive Summary

People are more hesitant to look for medical care even when it might be necessary[10]. The decision to seek medical care is often based on how expensive it can be[3] and how bad the individual's illness is[8]. Since wealth and health status have an influence on whether an individual seeks medical care, we studied whether measurements of wealth and health status can be used to predict the number of visits to a doctor/health professional within the past two months. The data set that we analyzed contained various measurements of wealth and health status that could be associated with the number of visits. The purpose of this study was to create a model that predicts the number of visits to a doctor/health professional within the past two months based on sex, age, annual income, score on a health questionnaire, whether the individual has a chronic condition but not limited in activity, whether the individual has a chronic condition and limited in activity, and the type of health insurance the individual has. The final model utilized was a zero-inflated Poisson regression model due to excess 0s from the response variable and discrete outcomes. We found that every predictor variable has a significant association with the number of visits to a doctor/health professional and can conclude that these measurements can be used to predict the number of visits to a doctor/health professional within the past two months. These findings can motivate future research by investigating whether other factors of wealth, such as an individuals assets, or other factors of health, such as different types of chronic diseases, can be used to predict the number of visits to a doctor/health professional an individual makes within the past two months.

2 Introduction

The number of hospital visits to doctors and/or health professionals has continued to decrease in recent years.[4,6,10]. Reasons for this decrease include a negative past experience with a medical professional[4,6], fear and anxiety of medical procedures[4,10], cost[10], and because individuals do not believe their illness is severe enough to seek medical attention[10]. Even though visits to a medical professional are decreasing, individuals still visit medical professionals when seeking immediate medical attention[11], checking up for their chronic disease[8], or just want a regular checkup[11]. Although all of these factors are known to affect an individual's decision of whether or not they seek medical attention, which of these factors has the greatest impact on whether an individual visits a doctor/health professional? Do all of these factors influence how often an individual visits a doctor/health professional? With this in mind, we aim to answer the question of whether wealth and/or health status are significantly associated with an individual's visit to a doctor/health professional.

Since previous research[3,6,8-9,11] indicates that wealth and health status affect an individual's decision to visit a doctor/health professional, we hypothesize that wealth and health status are both significantly associated with the number of visits to a doctor/health professional. Although we expect both wealth and health status to be association with the number of visits, we expect to there to be a stronger association of number of visits with health status compared to wealth because health risks have a greater urgency for medical attention[8,11]. To test this hypothesis, we analyzed the number of visits to a doctors and/or health professional in the past two months, **Hvisits**, in a data set that contains 2861 observations with 8 variables. In addition to general information such as **Sex** and **Age**, the other variables considered for the model were those related to an individual's wealth and an individual's health. In order to determine if the factors for wealth and health status were significant, we employed a zero-inflated Poisson regression model (ZIP) to predict the number of hospital visits to doctors and/or health professionals in the past two months. A ZIP regression model was used due to the variable **Hvisits** containing only positive integer values and excess number of 0s. To apply the ZIP regression model, we must ensure that the assumptions for a Poisson distribution are met. We will compare the AIC of both the best Poisson regression model and the best ZIP regression model to determine which is the better predictor of visits to doctors/health professionals in the past two months.

3 Exploratory Data Analysis

3.1 Summary Tables

To begin the analysis, we first took a look at the data set. No entries were missing so all observations were included. All of the variables in the data set were included since they all had a relation to either health status or wealth. The full data set has a total of 2861 observations with 8 variables ranging from the number of hospital visits to doctors and/or health professionals in the past two months(**Hvisits**), **Sex**, **Age**, and more. This analysis will be focusing on **Hvisits** the number of hospital visits to doctors and/or health professionals in the past two months. The possible predictor variables of interest are: **Age**, **Sex**, an individual's annual income (**Income**, an individual's score of a 12 health questionnaire with a higher score indicating bad health **Hscore**, whether an individual has a chronic condition but not limited in activity **Chronic1**, whether an individual has a chronic condition and limited in activity **Chronic2**, and whether an individual is covered by private insurance or the government(**Private Insurance**). From the summary statistics in table 1, we can see that the median of the visits to a doctor/health professional and the score of the health questionnaire are both 0 and have extremely low means. This indicates that these variables have over half of the observations equal to 0 which could lead to a very right skewed distribution of the variables. The excess 0 responses can be explained due to individuals overestimating their health[2] which can lead to a lower amount of individuals

going to visit a doctor/health professional.

Summary Statistics	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max
Hvisits	14	0	0.5	1.3	0.0	0.0	14.0
Age	12	0	40.5	20.4	19.0	32.0	72.0
Income	14	0	582.7	363.9	0.0	550.0	1500.0
Hscore	13	0	1.2	2.1	0.0	0.0	12.0

Table 1: This is a summary table of the quantitative variables, hvisits, age, income, and hscore.

3.2 Visualizing the Response Variable and Predictor Variables

As shown in the histogram in figure 1, the outcomes for **Hvisits** are discrete and contain only non negative counts. There is also a high frequency of 0s. This high frequency of 0s can be problematic since it would be difficult to transform **Hvisits** so that it can become normally distributed. These characteristics of **Hvisits** indicate that a Poisson regression model can be used to predict the number of visits to a doctor/health professional in the past two month. The high frequency of 0s are likely to fail the criteria of mean being equal to variance for a Poisson Regression model. In order to take the excess 0s into account, we will also consider a ZIP regression model.

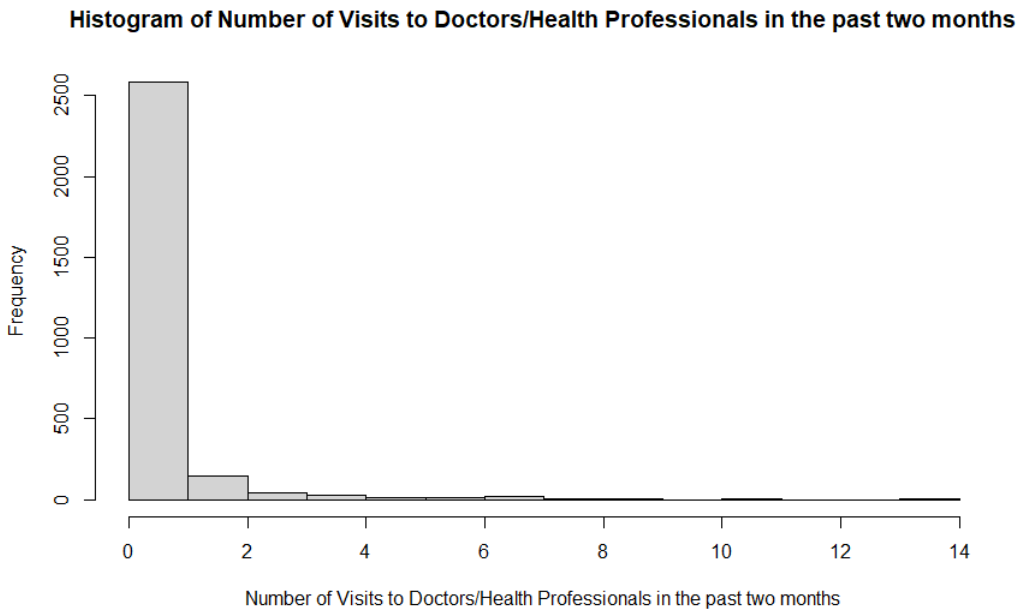

Figure 1: This histogram of Number of Visits to Doctors/Health Professionals in the past two months. The histogram is skewed to the right and shows an excess of the response 0.

The quantitative predictor variables are all discrete and are non positive. There appears to be a positive, yet small, correlation between visits to a doctor/health professional and both age and the health questionnaire score (Figure 2). An individual's income has a small negative correlation with the number of visits to a doctor/health professional (Figure 2). All of these quantitative variables appear to have a linear association with **Hvisits**. There is about equal proportion of each categorical variable in the data set with the exception of the categorical variables related to chronic conditions. When looking at plots of **Hvisits** vs **Hscore** separated by the groups in **Sex** and **Private Insurance**, there appears to be a difference in

trends between each group (Figure 3). The trends from the plots appear to show **Hscore** as a quadratic. The differences in the shapes also suggest that an interaction may be occurring with the two categorical variables and **Hscore**.The final regression model should reflect if both of are significant.

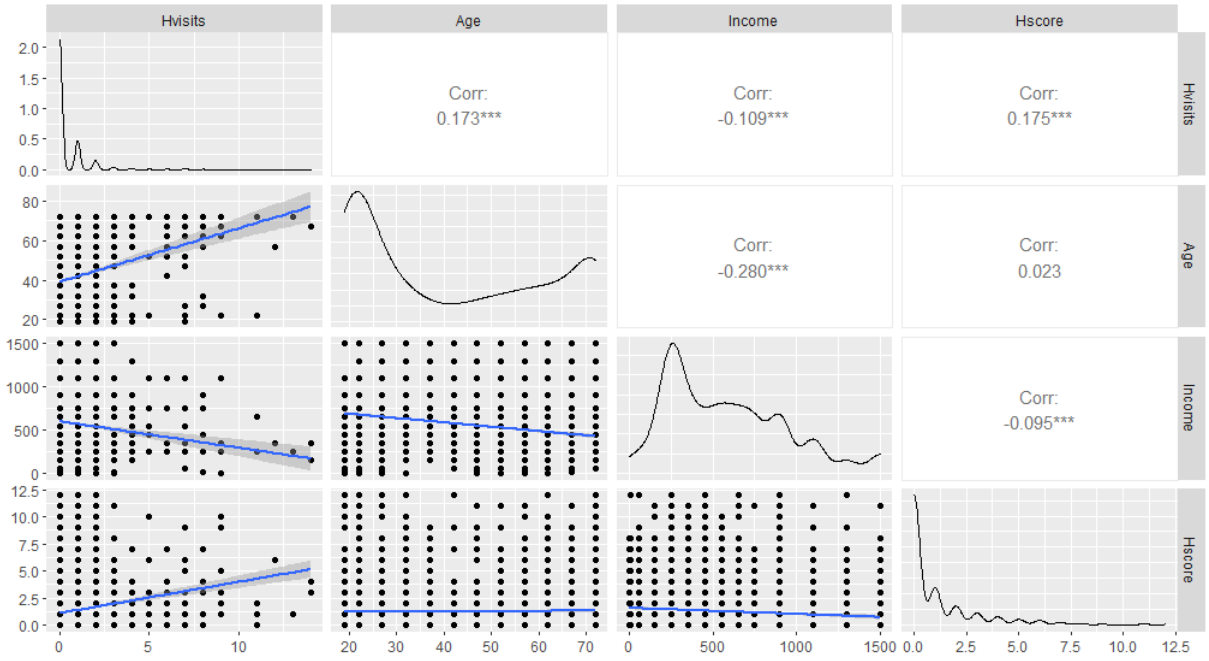

Figure 2: This is pairwise plot of each quantitative variable

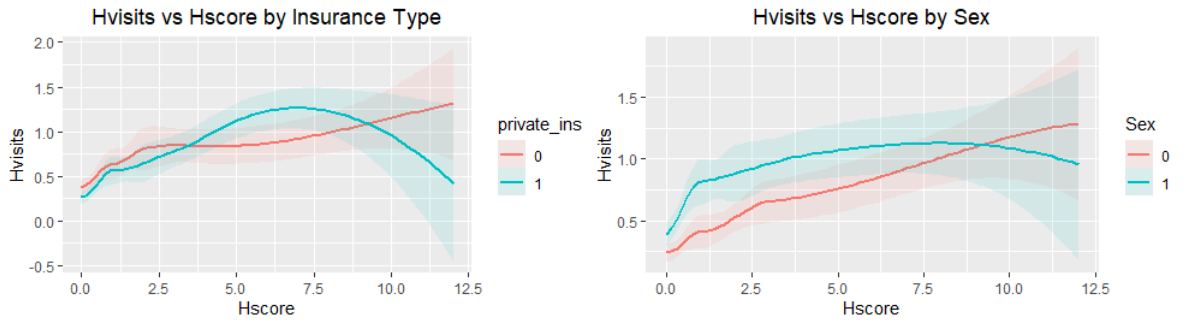

Figure 3: These plots of **Hvisits** vs **Hscore** show how the shape of each group for **Sex** and **Private Insurance** are different. Males are indicated by 0 and Females are indicated by 1 for the plot with **Sex**. Those that have government health insurance are indicated by 0 and those that have private health insurance are indicated by 1 for the plot with **PrivateInsurance**.

4 Statistical Analysis

4.1 Variable Selection

A full Poisson regression model was created with every variable of interest as a predictor variable. We performed a step-wise regression in both forward and backward direction to find the best model. This step-wise regression produced Akaike's An Information Criterion(AIC) which was used to determine the best predictor variables. The best fitting model with the smallest AIC value included every variable including the interactions with **Hscore** and both **Sex** and **Private Insurance**.

4.2 Assumptions for the Zero-Inflated Poisson Model

The criteria of a Poisson regression model includes: the response variable outcomes must being positive counts, they must be independent from each other, and the mean and variance must be equal. The **Hvisits** variable outcomes are all whole numbers that are greater than or equal to 0. We assumed that each individual response of visiting the doctor/health professional is independent since we were not given any information on how they were selected or how the information was collected. We also checked for over/under dispersion of the best fitting Poisson model. Table 2 shows how the Dispersion ratio is greater than 1 indicating overdispersion. With a p-value of <0.001, we can see that the variance is much greater than the mean. This means that we do not meet this criteria, but due to the high frequency of 0s, we implemented a ZIP regression model (ZIP). A ZIP regression model has two models to it, a Poisson count model and the logit zero-inflated model. The purpose of the zero-inflated model is to predict whether or not the participant did not visit a doctor/health professional within the past two months. The Poisson count model predicts the number of visits that were made if the individual was predicted to visit the doctor/health professional. From table 2, we compared the AIC values of the best Poisson regression model, a full ZIP regression model (all predictor variables are included), and a reduced ZIP regression model with only significant coefficients. The simple Poisson regression model had the greatest AIC value while the reduced ZIP regression model had the lowest AIC value (Table 2). This means that the best fitting model to predict the visits to a doctor/health professional was the reduced ZIP regression model.

Test for Overdispersion			
Data	Dispersion Estimate	Z-value	P-value
poisson_reg	2.77	5.74	< 0.001

Model Comparisons using AIC		
Models	df	AIC
poisson_reg	11.00	5920.67
full_zip	19.00	5327.43
reduced_zip	16.00	5324.69

Table 2: This table contains the values for the Test for Overdispersion and Model Comparisons of different Poisson regression models. The Test for Overdispersion includes the z-value and p-value for the original best Poisson. The Model comparisons include the model being compared and its respective degrees of freedom and AIC value.

4.3 Model Description, Inference, and Interpretation

We tested our coefficients for the final model to see if they are significantly different from 0 at an alpha value of 0.05. Our null hypothesis was that none of the values are significantly different from 0 while the alternative hypothesis was that at least one predictor variable is significantly different from 0. All of the coefficient estimates included have a p-value less than 0.05 meaning that the coefficient estimates for the predictor variables in final ZIP regression model are statistically significantly different from 0 (Table 3). All of the predictor variables related to health status and wealth were associated with the number of visits to a doctor/health professional in the past two months.

95% confidence intervals were created for each coefficient estimate and their incident risk ratios (IRR)/Odds ratio(OR). Table 4 shows how the 95% confidence intervals for the coefficient estimates do not include 0. The 95% confidence intervals for IRR and OR do not include 1 further supporting how the coefficient estimates in the final model are significantly different from 0. The predictor variable in the final count model that appears to have the greatest impact on

Count Model Coefficients	Estimate	Std. Error	Z-value	P-value
Intercept	0.103	0.132	0.78	0.435
Age	0.006	0.002	3.57	<0.001
Income	-0.001	<0.001	-3.11	0.002
Hscore	0.158	0.034	4.73	<0.001
$Hscore^2$	-0.016	0.003	-4.76	<0.001
Chronic2	0.297	0.075	3.93	<0.001
Private Insurance	-0.462	0.101	-4.56	<0.001
Hscore:Private Insurance	0.095	0.002	3.95	<0.001
Zero-Inflation Model Coefficients				
Intercept	1.987	0.172	11.57	<0.001
Sex	-0.513	0.134	-3.84	<0.001
Age	-0.011	0.003	-3.55	<0.001
Hscore	-0.179	0.037	-4.87	<0.001
Chronic1	-0.527	0.123	-4.29	<0.001
Chronic2	-0.930	0.172	-5.41	<0.001
Private Insurance	-0.505	0.136	-3.71	<0.001
Sex:Hscore	0.096	0.047	2.02	0.043
Log-likelihood:	-2646 on 16 DF			

Table 3: This is a summary table of the final zero-inflated regression model. It includes the estimates of all coefficients along with their standard errors, z-values, and p-values.

the number of visits to a doctor/health professional appears to be **Private Insurance**(Table 4). Among those that visited a health professional with every other variable remaining constant, if the individual has private health insurance the expected number of visits to a doctor/health professional in the past two months decreases by a factor of 0.63 (Table 4). The interaction for **Hscore** and **Private Insurance** indicates that when an individual has private health insurance, as their **Hscore** increases by 1 unit, the expected number of visits by to a doctor/health professional in the past two months decreases by a factor of 0.81 (Table 4). This factor was calculated by multiplying all of the LRRs related to the interaction together. The results from the final count regression model indicate that variables related to both wealth and health status are associated with the number of visits to a doctor/medical professional.

The final zero-inflation regression model has every coefficient estimate negative except for the interaction (Table 3). This indicates that almost every unit increase of the predictor variables for the zero-inflation regression model decreases the odds of not visiting a doctor/health professional. The variable with the lowest odds ratio was **Chronic2**. This means that the odds of not visiting a doctor/health professional in the past two months if you have a chronic condition(s) and limited in activity decreases, on average, decreases by a factor of 0.39 compared to an individual who does not have a chronic condition(s) but limited in activity (Table 4). The interaction of **Sex** and **Hscore** indicates that if the individual is a female, as their **Hscore** increases by 1 unit, the odds of not visiting a doctor/health professional in the past two months decreases by a factor of 0.55(Table 4). The results from the zero-inflation regression model supports our hypothesis since the variables related to health status appear to have the biggest impact on whether an individual visits a doctor/medical professional in the past two months.

5 Conclusions

Health status appears to have an influence in visits to a health professional. Although individuals that do not feel healthy are more likely to visit a doctor/health professional, the amount of times they visit in the past two months appears to decrease after they score a 5 in the health questionnaire. This drop could be due to the severity of their illness where the individual is in a

Count Model Coefficient	95% CI for Estimates	IRR	IRR 95% CI
Age	[<0.01,0.01]	1.01	[1.00*,1.01]
Income	[<-0.01,<-0.01]	1.00**	[.99,1.00**]
Hscore	[0.09,0.22]	1.17	[1.10,1.25]
$Hscore^2$	[-0.02,-0.01]	0.98	[0.98,0.99]
Chronic2	[0.15,0.44]	1.35	[1.16,1.56]
Private Insurance	[-0.66,-0.26]	0.63	[0.52,0.77]
Hscore:Private Insurance	[0.05,0.14]	1.10	[1.05,1.15]

Zero-Inflation Coefficient	95% CI for Estimate	OR	OR 95% CI
Sex	[-0.78,-0.25]	0.60	[0.46,0.78]
Age	[-0.02,-0.00]	0.99	[0.98,1.00**]
Hscore	[-0.25,-0.11]	0.84	[0.78,0.90]
Chronic1	[-0.77,-0.29]	0.59	[0.46,0.75]
Chronic2	[-1.27,-0.59]	0.39	[0.28,0.55]
Private Insurance	[-0.77,-0.24]	0.60	[0.46,0.79]
Sex:Hscore	[0.01,0.19]	1.10	[1.00*,1.21]

Table 4: This table provides all of the 95% confidence intervals for the coefficients estimates alonge with the IRRs/ORs and their 95% confidence intervals of the final ZIP regression model.*The number is rounded to the nearest hundredth but greater than 1.**The number is rounded to the nearest hundredth but less than than 1.

fatal stage. The low **Hscore** values near 1 could be from the ability to treat their illness within the first checkup with a health professional. Individuals with a chronic condition(s) are more likely to visit a health professionals due to the importance of checking up on chronic conditions[2] and the ease of accessibility of these checkups in modern times[1]. Since onset of chronic diseases increases as an individual gets older[9], it makes sense that age is another factor that affects whether an individual visits a doctor/health professional. The CDC's findings that women are more likely to go visit a doctor and have annual exams[7] is reflected in our final model since being a female influences whether they visit the doctor, but not how often.

All variables related to wealth were also significant as indicated by the final model. Even though the result for income was significant, an individual's income did not have a large influence on the increased number of visits. This is surprising since health care is expensive and is a factor when seeking medical attention[3]. This hesitancy is also based on the type of health insurance an individual has since health insurance provided by the government is cheaper than private insurance and allows people to obtain medical attention at a much lower value[3]. The model does show how the likeliness of visiting a health professional does increase when an individual is in bad health regardless of insurance.

A limitation of the data is our lack of information of how the data was collected. Although we assumed that each participant's response is independent from each other, it is possible that some of the participants might have affected each other's response. This could be from multiple participants living within the same household, working in close quarters, or being close enough to influence each other's responses.

The findings of this model show that wealth and health status can be used to predict the number of visits to a doctor/health professional within the past two months. Future research can look into other factors of wealth like an individual's assets and include whether an individual has health insurance at all. Additionally, a future study can look into more specific health status ailments such as cancer, cardiovascular disease, COPD, and other chronic conditions to see if a specific chronic condition has a greater influence on the number of visits to a health professional. An additional variable not included in this data set that can affect the number of visits to a doctor/health professional is the individual's region or city since a highly populated area may have increased availability of health professionals[5].

6 References

1. Dixon, Ronald F, and Latha Rao. "Asynchronous Virtual Visits for the Follow-up of Chronic Conditions." Telemedicine Journal and e-Health : the Official Journal of the American Telemedicine Association, U.S. National Library of Medicine, 27 July 2014, https://pubmed.
ncbi.nlm.nih.gov/24784174/.

2. Grauman, Åsa, et al. "Good General Health and Lack of Family History Influence the Underestimation of Cardiovascular Risk: A Cross-Sectional Study." OUP Academic, Oxford University Press, 22 Mar. 2021, https://academic.oup.com/eurjcn/article/20/7/676/6179500?login=true.

3. Kaestner, Robert, and Darren Lubotsky. "Health Insurance and Income Inequality." The Journal of Economic Perspectives, vol. 30, no. 2, American Economic Association, 2016, pp. 53–77, http://www.jstor.org/stable/43783707.

4. Lacy, Naomi L., et al. "Why We Don't Come: Patient Perceptions on No-Shows." Annals of Family Medicine, The Annals of Family Medicine, 1 Nov. 2004, https://www.annfammed.org/content/2/6/541.short.

5. Machado, Sara R., et al. "Physician Density in Urban and Rural Counties in the US, 2010 to 2017." JAMA Network Open, JAMA Network, 22 Jan. 2021, https://jamanetwork.com/journals /jamanetworkopen/fullarticle/2775404.

6. Meyerson, Beth E., et al. "I Don't Even Want to Go to the Doctor When I Get Sick Now: Healthcare Experiences and Discrimination Reported by People Who Use Drugs, Arizona 2019." International Journal of Drug Policy, Elsevier, 15 Jan. 2021, https://www.sciencedirect.com/science/article/abs/pii/S0955395921000116.

7. "NCHS Pressroom - 2001 News Release - Women Visit Doctor More Often than Men." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 6 Oct. 2006, https://www.cdc.gov/nchs/pressroom/01news/newstudy.htm.

8. Østbye, Truls, et al. "Is There Time for Management of Patients with Chronic Diseases in Primary Care?" Annals of Family Medicine, Copyright 2005 Annals of Family Medicine, Inc., 2005, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1466884/.

9. Prasad, Sahdeo, et al. "Age-Associated Chronic Diseases Require Age-Old Medicine: Role of Chronic Inflammation." Preventive Medicine, Academic Press, 9 Dec. 2011, https://www.sciencedirect.com/science/article/abs/pii/S0091743511004749.

10. Taber, Jennifer M., et al. "Why Do People Avoid Medical Care? A Qualitative Study Using National Data - Journal of General Internal Medicine." SpringerLink, Springer US, 12 Nov. 2014, https://link.springer.com/article/10.1007/s11606-014-3089-1.

11. Varshini, Amirtha, S. Leslie Rani, and M. P. Brundha. "Awareness of annual doctor checkups among general population." Drug Invention Today 14.2 (2020).

7 Appendix

7.1 R Code:

```
%%begin novalidate
#load dataset
data <- read.csv("C:\\Users\\erick\\Documents\\SDSU\\DataAnalysis_Exam\\
DATA_DAE_Spring2022.csv", header  = FALSE)

#This renames the variables to proper names.
install.packages("dplyr")
library(dplyr)
data  <- rename(data,
                Hvisits = "V1", #Number of visits to doctors/health
                #professionals in the past two months
                Sex = "V2", #1 if female, 0 if male
                Age = "V3", #age in years
                Income = "V4", #annual income in U.S. dollars
                Hscore = "V5", #12-item health  questionnaire score, high score
                #indicates bad health
                Chronic1 = "V6", #1 if chronic conditions but not limited
                #activity 0 otherwise
                Chronic2 = "V7", #1 if chronic conditions and limited activity,
                #otherwise
                private_ins = "V8")  #if covered by private insurance, 0 by gov

#Now we change the values of 0 and 1 to a response easier to understand
#THIS IS ONLY DONE FOR AESTHETICS. IT CHANGES BACK IN THE NEXT STEP
data$Sex[data$Sex == "0"] <- "Male"
data$Sex[data$Sex == "1"] <- "Female"
head(data)

data$Chronic1[data$Chronic1 == "0"] <- "Yes"
data$Chronic1[data$Chronic1 == "1"] <- "No"
head(data)

data$Chronic2[data$Chronic2 == "0"] <- "Yes"
data$Chronic2[data$Chronic2 == "1"] <- "No"
head(data)

data$private_ins[data$private_ins == "0"] <- "Private"
data$private_ins[data$private_ins == "1"] <- "Government"

#This will give us the results shown in Figure A1
head(data)
dim(data)

#This is to change the categorical variables from integers into factors.
#BE SURE TO REPEAT THE LOADING IN DATA AND RENAME.
class(data$Sex)
data$Sex <- as.factor(data$Sex)
```

```r
class(data$Sex)

class(data$Chronic1)
data$Chronic1 <- as.factor(data$Chronic1)
class(data$Chronic1)

class(data$Chronic2)
data$Chronic2 <- as.factor(data$Chronic2)
class(data$Chronic2)

class(data$private_ins)
data$private_ins <- as.factor(data$private_ins)
class(data$private_ins)



#This creates the values for the summary table 1, but table itself was
#modified manually for aesthetics

library(modelsummary)
library(skimr)
datasummary_skim(select(data,
                        Hvisits,
                        Age,
                        Income,
                        Hscore),
                 histogram = TRUE, output="latex")

#This is the histogram for figure 1
hist(data$hvisits,
main = "Histogram of Number of Visits to
    Doctors/Health Professionals in the past two months",
     xlab = "Number of Visits to Doctors/Health Professionals
     in the past two months",
     output = "latex")
hist(data$hscore)

#This makes the scatterplots in figure 2
library(GGally)
options(repr.plot.width=10, repr.plot.height=8)
plot_frame_dae <- data.frame("Hvisits" = data$Hvisits,
                             "Age" = data$Age,
                             "Income" = data$Income,
                             "Hscore" = data$Hscore)

ggpairs(plot_frame_dae,lower = list(continuous = wrap(my_fn, method="lm")))

#This creates the graphs with loess smooths in Figure 3
options(repr.plot.width=8, repr.plot.height=5)
hscore_sex <- ggplot(data, aes(x=Hscore, y=Hvisits, color=Sex, alpha = 0.1)) +
  #geom_point() +
```

```
  geom_smooth(method=loess, aes(fill=Sex),  alpha = 0.1)

hscore_sex + ggtitle("Hvisits vs Hscore by Sex") +
  theme(plot.title = element_text(hjust = 0.5))


options(repr.plot.width=8, repr.plot.height=5)
hscore_ins <- ggplot(data, aes(x=Hscore, y=Hvisits, color=private_ins,
    alpha = 0.1)) +
  #geom_point() +
  geom_smooth(method=loess, aes(fill=private_ins), alpha = 0.1)

hscore_ins + ggtitle("Hvisits vs Hscore by Insurance Type")+
  theme(plot.title = element_text(hjust = 0.5))

#This creates the best simple Poisson regression model
#The raw results are in figure A2
poisson_reg <- glm(Hvisits ~ Sex +
                     Age +
                     Income +
                     Hscore +
                     I(Hscore^2)+
                     Chronic1 +
                     Chronic2+
                     private_ins*Hscore+
                     Sex*Hscore+
                     private_ins,
                   data = data, family = "poisson")
summary(poisson_reg)

#This is the stepwise regression to find the best poisson model
#the raw results are in figure A4
pois <- step(poisson_reg)
#This tests for overdispersion in table 3
library(AER)
dispersiontest(poisson_reg)


#This creates a full ZIP regression model with every variable
#The raw results are in Figure  A5
library(pscl)
full_zip <- zeroinfl(Hvisits ~ Sex +
                     Age +
                     Income +
                     Hscore +
                       I(Hscore^2)+
                     Chronic1 +
                     Chronic2 +
                     private_ins+
                       private_ins*Hscore|
                     Sex +
```

```
                    Age +
                    Income+
                    Hscore +
                    Chronic1 +
                    Chronic2 +
                    private_ins+
                    Sex*Hscore, data=data)
summary(full_zip)

#This creates the best reduced ZIP regression model.
#the raw results are in Figure A6
best_zip <- zeroinfl(Hvisits ~
                       Age +
                       Income +
                       Hscore +
                         I(Hscore^2)+
                         Chronic2+
                       private_ins+
                         private_ins*Hscore
                     |
                       Sex +
                       Age +
                       Hscore+
                       Chronic1+
                       Chronic2 +
                       private_ins+
                       Sex*Hscore,
                     data=data)

#This is the values of the coefficients for table 4
summary(best_zip)

#AIC comparison of the simple poisson model, full ZIP,
#and reduced ZIP in table 3
#the raw results are in Figure A7
AIC(poisson_reg, full_zip,best_zip)

#This creates the IRRs and ORs
#the raw results are in Figure A8
zip_coef <- best_zip$coefficients
zip_coef
unlisted_zip_coef <-unlist(zip_coef)
exp(unlisted_zip_coef)

#This creates the 95% confidence intervals for the estimates and IRRs/ORs
#the raw results are in Figure A9
confint(best_zip)
exp(confint(best_zip))

## Error:  <text>:1:1:  unexpected SPECIAL
## 1:  %%
```

```
##      ^
```

7.2 Raw Output from R Code:

	Hvisits	Sex	Age	Income	Hscore	Chronic1	Chronic2	private_ins
1	0	Male	62	250	2	Yes	Yes	Private
2	0	Female	22	650	3	No	Yes	Government
3	0	Female	22	900	0	Yes	Yes	Private
4	0	Female	27	1100	0	Yes	Yes	Government
5	0	Male	32	750	0	Yes	Yes	Government
6	0	Female	22	150	0	Yes	Yes	Government

Figure A1: This the raw summary table output of the data from R.

```
Call:
glm(formula = Hvisits ~ Sex + Age + Income + Hscore + I(Hscore^2) +
    Chronic1 + Chronic2 + private_ins * Hscore + Sex * Hscore +
    private_ins, family = "poisson", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5474  -0.9630  -0.6943  -0.3529   8.1173

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.874e+00  1.106e-01 -16.939  < 2e-16 ***
Sex1                3.289e-01  7.411e-02   4.438 9.08e-06 ***
Age                 1.302e-02  1.454e-03   8.954  < 2e-16 ***
Income             -2.416e-04  8.983e-05  -2.690 0.007140 **
Hscore              2.751e-01  3.061e-02   8.988  < 2e-16 ***
I(Hscore^2)        -1.848e-02  3.005e-03  -6.150 7.73e-10 ***
Chronic11           3.750e-01  6.674e-02   5.618 1.93e-08 ***
Chronic21           8.270e-01  7.668e-02  10.786  < 2e-16 ***
private_ins1       -1.579e-01  7.190e-02  -2.196 0.028119 *
Hscore:private_ins1  6.743e-02  1.986e-02   3.395 0.000687 ***
Sex1:Hscore        -4.559e-02  1.999e-02  -2.281 0.022568 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4823.8  on 2860  degrees of freedom
Residual deviance: 4116.5  on 2850  degrees of freedom
AIC: 5920.7

Number of Fisher Scoring iterations: 6
```

Figure A2: This the raw summary table output for the simple Poisson Regression Model from R.

```
        Overdispersion test

data:  poisson_reg
z = 5.7426, p-value = 4.663e-09
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  2.770787
```

Figure A3: This the raw result output for the simple Poisson Regression Model from R.

set.seed(73)

$s2_1 < -14.42495 mun < -9.292308 s2.postsample < -1/rgamma(10000, (v0 + n1)/2, s2_1 * (v0+n1)/2) s.postsample < -sqrt(s2.postsample) theta.postsample1 < -rnorm(10000, mun, sqrt(s2.postsamp n1))) theta.postsample1$

$s2_2 < -18.17819 mun < -6.94875 s2.postsample < -1/rgamma(10000, (v0 + n2)/2, s2_2 * (v0+n2)/2) s.postsample < -sqrt(s2.postsample) theta.postsample2 < -rnorm(10000, mun, sqrt(s2.postsamp n2))) theta.postsample2$

```
Start:  AIC=5920.67
Hvisits ~ Sex + Age + Income + Hscore + I(Hscore^2) + Chronic1 +
    Chronic2 + private_ins * Hscore + Sex * Hscore + private_ins

                     Df Deviance    AIC
<none>                    4116.5 5920.7
- Sex:Hscore          1   4121.7 5923.8
- Income              1   4123.9 5926.0
- Hscore:private_ins  1   4127.8 5929.9
- Chronic1            1   4148.7 5950.9
- I(Hscore^2)         1   4157.8 5960.0
- Age                 1   4199.0 6001.2
- Chronic2            1   4230.1 6032.2
```

Figure A4: This the raw result output for the stepwise regression for the simple Poisson regression model from R.

```
Call:
zeroinfl(formula = Hvisits ~ Sex + Age + Income + Hscore + I(Hscore^2) + Chronic1 + Chronic2 +
    private_ins + private_ins * Hscore | Sex + Age + Income + Hscore + Chronic1 + Chronic2 +
    private_ins + Sex * Hscore, data = data)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.3343 -0.5134 -0.3800 -0.2247 18.6189

Count model coefficients (poisson with log link):
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.1673207  0.1436803   1.165 0.244207
Sex1                -0.0122661  0.0751776  -0.163 0.870391
Age                  0.0052595  0.0019459   2.703 0.006875 **
Income              -0.0004230  0.0001177  -3.593 0.000326 ***
Hscore               0.1591954  0.0334956   4.753 2.01e-06 ***
I(Hscore^2)         -0.0159691  0.0033470  -4.771 1.83e-06 ***
Chronic11            0.0707000  0.0957790   0.738 0.460418
Chronic21            0.3375038  0.0993226   3.398 0.000679 ***
private_ins1        -0.4302443  0.1018002  -4.226 2.38e-05 ***
Hscore:private_ins1  0.0922107  0.0239283   3.854 0.000116 ***

Zero-inflation model coefficients (binomial with logit link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.2273823  0.2349495   9.480  < 2e-16 ***
Sex1         -0.5606020  0.1491732  -3.758 0.000171 ***
Age          -0.0128306  0.0032627  -3.933 8.41e-05 ***
Income       -0.0003523  0.0002296  -1.534 0.125023
Hscore       -0.1820985  0.0370203  -4.919 8.70e-07 ***
Chronic11    -0.4663229  0.1464532  -3.184 0.001452 **
Chronic21    -0.9118830  0.1815326  -5.023 5.08e-07 ***
private_ins1 -0.4333239  0.1409491  -3.074 0.002110 **
Sex1:Hscore   0.0947114  0.0476376   1.988 0.046793 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 27
Log-likelihood: -2645 on 19 Df
```

Figure A5: This the raw result output for the Full ZIP Poisson Regression Model from R.

$s2_3 < -13.09347 mun < -7.812381 s2.postsample < -1/rgamma(10000, (v0 + n3)/2, s2_3 * (v0+n3)/2) s.postsample < -sqrt(s2.postsample) theta.postsample3 < -rnorm(10000, mun, sqrt(s2.postsamp n3))) theta.postsample3$

mean(theta.postsample1 ¡ theta.postsample2 theta.postsample2 ¡ theta.postsample3)

```
Call:
zeroinfl(formula = Hvisits ~ Age + Income + Hscore + I(Hscore^2) + Chronic2 + private_ins +
    private_ins * Hscore | Sex + Age + Hscore + Sex * Hscore + Chronic1 + Chronic2 + private_ins,
    data = data)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.3445 -0.5142 -0.3785 -0.2278 18.2425

Count model coefficients (poisson with log link):
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.026e-01  1.315e-01   0.780 0.435430
Age                  6.366e-03  1.782e-03   3.572 0.000354 ***
Income              -2.973e-04  9.553e-05  -3.112 0.001859 **
Hscore               1.584e-01  3.351e-02   4.727 2.28e-06 ***
I(Hscore^2)         -1.592e-02  3.346e-03  -4.756 1.98e-06 ***
Chronic21            2.967e-01  7.547e-02   3.931 8.44e-05 ***
private_ins1        -4.624e-01  1.014e-01  -4.562 5.08e-06 ***
Hscore:private_ins1  9.478e-02  2.397e-02   3.954 7.67e-05 ***

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.986501   0.171636  11.574  < 2e-16 ***
Sex1        -0.513355   0.133734  -3.839 0.000124 ***
Age         -0.010852   0.003061  -3.546 0.000391 ***
Hscore      -0.179302   0.036821  -4.870 1.12e-06 ***
Chronic11   -0.527341   0.122894  -4.291 1.78e-05 ***
Chronic21   -0.929801   0.172014  -5.405 6.47e-08 ***
private_ins1 -0.504752   0.136131  -3.708 0.000209 ***
Sex1:Hscore  0.095793   0.047343   2.023 0.043036 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 23
Log-likelihood: -2646 on 16 Df
```

Figure A6: This the raw result output for the Final Reduced ZIP Poisson Regression Model from R.

```
             df      AIC
poisson_reg 11 5920.673
full_zip    19 5327.425
best_zip    16 5324.690
```

Figure A7: This the raw result output for the comparisons of the AIC between the simple Poisson regression model, the full zip regression model, and the reduced zip regression model from R.

```
 count.(Intercept)              count.Age           count.Income          count.Hscore
         1.1080438              1.0063858              0.9997028             1.1716398
 count.I(Hscore^2)       count.Chronic21   count.private_ins1 count.Hscore:private_ins1
         0.9842105              1.3454213              0.6297935             1.0994165
  zero.(Intercept)              zero.Sex1               zero.Age            zero.Hscore
         7.2899787              0.5984843              0.9892062             0.8358535
    zero.Chronic11       zero.Chronic21    zero.private_ins1      zero.Sex1:Hscore
         0.5901719              0.3946323              0.6036552             1.1005309
```

Figure A8: This the raw result output for the IRRs and ORs of the reduced zip regression model from R.

SDSU Red Model Coefficients	Coefficient Values	Standard Error	P-Values
Intercept	-1.51	0.64	0.018
Eco Round	-1.51	0.64	0.021
Set Play Before Round	-2.81	1.21	0.832
First Blood	3.83	0.71	0.002
Bomb Planted	1.59	1.23	0.011
Set Play and First Blood	-2.99	1.42	0.035
SDSU Red Model Coefficients	**Odds Ratios**	**95% Confidence Intervals**	
Intercept	0.22	[0.055-0.701]	
Eco Round	0.06	[0.003-0.440]	
Set Play Before Round	0.86	[0.210-3.513]	
First Blood	45.85	[5.848-1041.779]	
Bomb Planted	4.89	[1.524-18.139	
Set Play and First Blood	0.05	[0.002-0.621]	

	2.5 %	97.5 %
count_(Intercept)	-0.1552269905	0.360419264
count_Age	0.0028728223	0.009858204
count_Income	-0.0004844964	-0.000110037
count_Hscore	0.0927254358	0.224083187
count_I(Hscore^2)	-0.0224743910	-0.009356560
count_Chronic21	0.1487868922	0.444627578
count_private_ins1	-0.6610283399	-0.263698280
count_Hscore:private_ins1	0.0478024211	0.141756834
zero_(Intercept)	1.6501011094	2.322900142
zero_Sex1	-0.7754690876	-0.251240872
zero_Age	-0.0168512606	-0.004853635
zero_Hscore	-0.2514703195	-0.107133458
zero_Chronic11	-0.7682093055	-0.286473425
zero_Chronic21	-1.2669423917	-0.592659035
zero_private_ins1	-0.7715649456	-0.237939298
zero_Sex1:Hscore	0.0030013580	0.188584053

	2.5 %	97.5 %
count_(Intercept)	0.8562208	1.4339305
count_Age	1.0028770	1.0099070
count_Income	0.9995156	0.9998900
count_Hscore	1.0971605	1.2511751
count_I(Hscore^2)	0.9777763	0.9906871
count_Chronic21	1.1604257	1.5599091
count_private_ins1	0.5163201	0.7682053
count_Hscore:private_ins1	1.0489634	1.1522964
zero_(Intercept)	5.2075063	10.2052280
zero_Sex1	0.4604877	0.7778350
zero_Age	0.9832899	0.9951581
zero_Hscore	0.7776565	0.8984058
zero_Chronic11	0.4638429	0.7509070
zero_Chronic21	0.2816916	0.5528553
zero_private_ins1	0.4622890	0.7882505
zero_Sex1:Hscore	1.0030059	1.2075386

Figure A9: This the raw result output for the 95% confidence intervals of the estimates of the coefficients and the IRRs/ORs of the reduced zip regression model from R. The 95% confidence intervals of the estimates are on the left and the 95% confidence intervals of the IRRs/ORs are on the right.