Predicting number of doctor visits with a generalized linear model

[redacted]

1 Introduction

The data consist of the number of visits to a family doctor for 1204 individuals. Ten demographic and health insurance related characteristics about the individuals are provided, of which three are numerical, e.g. income; and six are binary, e.g., public or private insurance. We investigate how the number of doctor visits depends on these characteristics under a generalized linear model.

2 Data Exploration

The six binary variables are: sex, marriage status (single or married), employment status (employed or unemployed), health insurance type (private or public), whether or not the individual has additional health insurance, and lastly whether or not the individual is living in a household with kids (under the age of sixteen). The three numerical variables are: age, net monthly household income (measured in thousands of German Marks), and years of schooling. There is also an ordinal variable that provides the highest degree of schooling obtained, such that zero corresponds to 'none', one corresponds to 'high school', and five corresponds to 'university'.

The number of individuals within each category of the six binary variables is shown in Figure 1. It is apparent that very few individuals have additional insurance, which may be important when fitting a model. The distributions of individuals across the ordinal and numerical variables are revealed to be positively skewed in Figure 2. This skewness has the potential to engender models that are overly weighted by a small number of observations within the upper range.

The fifteen possible pairs of the binary variables have a couple of moderate associations. A greater proportion of females were unemployed (47%) than males (14%), (phi coefficient of 0.37), and a greater proportion of married individuals lived in a household with kids (48%) than single individuals (14%), (phi coefficient of 0.31). It is also worth noting that only individuals with public health insurance had additional insurance. A couple of weak associations were found between the binary and numerical variables: older individuals tended to live in a household with kids (point biserial correlation of 0.30), and individuals with more years of education tended to have public health insurance (point biserial correlation of 0.32). Lastly, a weak positive association was found between household income and years of education (Pearson correlation coefficient of 0.25).

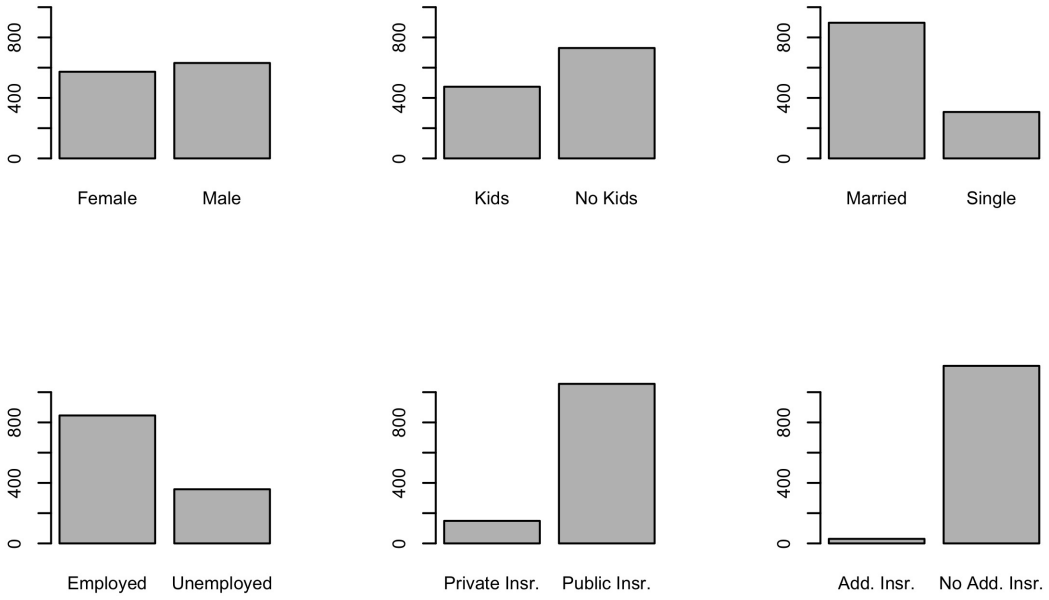

Figure 1: Barplots of the number of individuals across the binary variables.

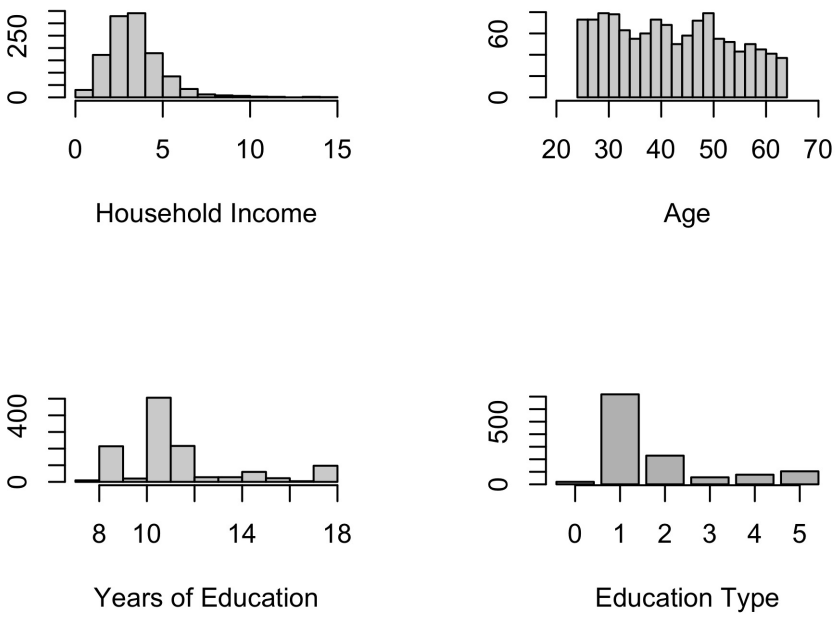

Figure 2: Histograms of the number of individuals across the ordinal (education type) and numerical variables.

While these aforementioned associations do not raise concerns of multicollinearity, a near linear dependence between education type and years of education is revealed in Figure 3. We might therefore wish to remove one of these education variables. The education type variable has some clear issues. Firstly, the precise definitions of education types one through four are not provided. Secondly, the wide variation in years of education for the same education type suggests that individuals may have interpreted the types differently. For example, some may have interpreted them as having *completed* that level of education, whereas others may have interpreted them as merely having *attended* that level of education. These ambiguities could impair the interpretability of our model, therefore we choose to exclude education type from the analysis.

The observed years of education were largely integers or multiples of 0.5. However, six individuals, all of whom had education type 0, had puzzling non-integer values, e.g., 10.80549. As these observations may have been errors, they can potentially be excluded if they later are found to be outliers that highly influence the fit.

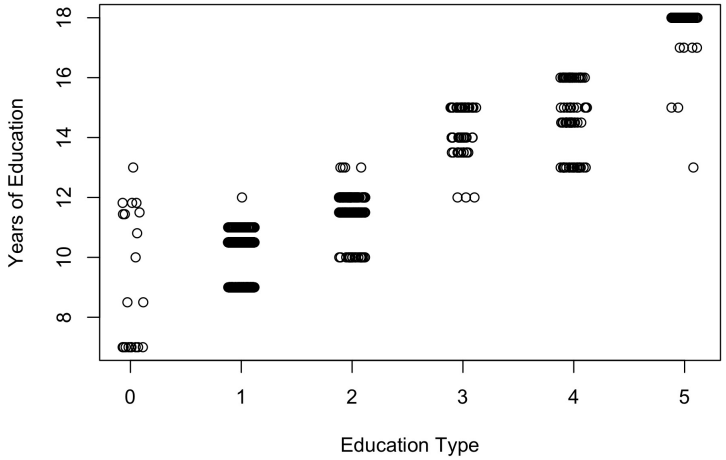

Figure 3: Scatter plot of education type against years of education, shown with random horizontal 'jitter' in education type.

We would like to observe the overall association the variables have with the number of doctor visits. For each category of the six binary variables, the distributions of the number of doctor visits are shown in Figure 4. Most prominently, the distribution of doctor visits is clearly shifted to the left for males as compared to females, such that males are far more likely to have zero doctor visits. In contrast, there does not appear to be any clear association between doctor visits and marriage status. It is also clear from Figure 4 that the most frequent number of doctor visits was zero, with a general trend of decreasing frequency as the number of doctor visits increases.

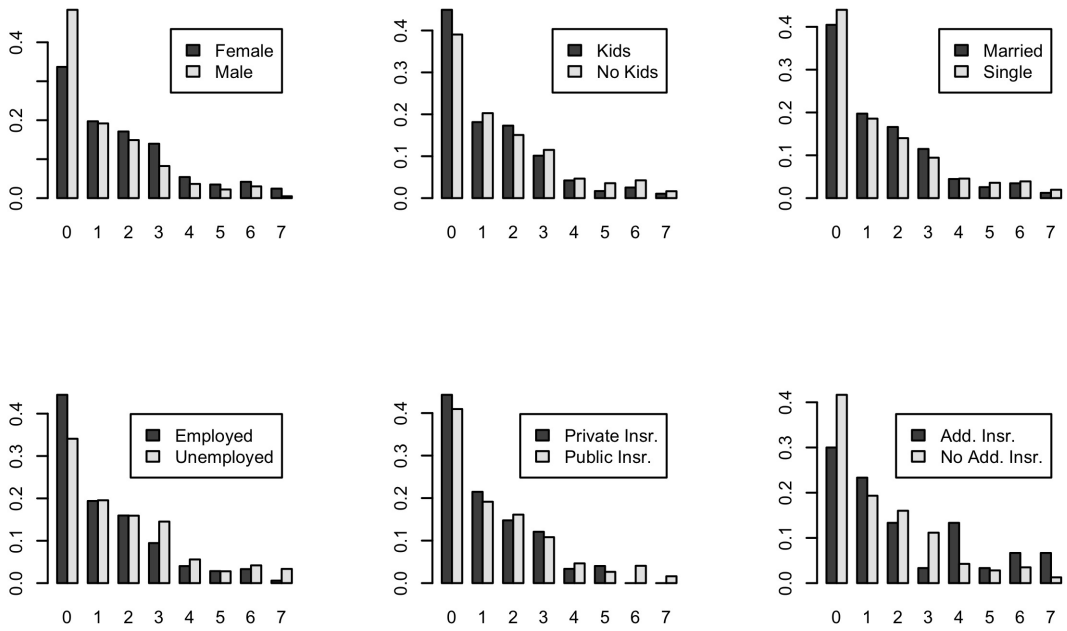

Figure 4: Barplots of the distribution of doctor visits across the six binary variables (shown as the proportion of individuals within each binary category).

For the three numerical variables, boxplots against the number of doctor visits are shown in Figure 5. Age appears to be positively associated with doctor visits. The median age for individuals with 5 – 7 doctor visits (49) is greater than the median age for individuals with 0 – 2 doctor visits (41). Household income is less clearly associated with doctor visits, although a decreasing trend is observed in the outliers, and the median household income for individuals with 0 – 2 doctor visits (3.4 thousand German Marks) is slightly greater than the median household income for individuals with 5 – 7 doctor visits (3 thousand German Marks). The association between years of education and doctor visits is similarly unclear, although there is a noticeable decreasing trend in the first and third quartiles.

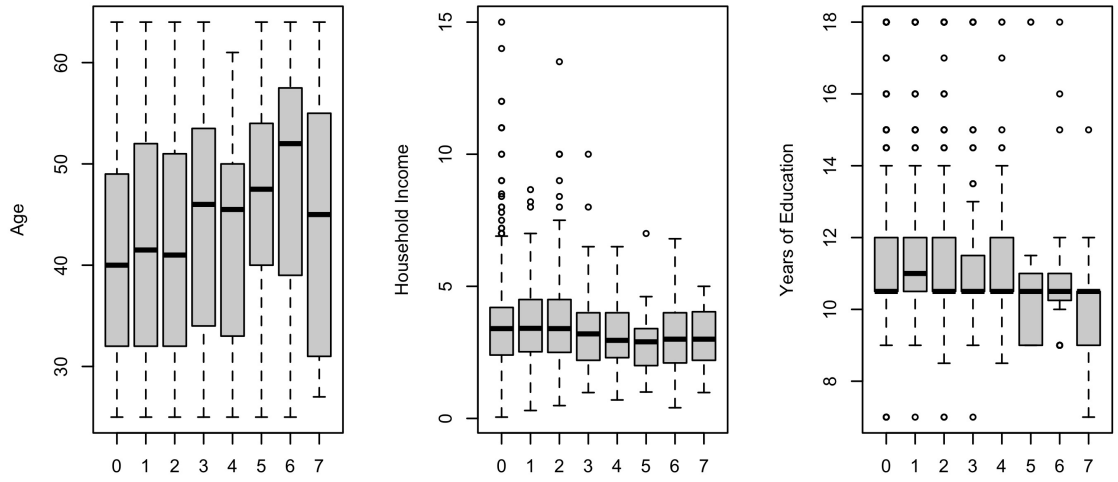

Figure 5: Boxplots of the number of doctor visits against the three numerical variables.

3 Modeling

3.1 Model selection

The number of doctor visits is a count of events, presumably within a fixed time interval. Therefore, the Poisson distribution is a reasonable choice for the generalized linear model. A natural way to link the linear combination of predictor variables η_i and the mean rate of occurrence for each observation μ_i is the canonical link function. Thus, our model has the following form:

$$Y_i \sim \text{Poisson}(\mu_i)$$

$$\mu_i = \exp(\eta_i)$$

In order to determine the linear predictor, we need to decide which terms to include, and of those terms, which could be combined into interactions. It is not practically possible to iterate over all possible models and choose the one with the highest quality. For illustration, there are $\binom{9}{2} = 36$ possible pairs of two-way interactions of the nine predictor variables. With a reasonable limit of, say, 5 two-way interactions, there are then at least $\binom{36}{5} > 350,000$ possibilities, an infeasible number of computations. An efficient alternative is to use a search algorithm, which uses some metric by which to judge the 'quality' of a model.

One reasonable metric is Akaike information criterion (AIC), which corresponds to the number of predictor variables minus the log likelihood (both scaled by 2). Minimizing AIC results in simultaneously maximizing both model fit and simplicity, the latter of which both promotes interpretability and prevents overfitting. The R function 'stepAIC', described in Zhang, 2016, provides an AIC search algorithm that carries out forward selection and backwards elimination, wherein terms are continuously added or removed, respectively, to obtain the largest possible reduction in AIC. Stepwise selection combines both forward and backward steps to find the model with the lowest AIC.

An initial model, a lower bound model, and an upper bound model are needed to run stepwise selection. The linear combination of all explanatory variables (wherein binary variables are treated as indicators) is chosen for the initial model. To allow for the removal of some or all of these terms, the intercept is chosen for the lower bound model. When three-way interactions are included in the upper bound model, the model is frequently overfit, such that certain observations are linearly separated by the predictor variables, causing a few of the estimated coefficients to diverge. Therefore, only two-way interactions are considered for the upper bound model. Because there are only thirty individuals with additional health insurance, two-way interaction terms involving this variable are excluded to prevent overfitting. The hierarchy principle is observed throughout the search algorithm.

The stepwise AIC algorithm obtains a linear combination of the following terms, which corresponded to an AIC of 4310 :

Intercept + Age + Income + Yrs. of Edu. + Sex + No Kids + Unemployed + Public Insr. + No Add. Insr. + Yrs. of Edu.:Public Insr. + Age:Male

Most notably, marriage status is excluded and there are two interactions, resulting in $p = 11$ terms in the linear predictor. It may be worth remarking that using education type in place of years of education for the stepwise AIC resulted in a marginally lower AIC (4308), though at the expense of interpretability for the reasons described earlier.

3.2 Model assessment and outlier analysis

A likelihood ratio test can be run to check whether there is no relation between the mean μ_i and the explanatory variables in the model. This test is given by the following equation where $D^{(0)}(y)$ is the deviance of the null (intercept-only) model (2553) and $D(y)$ is the deviance of the obtained model (2407):

$$D^{(0)}(y) - D(y) \sim \chi^2(p-1)$$

Under this test, the null hypothesis that all of the coefficients are equal to zero was rejected with a p-value of 0.

Because the number of doctor visits are relatively small (maximum of seven), the goodness of fit comparing the model to the saturated model is not suitable as the parameter estimates in the saturated model cannot converge normal distributions. An alternative goodness of fit is the deviance-based R^2, described in Cameron and Windmeijer (1997):

$$R^2_D = 1 - \frac{D(y)}{D^{(0)}(y)}$$

The model obtained an $R^2 = 0.0570$, suggesting a low degree of explanatory power.

Because the Poisson distribution does not approach normality for small count sizes, the residuals are not approximately normal. Nonetheless, the residuals are plotted in Figure 6 against the linear predictor to check for misfitting cases or patterns that might reveal problems with the model. Eight bands of residuals are observed, which correspond to the eight possible numbers of doctor visits. The bands are downward sloping, which just means that for the same number of doctor visits, the residual decreases as the linear predictor increases, as expected.

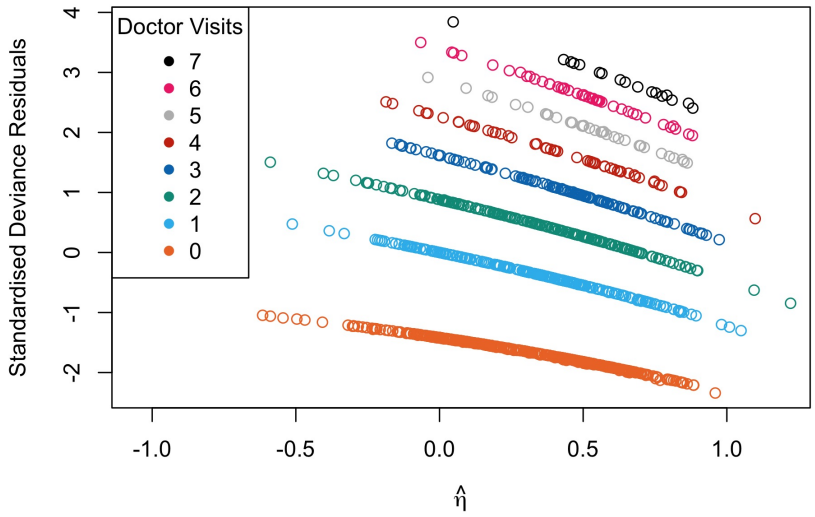

Figure 6: The linear predictor plotted against the residuals, coloured by the number of doctor visits.

The outliers can be identified with Cook's distance, which measures the influence data points have on the linear fit. Points with influence greater than the threshold $\frac{8}{n-2p}$ can be considered to have 'high' influence. Around 5.0% of observations had influence above this threshold. Similarly, around 8.3% of observations had 'high' leverage, which can be considered as greater than twice the mean leverage $\frac{2p}{n}$. Three observations had much larger influence than the other influential observations and are coloured in Figure 7.

It was discussed earlier how the positive skewness of the numerical variables could cause the predictions for the numerical variables to be highly influenced by a small number individuals in the high-end of these variables. It is therefore worth reporting that none of three coloured outliers had very large numerical values, with the exception that one had an age of 62. Still, the three variables all had additional health insurance, indicating that they might be strongly influencing the effect of having additional health insurance. Although a far lower AIC is obtained with these outliers removed (4274), there is no reason to doubt the validity of these observations, thus they shall remain in the model. Lastly, none of the observations with puzzling non-integer values for years of education had high influence or leverage.

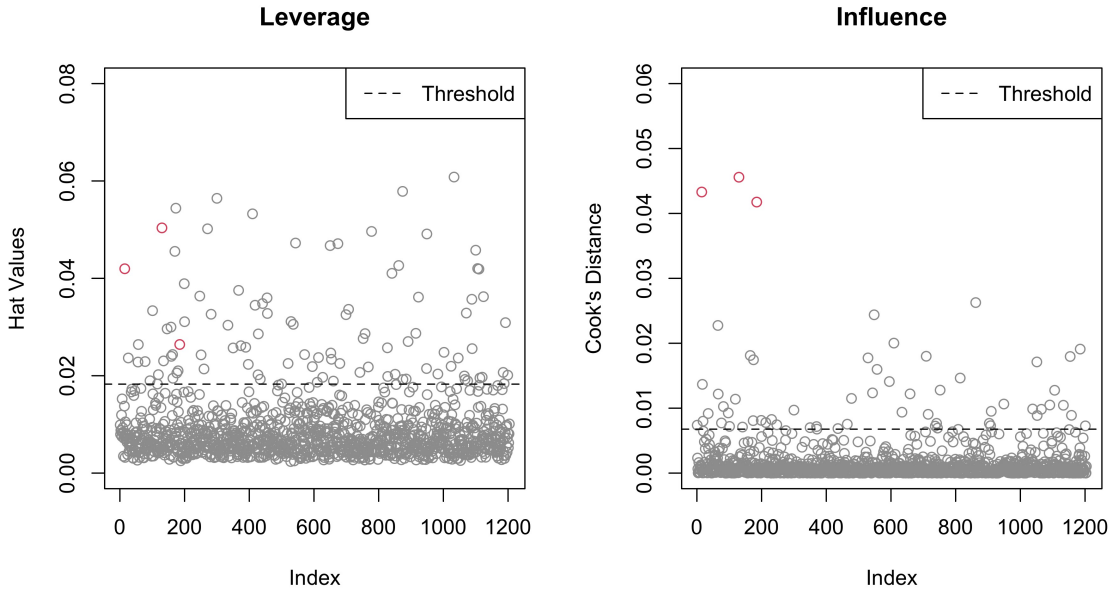

Figure 7: The leverage and influence of the observations. The threshold for leverage is given by $\frac{2p}{n}$ and the threshold for influence is given by $\frac{8}{n-2p}$. Three observations with relatively high influence are coloured red.

3.3 Model interpretation

We would like to interpret how the generalized linear model makes predictions. To do this, we turn to the coefficients, which are shown in Table 1. The intercept provides the predicted number of doctor visits for a hypothetical female who is employed, living in a household with kids, has private health insurance, and has additional insurance. Furthermore, this individual has the average age, income, and years of education in the dataset (as the numerical variables were centred to their means). The multiplicative factors are given relative to this baseline, and are multiplied together to obtain a prediction.

The binary multiplicative factors are the ratio in predictions between the two categories. For example, the marginal effect of an individual being male rather than female, all else equal, is that the Poisson expectation is multiplied by a factor of 0.728 (or equivalently divided by 0.728 for being female rather than male). The numerical multiplicative factors are the ratio in predictions when the numerical variable increases by one. For example, the marginal effect of an individual having one additional year of education is an increase in the expectation by a factor of 1.029. For the interactions between numerical and binary variables, the multiplicative factor is the additional marginal effect of the numerical variable increasing by one, when the binary category is present. For instance, when an individual has public insurance, an additional year of education means the prediction is multiplied by 1.029 and 0.933, which will be demonstrated shortly.

The 95% confidence intervals for the multiplicative factors are obtained from the asymptotic normality of the coefficient estimates (Wald test). The significance codes are provided by the likelihood ratio tests in the analysis of deviance, which depends on the order of the variables.

Table 1: The coefficients of the model given as multiplicative factors with 95% confidence intervals. Significance codes for the sequential likelihood ratio tests are given as follows: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Type	Variable	Mult. Factor	Conf. Int.	Signif. Code
Intercept	-	1.968	(1.417, 2.732)	-
Numer.	Age	1.003	(0.996, 1.009)	***
	Income	0.940	(0.911, 0.971)	***
	Years of Education	1.029	(0.985, 1.076)	**
Bin.	Male	0.728	(0.657, 0.806)	***
	No Kids	1.147	(1.032, 1.276)	*
	Unemployed	1.088	(0.974, 1.215)	
	Public Insr.	1.119	(0.930, 1.347)	
	No Add. Insr.	0.674	(0.525, 0.866)	**
Numer. & Bin.	Yrs. of Edu. & Public Insr.	0.933	(0.887, 0.981)	**
	Age & Male	1.009	(1.001, 1.017)	*

The multiplicative factors confirm our earlier observations in the data exploration, e.g., that age is positively associated with doctor visits. At first glance, it might appear that the model has years of education positively associated with the numbers of doctor visits, though this multiplicative factor is for individuals with private insurance. The vast majority of individuals have public insurance, for whom multiplying by the interaction ($1.029 \times 0.933 = 0.96$) results in decreasing predicted doctor visits with increasing years of education.

Some predictions of these coefficients can be visualized in Figure 8. The left plot shows the predicted number of doctor visits as a function of household income, for which the predictions are slightly greater (by a factor of 1.119) for those with public insurance. The right plot shows a strong correspondence with the observed data, as the number of doctor visits for females living in a household without kids is predicted to be ($1.147 \div 0.728 = 1.58$) times greater than for males living in a household with kids, all else equal.

3.4 Estimating the dispersion parameter

The dispersion parameter ϕ affects the standard errors of the coefficients in the linear predictor. Under the Poisson model, it is assumed to be one, though the actual value can be estimated by the following equation involving the variance function $V(\mu)$:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

If the estimation $\hat{\phi}$ is greater than 1, then the estimates for the standard errors of the coefficients in the linear predictor must be adjusted to be a factor of $\hat{\phi}^{1/2}$ greater. In consequence, the confidence intervals widen and the coefficients may lose significance under the Wald test.

In our case, $\hat{\phi} = 1.92$ was obtained. After adjustment, three terms in the linear predictor (the effect of living in a household with kids and the two interaction terms) lost significance at the 5% level under the Wald test, which for Table 1 means their adjusted confidence intervals included 1. After this adjustment, only the intercept, household income, sex, and additional insurance were significant under the Wald test.

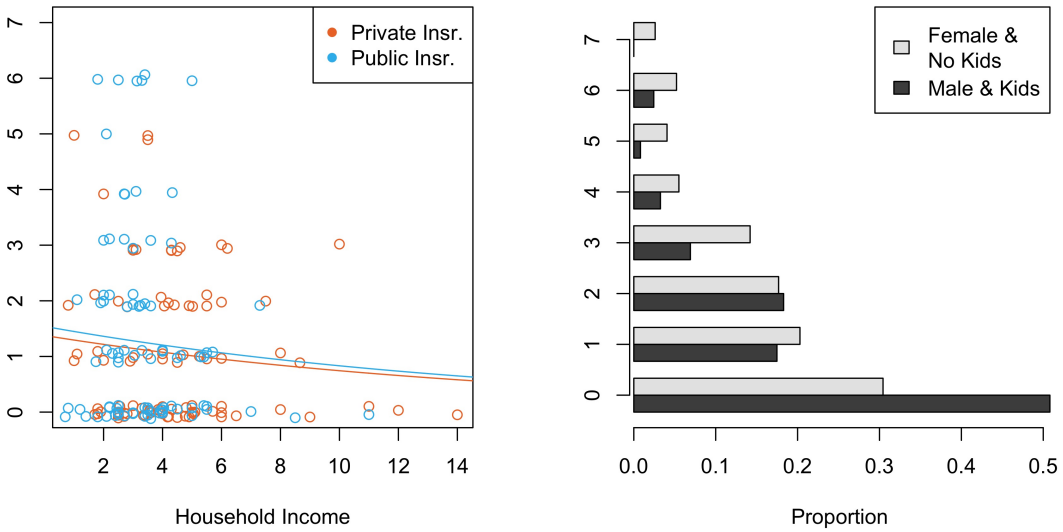

Figure 8: Visualizations of model coefficients. The left plot shows predictions for number of doctor visits against household income, separated by type of health insurance for ninety randomly selected males. (The predictions assume the mean value for the other two numerical variables and the majority category for each binary variable.) In the right plot, the differing distributions of doctor visits for females who do not live with kids versus males who live with kids.

4 Conclusions

A generalized linear model was fit between the number of doctor visits made by an individual and characteristics of the individual. For improved interpretability, years of education were used in place of education type. The model was obtained by a stepwise AIC search algorithm. The obtained model did not include marriage status and contained two interaction terms. The model had a low degree of predictive power given by a deviance based $R^2 = 0.057$. The mean absolute error of the model was 1.34, which was slightly lower than that of the null model 1.405. The model contained many outliers and had an estimated dispersion parameter larger than one, such that three predictor variables lost significance by the Wald test after adjustment.

Perhaps more important than the performance is the model's estimates of the effects of the explanatory variables. A majority of the model's predictor variables were significant by likelihood ratio tests. The most robust findings of the model were that males were predicted to have fewer doctor visits and that greater household incomes were associated with fewer doctor visits. These effects remained significant even after the dispersion parameter adjustment. The strength of these associations warrant further investigation into the causal factors behind them.

Appendix

```
# R code


# Data Exploration


## Load Data

d1 <- read.csv('dvis.csv')
d2 <- d1[c("docvis", "age", "hhninc", "educyrs", "eductype")]


### Binary variable renaming and make factors

d2$sex <- as.factor(sapply(d1$female, function(x) {
  if (x == 1) {'Female'} else {'Male'}}))

d2$kids <- as.factor(sapply(d1$hhkids, function(x) {
  if (x == 1) {'Kids'} else {'No Kids'}}))

d2$married <- as.factor(sapply(d1$married, function(x) {
  if (x == 1) {'Married'} else {'Single'}}))

d2$employed <- as.factor(sapply(d1$employed, function(x) {
  if (x == 1) {'Employed'} else {'Unemployed'}}))

d2$insurance <- as.factor(sapply(d1$privateins, function(x) {
  if (x == 1) {'Private Insr.'} else {'Public Insr.'}}))

d2$addins <- as.factor(sapply(d1$addins, function(x) {
  if (x == 1) {'Add. Insr.'} else {'No Add. Insr.'}}))


## Save edited data

write.csv(d2, 'd2.csv')


## Strange numbers in years of education

strange_nums <- d1$educyrs %% 0.5 != 0

options(digits=5)
d1[strange_nums, ]$educyrs

sn2 <- (d2$eductype == 0) * (d2$educyrs >= 10)
```

```
sn2 <- as.logical(sn2)

d2t <- d2[!sn2, ]


## Numerical frequency box plots

jpeg("plots/numerical_hist.jpg", width = 5, height = 4, units = "in", res = 400)
par(mfrow = c(2, 2))
hist(d2$hhninc, breaks=15,
     main=NULL, ylab=NULL, xlab='Household Income')

hist(d2$age, breaks=15, xlim=c(20, 70),
     main=NULL, ylab=NULL, xlab='Age')

hist(d2$educyrs, breaks=15,
     main=NULL, ylab=NULL, xlab='Years of Education')

barplot(table(d2$eductype), xlab='Education Type', axis.lty=1)
dev.off()


## Education correlation

jpeg("plots/education_correlation.jpg", width = 6.4, height = 4.8, units = "in", res = 300)
plot(educyrs ~ jitter(eductype, 0.6),
     xlab='Education Type', ylab='Years of Education',
     data = d2)
dev.off()


## Numerical relation

jpeg("plots/numerical_boxplot_relation.jpg", width = 7, height = 3.8,
units = "in", res = 400)
par(mfrow=c(1, 3))
boxplot(age ~ docvis, data=d2, xlab=NULL, ylab='Age')

boxplot(hhninc ~ docvis, data=d2, xlab=NULL, ylab='Household Income')

boxplot(educyrs ~ docvis, data=d2, xlab=NULL, ylab='Years of Education')
dev.off()


## Binary frequency box plots

binary_cols <- c("sex", "kids", "married", "employed", "insurance", "addins")

jpeg("plots/binary_boxplots.jpg", width = 6, height = 4, units = "in", res = 400)
```

```
par(mfrow = c(2, 3), cex.axis=.85)
for (col in binary_cols) {
  tab <- table(d2[col])
  print(tab / sum(tab))
  barplot(tab, beside = TRUE, ylim=c(0, 1000))
}
dev.off()


## Binary relation

jpeg("plots/binary_relation_hist.jpg", width = 6, height = 4, units = "in", res = 400)
par(mfrow = c(2, 3), cex=0.55)
for (col in binary_cols) {
  tab <- table(d2[, col], d2[, 'docvis'])
  print(tab <- tab / rowSums(tab))
  barplot(tab, beside=TRUE, legend=row.names(tab))
}
dev.off()

tab / rowSums(tab)


## Binary two-way frequencies (verbally summarize)

library(psych)
factor_combos <- combn(binary_cols, 2)

options(digits=3)
for (i in 1:dim(factor_combos)[2]){
  combo <- factor_combos[, i]
  tab <- table(d2[, combo[1]], d2[, combo[2]])
  print(tab / rowSums(tab))
  print(abs(phi(tab)))
}


## Point biserial correlations

library(ltm)

numer_cols <- c("educyrs", "hhninc", "age")
for (col in binary_cols) {
  for (col2 in numer_cols) {
    print(paste(col, col2))
    print(biserial.cor(d2[, col2], d2[, col]))
  }
}
```

```
## Pearson correlations

cor(d2[, numer_cols])

cor(d2[, c("educyrs", "eductype")])


## Mean differences for binary

for (col in binary_cols){
  for (val in unique(d2[, col])){
    print(val)
    print(mean(d2[d2[, col] == val, ]$docvis))
  }
  print("")
}


## Doctor visit summaries

barplot(table(d2$docvis))
mean(d2$docvis)
median(d2$docvis)
sd(d2$docvis)


## Median values for the boxplots

median(d2[d2$docvis < 3, ]$hhninc)
median(d2[d2$docvis > 4, ]$hhninc)

median(d2[d2$docvis < 3, ]$age)
median(d2[d2$docvis > 4, ]$age)



# Model Selection


## Base models for stepAIC

base1 <- glm(docvis ~ (age + hhninc + educyrs + sex + kids +
                   married + employed + insurance + addins), data=d2, family=poisson)

base1t <- glm(docvis ~ (age + hhninc + eductype + sex + kids +
                   married + employed + insurance + addins), data=d2, family=poisson)

base2 <- glm(docvis ~ (age + hhninc + educyrs + sex + kids +
```

```
                  married + employed + insurance)^2 + addins, data=d2, family=poisson)


## Start from two-way terms

library(MASS)

search2 <- stepAIC(base2,
             scope = list(
                upper = ~ (sex + age + hhninc + kids + educyrs +
                             married + employed + insurance + addins) +
                  (sex + age + hhninc + kids + educyrs +
                             married + employed + insurance)^2,
                lower = ~ 1),
             trace=FALSE, steps=10000, direction='both')

search2$aic


## Start from singular terms (selected)

search1 <- stepAIC(base1,
             scope = list(
                upper = ~ (age + hhninc + educyrs + sex + kids +
                             married + employed + insurance + addins) +
                  (age + hhninc + educyrs + sex + kids +
                             married + employed + insurance)^2,
                lower = ~ 1),
             trace=TRUE, steps=10000, direction='both')

summary(search1)


## Using education type

search1t <- stepAIC(base1t,
             scope = list(
                upper = ~ (age + hhninc + eductype + sex + kids +
                             married + employed + insurance + addins) +
                  (age + hhninc + eductype + sex + kids +
                             married + employed + insurance)^2,
                lower = ~ 1),
             trace=FALSE, steps=10000, direction='both')

summary(search1t)$aic


## Set colors
```

```
cblue <- "#0077BB"
ccyan <- "#33BBEE"
cteal <- "#009988"
corange <- "#EE7733"
cred <- "#CC3311"
cmagenta <- "#EE3377"
cgrey <- "#BBBBBB"
cblack <- "#000000"
ccol <- c(corange, ccyan, cteal, cblue, cred, cgrey, cmagenta, cblack)



# Model Analysis


## Define model

m <- glm(formula = docvis ~ age + hhninc + educyrs + sex + kids +
    employed + insurance + addins + educyrs:insurance + age:sex,
    family = poisson, data = d2)


## Run tests

summary(m)

anova(m, test="Chisq")


## Likelihood ratio test

(n <- dim(d2)[1])
(p <- m$rank)

(D0 <- m$null.deviance)
(D <- m$deviance)

(Lambda <- D0 - D)

# Statistical test of all parameters equal to zero
options(digits=22)
1 - pchisq(Lambda, p-1)


## Deviance-based R^2

(R2 <- 1 - D/D0)

library('rsq')
```

```
rsq.kl(m)


## Deviance residuals

eta_hat <- predict(m, type='link')

rd_stand <- rstandard(m)

var(rd_stand)


## Residuals versus linear predictors

num_visits <- as.character(c(0:7))

jpeg("plots/misfit.jpg", width = 6.4, height = 4.8, units = "in", res = 300)
par(mfrow=c(1, 1))
plot(eta_hat, rd_stand, col=ccol[d2$docvis+1], xlim=c(-1.05, 1.2),
     ylab='Standardised Deviance Residuals', xlab=expression(hat(eta)))
legend("topleft", rev(num_visits), title="Doctor Visits",
       col = rev(ccol), pch = 16, bg="transparent")
dev.off()


## Leverage

h_vals <- hatvalues(m)

# twice the mean leverage
lev_thres <- (2*p/ n)

lev_vec <- h_vals > lev_thres

sum(lev_vec) / n


## Influence

inf_thres <- 8 / (n - 2*p)

cooks_dist <- cooks.distance(m)

inf_vec <- cooks_dist > inf_thres

sum(inf_vec) / n


## Check the strange education years values
```

```
sum(inf_vec * sn2)

sum(lev_vec * sn2)


## Get the (three) extreme outliers

extreme_outliers <- cooks_dist > 0.03

options(digits=10)
d2[extreme_outliers, ]

# plot the misfit
plot(eta_hat, rd_stand, col=8-6*extreme_outliers,
     ylab='Standardised Deviance Residuals', xlab=expression(hat(eta)))


## AIC of model without these outliers

summary(glm(formula = docvis ~ age + hhninc + educyrs + sex + kids +
    employed + insurance + addins + educyrs:insurance + age:sex,
    family = poisson, data = d2[!extreme_outliers, ]))


## Plot leverage and influence

jpeg("plots/lev_inf.jpg", width = 1.1*8.3, height = 1.1*4.8, units = "in", res = 400)
par(mfrow=c(1, 2), pty='s')

plot(h_vals, col=8-6*extreme_outliers,
     main='Leverage', ylab = "Hat Values", ylim=c(0, 0.08))
legend("topright", c("Threshold"), lwd=1, lty = 2, bg="transparent")
abline(lev_thres, 0, lty = 2)


plot(cooks_dist, col=8-6*extreme_outliers,
     main = "Influence", ylab = "Cook's Distance", ylim=c(0, 0.06))
legend("topright", c("Threshold"), lwd=1, lty = 2)
abline(inf_thres, 0, lty = 2, col=1)
dev.off()


## Dispersion parameter

expect <- predict(m, type="response")

y <- d2$docvis
```

```
(phi_hat <- sum((y - expect)^2 / expect) / (n-p))

sqrt(phi_hat)



# Model Interpretation


## Center numerical values

d2i <- d2

d2i$hhninc <- d2i$hhninc - mean(d2i$hhninc)
d2i$age <- d2i$age - mean(d2i$age)
d2i$educyrs <- d2i$educyrs - mean(d2i$educyrs)

mi <- glm(formula = docvis ~ age + hhninc + educyrs + sex + kids +
    employed + insurance + addins + educyrs:insurance + age:sex,
    family = poisson, data = d2i)


## Interpret coefficients and adjust by the dispersion parameter

coefs <- data.frame(summary(mi)$coefficients)

colnames(coefs) <- c("B_est", 'std_err', 'z', 'p')

coefs$adj_std_err <- sqrt(phi_hat) * coefs$std_err

z_alpha <- qnorm(1 - 0.05/2)  # 1.96

coefs$adj_ci_upper <- coefs$B_est + z_alpha * coefs$adj_std_err
coefs$adj_ci_lower <- coefs$B_est - z_alpha * coefs$adj_std_err


coefs$ci_upper <- coefs$B_est + z_alpha * coefs$std_err
coefs$ci_lower <- coefs$B_est - z_alpha * coefs$std_err


## Get multiplicative factors (including adjusted )

coefs$e_B_est <- exp(coefs$B_est)

coefs$e_std_err <- exp(coefs$std_err)

coefs$e_adj_std_err <- exp(coefs$adj_std_err)

coefs$e_adj_ci_upper <- exp(coefs$adj_ci_upper)
```

```
coefs$e_adj_ci_lower <- exp(coefs$adj_ci_lower)


coefs$e_ci_upper <- exp(coefs$ci_upper)
coefs$e_ci_lower <- exp(coefs$ci_lower)


## Get adjusted p-values

coefs$adj_z <- coefs$B_est / coefs$adj_std_err
coefs$adj_p <- 2 * (1 - pnorm(abs(coefs$adj_z)))

## Save coefficients to disk (for making a table in excel)

write.csv(coefs, 'out-data/coefs.csv')


## Prepare plot of doctor visits by income and insurance

cosi <- exp(coefficients(m))

(intercept_k <- cosi[1])
(income_k <- cosi[3])
(public_k <- cosi[8])

(male_k <- cosi[5])
(kids_k <- cosi[6])

(age_k <- cosi[2]^mean(d2$age))
(educyrs_k <- cosi[4]^mean(d2$educyrs))

(addins_k <- cosi[9])

(cosi_pe <- cosi[10]^mean(d2$educyrs))
(cosi_as <- cosi[11]^mean(d2$age))

x_inc <- seq(0, 15, by=0.01)

private_eq <- intercept_k*(income_k^x_inc)
public_eq <- intercept_k*public_k*(income_k^x_inc)

private_eq <- private_eq*age_k*educyrs_k*addins_k*male_k*cosi_as*kids_k
public_eq <- public_eq*age_k*educyrs_k*addins_k*male_k*cosi_as*kids_k

public_eq <- public_eq*cosi_pe


## Prepare plot of Male and Kids
```

```
s1a <- subset(d2, sex == 'Male' & kids == 'Kids')
s1b <- subset(d2, sex != 'Male' & kids != 'Kids')


t1 <- table(s1a$docvis)
t2 <- table(s1b$docvis)

t1 <- c(t1, 0)

matrix1 <- matrix(c(t1, t2), nrow=2, byrow=TRUE)
matrix1 <- matrix1 / rowSums(matrix1)

colnames(matrix1) <- c(0:7)


## Plot visualizations of interpretation

d2i <- subset(d2, sex == "Male")

# jpeg("plots/interpret.jpg", width = 1.1*8.3, height = 1.1*4.8,
# units = "in", res = 400)
par(mfrow=c(1, 2), pty='s')
# set.seed(12)
# di1 <- sample_n(sample_n(d2[(d2$insurance == 'Private Insr.'), ], 100), 90)
# di2 <- sample_n(sample_n(d2[(d2$insurance == 'Public Insr.'), ], 100), 90)

set.seed(0)
di1 <- sample_n(d2i[(d2i$insurance == "Private Insr."), ], 90)
di2 <- sample_n(d2i[(d2i$insurance == "Public Insr."), ], 90)

plot(di1$hhninc, jitter(di1$docvis, 0.6), col=ccol[1],
     ylab="",
     xlab="Household Income",
     ylim=c(0, 7))
lines(x_inc, private_eq, col=ccol[1])

points(di2$hhninc, jitter(di2$docvis, 0.6), col=ccol[2])
lines(x_inc, public_eq, col=ccol[2])
legend("topright", c("Private Insr.", "Public Insr."),
       # title="Insurance Type",
   col = ccol[1:2], pch = 16, bg="transparent")

barplot(matrix1, beside=TRUE, axis.lty=1,
        legend=c('Male & Kids', 'Female &\nNo Kids'),
        args.legend = list(x='topright', bg="transparent"),
        xlab='Proportion',
        horiz=TRUE)
# dev.off()
```

```
## Multiplicative factor for discussion

mi_female <- 1 / exp(coefficients(mi))[5]
mi_nokids <- exp(coefficients(mi))[6]

mi_female * mi_nokids
# 1.58


## Mean absolute error

just_intercept <- glm(docvis ~ 1, data=d2, family=poisson)

mean(abs(residuals(just_intercept, type="response")))
mean(abs(residuals(m, type="response")))
```