PSTAT 220B Final Project

Jordan Garrett

12/3/19

Introduction

Communicating with one's healthcare practitioner serves as an integral component to staying healthy. Relative to physicians, pharmacists are arguably the most accessible practitioners who can provide advice on proper substance consumption. Further, they are capable of quickly and accurately diagnosing symptoms and prescribing necessary medications. Despite their utility, some may not take advantage of the services provided by their local pharmacist. Here, I analyzed data from a survey conducted in San Francisco to investigate the number of times an individual consulted with a pharmacist. More specifically, I applied a generalized linear model (GLM) to determine the characteristics of an individual that are most predictive of the number of times an they consult the pharmacist.

Methods

The following characteristics were collected from each individual (N = 500) who completed the survey (variable names are in bold):

- if they are a male or female (***sex***)
- ***age***
- annual income (***income***)
- if they have both private health insurance and pharmacy coverage (***lp***)
- if they have private health insurance without pharmacy coverage (***fp***)
- if they are not covered by private health insurance (***fr***)
- number of illnesses in the past 4 weeks (***ill***)

- number of self-reported days of reduced activity in the past 4 weeks due to illness or injury (***ad***)
- general health questionnaire score, where high scores indicate poor health (***hs***)
- has a chronic medical condition that does not limit activity (***ch1***)
- has a chronic medical condition that limits activity (***ch2***)

Each of these measures served as a potential covariate in the GLM to predict the number of times the individual visited the pharmacist in a 4 week period (***pc***).

When applying a GLM, the following protocol is recommended to avoid statistical fallacy: exploratory data analysis; formulation of hypotheses; model assessment, diagnostics, comparisons, selection and interpretation. This report assumes that the reader is familiar with the basics of a simple linear regression, thus proofs for equations are not presented.

1 Generalized Linear Models Primer

GLM refers to large class of regression models where a response variable $\boldsymbol{y_i}$ is assumed to follow an exponential family distribution (e.g., Normal, Poisson, Binomial, Bernoulli) with a mean of $\boldsymbol{\mu_i}$. This mean is assumed to be some function of $\boldsymbol{x_i^T \beta}$. There are three components of a GLM. The first is the random component, which represents the probability distribution of the response variable which has a density

$$f(y_i|\theta_i, \phi) = exp\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\} \quad for\ i = 1, 2, ...n \tag{1}$$

where $E(Y_i) = \mu_i$. This equation represents the general density function form of an exponential family distribution. Further, it is assumed that y_i is distributed independently. The second component is the systematic component, which specifies the linear combination of explanatory variables ($\boldsymbol{x_1, x_2, ...x_n}$) in the model:

$$\eta_i = \sum_{j=0}^{p-1} \beta_j x_{ij} \quad for\ i = 1, 2, ...n \tag{2}$$

The last component is the link function $g(\mu_i)$, which represents how the random and systematic components are related to one another:

$$\eta_i = g(\mu_i) \tag{3}$$

In other words, it indicates how the expected value of the response variable is related to the linear prediction. Each exponential family distribution has a canonical link function, which means $g(\mu_i) = \theta_i$. For example, the link function of a response variable that follows a binomial distribution is $\eta_i = g(\mu_i) = logit(E(Y_i))$. Note, a simple linear regression contains these three components, where the random component is $Y_i \overset{indep}{\sim} N(\mu_i, \sigma^2)$, the systematic component is the same, and $g(\mu_i) = I(\cdot)$. Multiple assumptions are made when using the GLM: y_i is distributed independently and follows an exponential family distribution, there is a linear relationship between the link function and explanatory variables, and errors are distributed independently. To determine the nature of each component when modeling the response variable *pc*, I first explored the distributions of each survey measure.

2 Exploratory Data Analysis

Prior to modeling, it is important to explore the data set and determine how each variable is distributed and related to one another. Descriptive statistics for each of the aforementioned variables are presented in **Table 1** below. The abundance of zeroes may be concerning at first sight, but note that the except for age, income, ill, and hs, all of the other variables are dummy coded factors. Thus, 1 indicates the first level of the factor (e.g., fr = 1 means they are not covered by private health insurance), while 0 indicates the second level (e.g., fr = 0 means they are covered by private health insurance). Looking at the spread of the data (**Figure 1A**) it is apparent that oultiers are present. Although this may just be a result of the dummy coding, extreme observations have the potential of influencing the relationship between a response variable and independent variables when using the GLM.

Survey Scores				
Variable	Mean	Median	Mode	Variance
pc	0.30	0	0	0.86
sex	0.52	1	1	0.25
age (yr.)	40.91	32	22	411.39
income ($)	571.78	550	250	130526.48
lp	0.40	0	0	0.24
fp	0.05	0	0	0.05
fr	0.24	0	0	0.18
ill	1.42	1	1	1.92
ad	0.94	0	0	8.94
hs	1.12	0	0	4.36
ch1	0.41	0	0	0.24
ch2	0.12	0	0	0.10

Table 1: Quantitative descriptive measures of survey participant's characteristics.

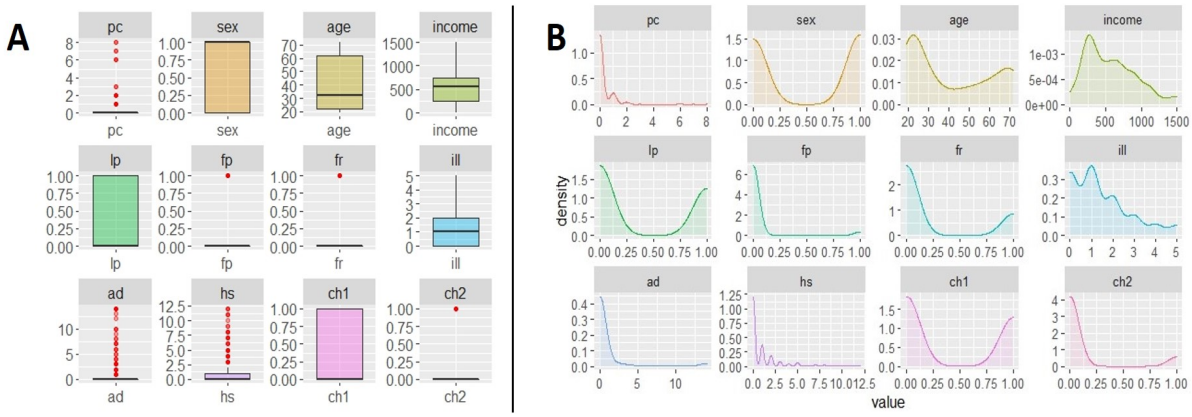

Figure 1: Exploratory data analysis. *a*) Box plots of survey measures. Red points represent outliers. *b*) Kernel density plots of distributions for each measure.

If not handled properly, it is possible to miss the true relationship between the response and independent variable, resulting in an inaccurate model. Decisions on removing these and other suspicious observations were later made when computing their influence based off their corresponding residual values (see **Model Diagnostics** below).

Next, I assessed the structure of how each measure is distributed using kernel density plots (**Figure 1B**). Focusing on the response variable *pc*, it is obvious that its distribution is non-normal. Thus, a standard linear regression that relies on the data to be normally distributed will likely fit the data poorly. Instead, a GLM with a random component that assumes the response variable follows a Poisson or Negative Binomial distribution was deemed to be appropriate for modeling. Note that the latter distribution is used to model count data when there is evidence for over-dispersion (i.e., $\mu_i < var(\mu_i)$). The mean and variance of *pc* suggest that over-dispersion may be present (**Table 1**). As expected, categorical variables followed a bimodal distribution. The distribution of continuous variables closely resembled that of a Poisson, suggesting that they may serve as the best predictors of *pc*. To assess how each measure was associated with one another, I computed the Pearson's correlation coefficient between each pair of variables (**2**). There was a significant positive

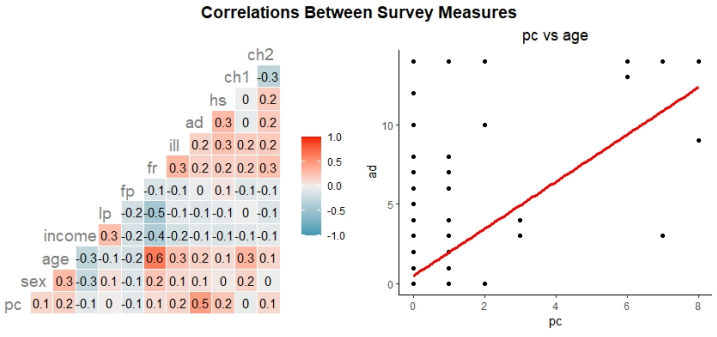

Figure 2: Correlations between survey measures. A strong positive linear association was found between the response variable *pc* and *ad*, suggesting that it will likely be a significant predictor.

correlation between *pc* and the covariate *ad* ($r(498) = 0.46, p < 0.001$), further suggesting that it is likely a strong predictor. All other correlations between *pc* and other covariates were fairly weak. In regards to the covariates, there were strong associations between fr with *income*, *age*, and *lp* ($r(498) = -0.39;\ 0.64;\ -0.46;\ p < 0.05$. respectively). In regression, correlations between covariates (i.e., multicollinearity) can be a concerning issue. Multicollinearity leads to an inflation in the variance of independent variables in the model, thus making it more difficult for their coefficients to be statistically significant. In addition, highly correlated regressors can be redundant and not make a significant contribution to the amount of variation explained by the model. After gaining a sense of the survey data, I proceeded to conduct both a Poisson and Negative Binomial regression to test the following null and alternative hypotheses:

$$H_0 : \beta_1 = \beta_2 = ... \ = \beta_p = 0 \quad for\ p = 1, 2, ..., 11 \tag{4}$$

$$H_1 : \text{at least one } \beta_p \neq 0 \tag{5}$$

In other words, the null hypothesis states that there is no relationship between covariates and *pc*, while the alternative hypothesis states that there is a relationship between *pc* and at least one survey measure.

3 Model Assessment

3.1 Analyses

There exists a multitude of potential models that explain the relationship between a response variable and predictors when applying the GLM. Therefore, it is important to exhaustively explore the model space to extract a subset of the best models. As previously mentioned, it was decided that either a Poisson or Negative Binomial distribution may serve as the best choice for the random component of the model. For the Poisson regression model this can be written as

$$Y_i \overset{indep}{\sim} Pois(\mu_i) \quad for\ i = 1, ..., 500 \tag{6}$$

with the density function

$$f(y_i|\mu_i) = \frac{exp(-\mu_i)\mu_i^{y_i}}{y_i!} \tag{7}$$

whereas the Negative Binomial distribution random component is written as

$$Y_i \overset{indep}{\sim} NB(\mu_i, \tau) \quad for\ i = 1, ..., 500 \tag{8}$$

Model Performance From Initial Analyses				
Model #	Random	Formula: $\boldsymbol{pc} =$	D_M	R^2
1	Poisson	$\beta_0 + \sum_{j=1}^{11} \beta_j x_{ij}$ for i = 1,2,...,500 and j = 1,2,...,11	402.55	0.33
2**	Poisson	$\beta_0 + \beta_1 x_{i1} + \beta_5 x_{i5} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_{i9} x_9$	329.33	0.33
3	Poisson	Model 1 + $\beta_{12} x_{i8} x_{i11} + \beta_{13} x_{i3} x_{i4} + \beta_{14} x_{i6} x_{i8} + \beta_{15} x_{i1} x_{i2} + \sum_{t=1}^{7} \beta_k x_{it} x_{i9}$ for $t = (1, 2, 4, 6, 7)$ and $k = (16, 17, ..., 20)$	342.96	0.43
4**	NB	$\beta_0 + \beta_1 x_{i1} + \beta_5 x_{i5} + \beta_7 x_{i7} + \beta_8 x_{i8}$	272.55	0.31

Table 2: Residual deviance (D_M) and total amount of deviance explained by each model (R^2). $x_1, x_2, ..., x_p = sex, age, income, lp, fp, fr, ill, ad, hs, ch1, ch2$, respectively. For each model except for the second and fourth, $i = 1, 2, ..., 500$. **For the second and fourth model four outliers were removed see (**Model Diagnostics**), making $N = 496$.

with the density function

$$f(y_i|\mu_i, \tau) = \frac{\Gamma(y_i + \tau)}{\Gamma(\tau) y!} \frac{\mu_i^y \tau^\tau}{(\mu_i + \tau)^{y_i + \tau}} \tag{9}$$

Note that Y_i = the number of times the i^{th} individual visited the pharmacist in a 4 week period. In the Negative binomial distribution, τ represents the shape parameter. For both of these random components, it is assumed that $E(Y_i) = \mu_i$. Furthermore, both models contain the same systematic component (Equation 2) and link function

$$g(\mu_i) = ln(\mu_i) = \eta_i \tag{10}$$

The residual deviance (D_M) of each model was computed to determine how well it accounted for the variation in pc. D_M represents how far the model in question is from a perfect idealistic model, and smaller values indicate that the tested model is closer to a perfect model. This measure is similar to the residual sums of squares in standard linear regression. D_M was then used to calculate the ratio of deviance accounted for by the model (R^2) from the equation

$$R^2 = 1 - \frac{D_M}{D_{H_0}} \tag{11}$$

where D_{H_0} is the amount of deviance accounted for by a null model with no parameters. Both goodness-of-fit statistics are presented in **Table 2** for a full model including all covariates (Model 1) and three of the top performing potential models. Note that ($\boldsymbol{x_1, x_2, ..., x_p = sex, age, income, lp, fp, fr, ill, ad, hs, ch1, ch2}$) respectively. Given that Model 2 and Model 4 yielded the smallest amount of residual deviance, they were se-

lected for diagnosis. Although these two models are the focus of this report, diagnostics and model selection measures were also performed on all other possible models.

3.2 Model Diagnostics

The nature of a model's residuals reveals the violation of assumptions, highly influential outliers, and the need to transform any variables. Diagnostic analyses focused on standardized residuals, which account for covariates that are on different scales and are better for testing non-constant variance. Both Model 2 and 4 satisfied the assumption of homogeneity and heteroskedasticity as indicated by the left column of plots in **Figure 3**. Further, four individuals were identified as potential outliers given the leverage and Cook's distance of their residuals. Although these residuals did not exceed conventional thresholds (e.g., Cook's distance > 1), each model fit improved according to model selection statistics discussed below (see **Model Comparisons and Selection**). In addition, a variance inflation test (VIF) was conducted using the R package "car" to determine if multicollinearity was an issue (Fox Weisburg, 2018). VIF scores greater than 5 indicate that the correlation between variables is inflating their estimated coefficients. For both Model 2 and 4, all covariates had a VIF score 1.

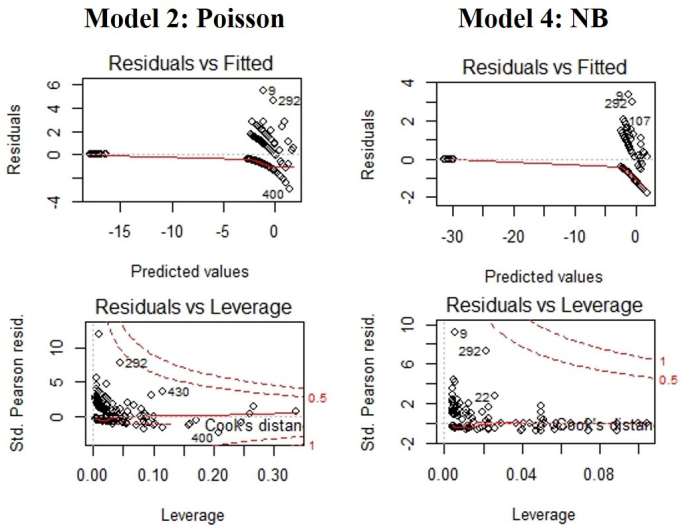

Figure 3: Residual plots for two top performing models. (*Top*) Residuals are plotted against predicted values. Red line represents a fitted loess model to determine if there are any deviations from linearity, indicating violations of constant variance. Neither model displayed overt violations of constant variance. (*Bottom*) Residuals plotted against leverage values, while dashed lines represent Cook's distance thresholds. Numerically labeled residuals depict those with the highest influence. Model fit tremendously improved after they were removed.

As aforementioned, over-dispersion was a potential concern. To determine its presence, $\hat{\phi}$ and $\widetilde{\phi}$ were estimated. When conducting a Poisson or Negative Binomial regression, it is assumed that $\phi = 1$, thus deviations from this assumption by its estimates indicates issues of over/under-dispersion. $\hat{\phi}$ was estimated by dividing the residual deviance of a model by its residual degrees of freedom, while $\widetilde{\phi}$ was estimated by calculating the sum of a model's squared Pearson residuals, and dividing by the residual degrees of freedom. These measures are presented in **Table 3**, and indicate that over-dispersion was not an issue for either model.

Measures of Model Complexity and Over-Dispersion				
Model #	AIC	BIC	$\widetilde{\phi}$	$\hat{\phi}$
2	534.68	559.92	1.013	0.672
4	543.66	568.91	0.86	0.55

Table 3: AIC and BIC were calculated for all possible models. Models 2 and 4 yielded the lowest values, corroborating their initial selection. When compared to one another, Model 2 was found to be superior given its lower scores. Estimates of ϕ are also provided to determine the presence of over-dispersion, which proved to not be a concern for either model.

Model Comparisons and Selection

There are two common statistics used when comparing models, Bayesian Information Criterion (AIC) and Akaike Information Criterion (AIC). Both were calculated for each possible combination of parameters in the model using the "step" function in R (Venables Ripley, 2002). These measures evaluate how well the model fits the data while also imposing a penalty for the number of parameters in the model (i.e., model complexity). Smaller BIC and AIC values indicate a parsimonious model that fits the data well. The lowest BIC and AIC scores were found for both Model 2 and 4, reaffirming their previous selection based off of the amount of deviance they accounted for. When comparing these models, Model 2 yielded a smaller AIC and BIC (**Table 3**). Thus, it was selected for prediction analyses. The full Poisson regression model can be expressed as

$$ln(pc) = -2.75 + \sum_{k=0}^{1} 0.74(1\{sex_i = k\}) - 15.26(1\{fp_i = k\}) + 0.27(ill_i) + 0.13(ad_i) + 0.09(hs_i) \tag{12}$$

where $i = (1, 2, ..., 496)$ and

$$1\{fp_i = k\} = \begin{cases} 1 & \text{if } fp_i = k \\ 0 & \text{otherwise} \end{cases} \qquad 1\{sex_i = k\} = \begin{cases} 1 & \text{if}_i = k \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Note that $fp_i = 1$ if covered by private health insurance without pharmacy coverage, 0 otherwise; $sex_i = 0$ if person is a male, and 1 if they are female. Interestingly, the Negative binomial model contained the same terms except for the covariate hs, reflecting the similarity in its components with the Poisson model. Predictions yielded by this model were also computed to determine the extent of its congruence with its Poisson counterpart.

4 Predictions

The intention of applying regression is being able to forecast future responses given a set of covariates. Thus, I determined the accuracy of Model 2 by predicting pc for the first individual in the dataset. Plugging their responses into the model yields the estimated prediction of $\hat{pc}$

$$0.134 = exp\{-2.75 + 0.74(1) - 15.26(0) + 0.27(0) + 0.13(0) + 0.09(0)\} \tag{14}$$

which is similar to their true value of $pc = 0$. Using the Poisson density formula (Equation 7) where $\hat{pc} = \mu_i$, I estimated the probability distribution for $y_i = 0, 1, 2, ..., V$ where V = number of pharmacist visits in a four week period (**Figure 4**). The same analysis was done for Model 4 as a final model comparison test. The prediction made by Model 2 yielded a smaller standard error and higher probability for the true value of pc, corroborating its selection.

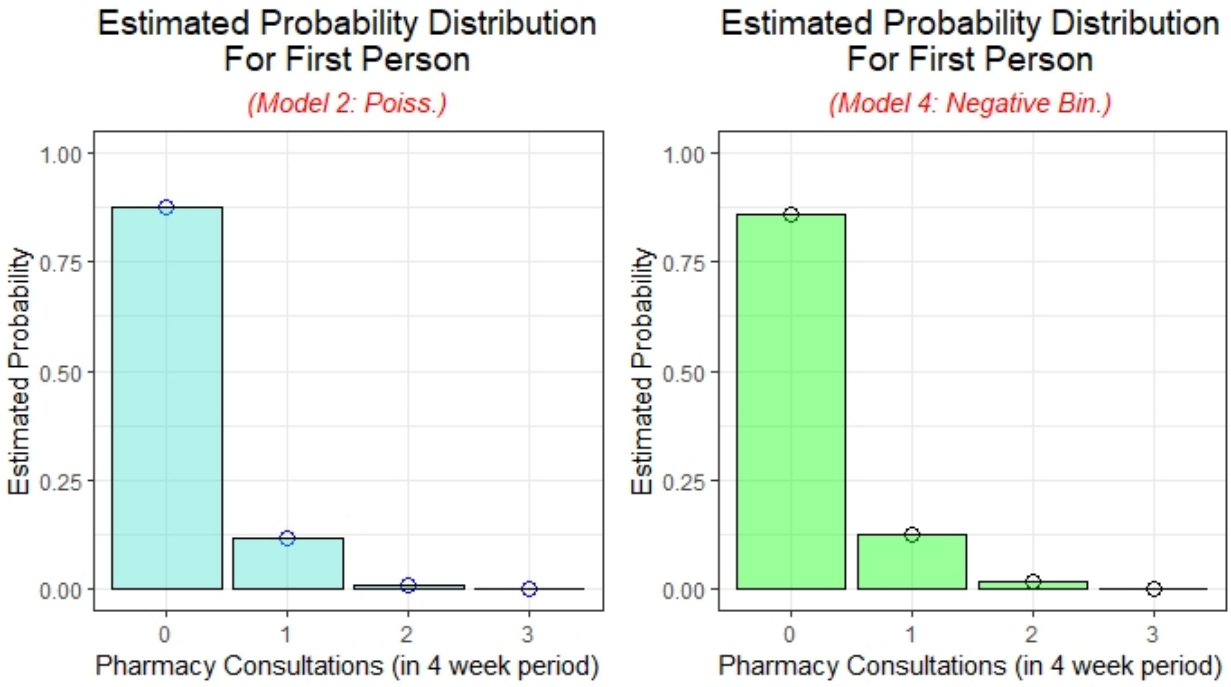

Figure 4: Probability distribution for the number of times an the first individual analyzed in the dataset visited the pharmacist. Note that the true vale of $pc = 0$. Both models accurately estimated $\hat{pc}$, with Model 2 performing slightly better.

Discussion

5 Interpretation

When compared to a null counterpart, a poisson regression model including the covariates ***sex***, ***fp***, ***ill***, ***ad***, and ***hs*** (Model 2) significantly yielded less residual deviance when accounting for the variation within ***pc*** ($D_M(490) = 329.33,\ p < 0.001$), prompting the rejection of the null hypothesis. Covariates were contrasted with one another using a Wald Chi-squared Test, which indicated that each were significantly different from one another ($p < 0.01$). When holding all other covariates constant: on average females visited their pharmacist about two times more than males ($\beta_s ex = 0.74, p < 0.001$), those who were covered with health insurance without pharmacy coverage were less likely to consult their pharmacist ($\beta_f p = -15.26, p > 0.05$), a one unit change in the number of illnesses in the past 4 weeks corresponded to an increase of 2 consultations ($\beta_i ll = 0.27, p < 0.001$), a one unit change in the number of self-reported days of reduced activity due to illness or injury resulted in an increase of 1 consultation ($\beta_a d = 0.13, p < 0.001$, while a change in general health questionnaire score corresponded to a an increase of 1 consultation ($\beta_h s = 0.09, p < 0.001$). The lack of *age* having an affect on pharmacy consultations was surprising, but this may be due to the greater number younger adults (20-30 years old) who completed the survey relative to the number of older adults (50-70 years old). If we split the data or added a variable for age group, then likely being an older individual would have significantly predicted the number of pharmacy consultations.

5.1 Issues

It is arguable that the Negative Binomial regression model (Model 4) was a better candidate for selection. The distribution of *pc* had a larger variance than mean, suggesting that over-dispersion was an issue. Despite this, estimates of ϕ for the Poisson model was close to its assumed value of 1, indicating that the reported results are not confounded by this issue. In addition, similar covariates were found to be significant in both models, suggesting that they may be comparable. The combination of these covariates yielded a smaller D_M for the Negative Binomial model relative to Poisson model. Although this provides evidence that the former is superior to the latter, R^2 was less for the Negative Binomial model. Further, prediction accuracy, AIC, and BIC were all found to be better for the Poisson model. Thus, it was deemed to be the preferred model. With more data, I would like to do a k-fold cross validation analysis to determine which model displayed greater generalization and accuracy. In regards to the collected data, representation seems to be a potential issue. Very few

individuals actually reported consulting with their pharmacist more that once. It may have been more fruitful to simply ask them if they visit their pharmacist or not, and then model the data using a binomial regression to predict the probability of consultation. As aforementioned, there was a bimodal distribution in the *age* of individuals who completed the survey. Therefore, it is likely that the latent variable of age group (e.g., young vs old) is masking the relationship between *age* and *pc*. Lastly, I am curious as to how this data was collected. If the survey was completed online, then there may be the representational issue of sampling only individuals with internet.

6 Concluding Remarks

All models have flaws, including the one proposed here. For instance, there was still a large amount of residual deviance that remains unaccounted for. Despite this, thorough diagnose of model performance suggested that Model 2 was the best candidate for predicting the number of times an individual visits their pharmacist in a 4 week period. Replication of these analyses and cross validation test are recommend to test the robustness of this model.

The cited R packages below were used to conduct the analyses above. Plots were created using the core "plot" function and the package ggplot.

References

[1] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

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

[3] Barret Schloerke, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Joseph Larmarange (2018). GGally: Extension to 'ggplot2'. R package version 1.4.0. https://CRAN.R-project.org/package=GGally

[4] Venables, W. N. Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[5] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr

[6] Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.2.5. https://CRAN.R-project.org/package=ggpubr

[7] H. Wickham. Reshaping data with the reshape package. Journal of Statistical Software, 21(12), 2007.

[8] Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2020). Hmisc: Harrell Miscellaneous. R package version 4.3-1. https://CRAN.R-project.org/package=Hmisc

[9] John Fox and Sanford Weisberg (2019). An R Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/