Practical for Michaelmas Term Week 7

P576

December 9 2020

1 Exploratory Analysis

The data consist of information related to doctor visits for either an individual or a household in which the individual resides in. Upon a small exploration, it is seen that the data can be categorized into 7 categorical and 4 integer-valued variables. The frequency tables for the binary variables are below:

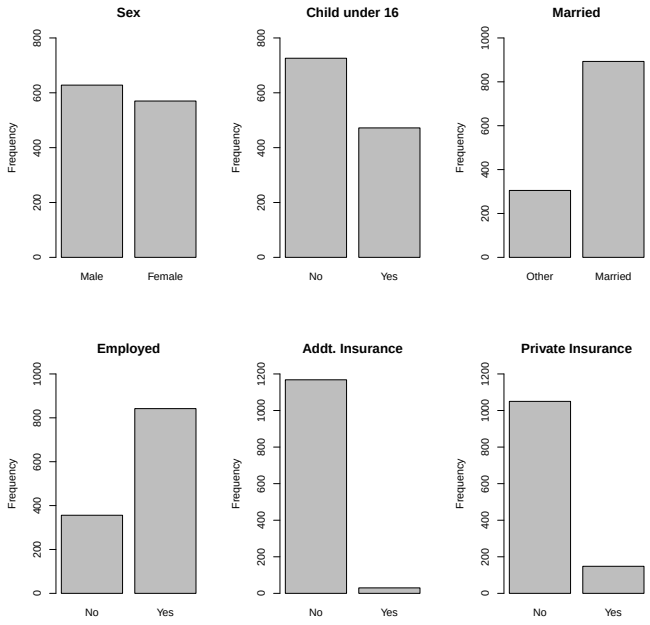

Figure 1: Binary Variables Frequency Tables

From Figure 1, we observe that there is a balanced distribution of sexes in the sample. However, when it comes to variables *Employed, Additional Insurance, Private Insurance* we see that these variables are not very balanced. We look into how employment status changes across sexes.We see that for females, the employment rate is 53% and for males the employment rate is 86%. So, we deduce that the imbalance in employment is caused by sex. Next, we look into the frequency tables for non-binary variables:

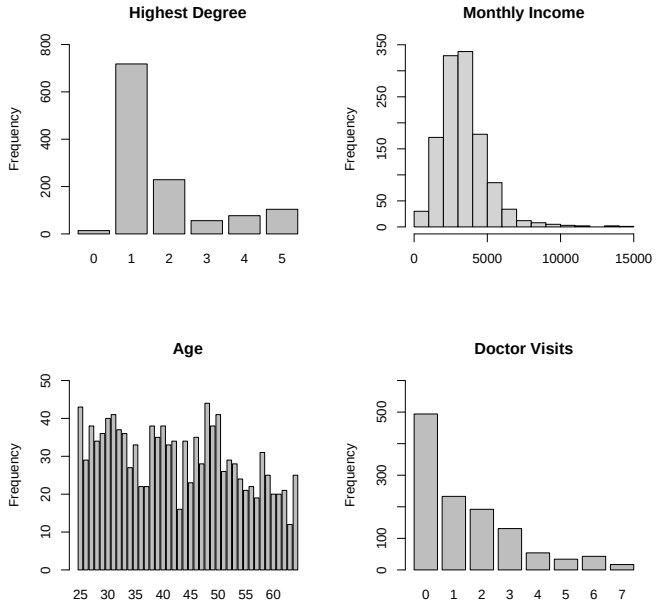

Figure 2: Non-binary Variables Frequency Tables

We observe that the most common degree attained is high school, the most common monthly income range is 1000-4000 German marks, and most common number of doctor visits is 0 in the sample. Age is distributed relatively equally across age groups. We see that there is slightly more young population than old. As you can notice, we dropped *educyrs* from the data. We reach this conclusion by first seeing that the correlation between *educyrs* and *eductype* is 94%. This would cause multicollinearity in our regression if we choose to include both *educyrs* and *eductype*. So, we look more into these variables to choose which one to drop. We first observe that there are entries with *educyrs* variable equalling 11.441, 10.805, and 11.818. These numbers are rather odd. We dig deeper into the relationship between *educyrs* and *eductype* by creating Figure 3. This figure plots the unique *educyrs, eductype* pairs.

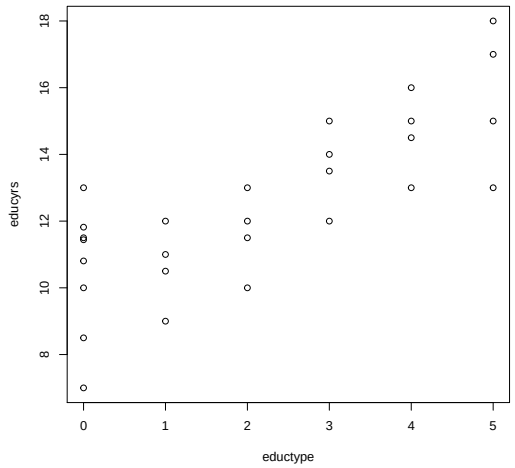

Figure 3: Years of education versus highest schooling attained

As can be seen from Figure 3, with 12 years of schooling, someone may have attained levels 0,1,2,3 and with 13 years of schooling, one may have attained 1,3,4,5. This is rather a large range. So, we decide that sticking to *eductype* is a good idea since these observations makes it obvious that *eductype* contains more granular information regarding an individual's quality of education.

We move on to explore the distribution of the number of doctor visits across different variables:

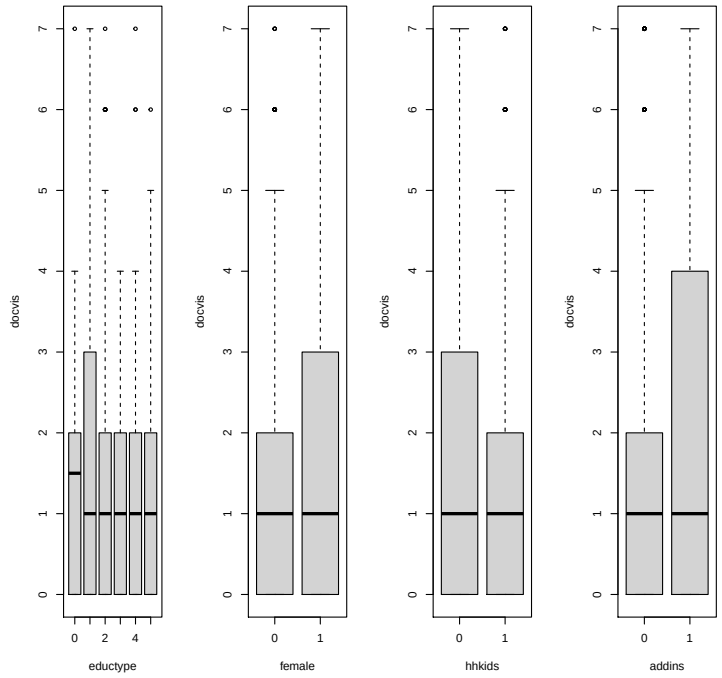

Figure 4: Doctor Visits against *eductype, female, hhkids,addins*

From Figure 4, we can deduce relationships between variables. For example, we see that there seems to be a negative relationship between number of doctor visits and education type, and children under 16. Also, there seems to be a positive relationship between number of doctor visits and being female, and add-in insurance. Since, we know from 1 that we have so few observations with additional insurance, we should be cautious about making any causal claims about the relationship between the number of doctor visits and additional insurance.

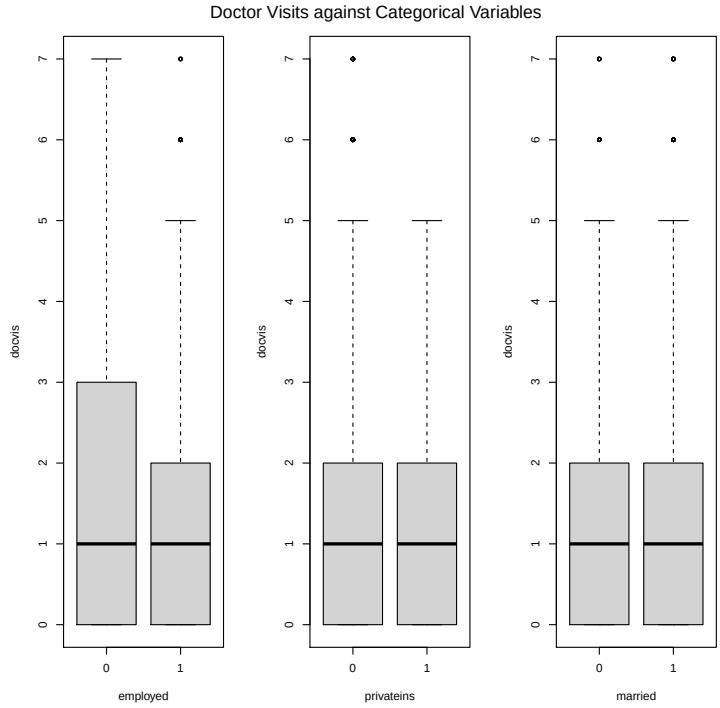

Figure 5: Doctor Visits against *employed, privateins, and married*

From figure 5, we see that there is a negative relationship between being employed and the number of doctor visits. We find it difficult to infer any relationship between number of doctor visits and private insurance and being married since the mean and the quartiles across categories for these two variables seem to fall in the same place.

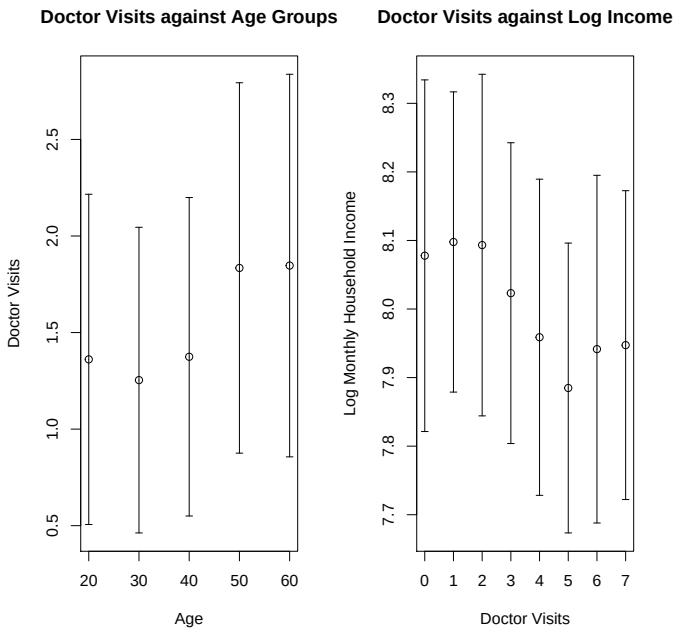

Figure 6: Doctor Visits against Numerical Variables

To obtain Figure 6, we aggregate doctor visits by age (in 10s), and log monthly income. Using the same aggregations, we obtain standard errors, and create intervals around the related means to have an idea about the standard deviation at each level. We use half standard deviation at each level to create the confidence intervals. From these plots, we can infer that there is a positive relationship between doctor visits and age. This is expected as people are more prone to ailments as they age. From the second plot, we can infer that there is a negative relationship between income and doctor visits. This can be rationalized since richer people tend to have safer jobs and are less prone to injuries and illnesses as a result e.g. a rich person may be a manager, and a poor person may be a construction worker.

2 Relation Modelling and Model Selection

We first create a baseline model by including all variables except *educyrs* and all of their two-way interactions. The R code to generate this model is as follows:

$$glm1 < -glm(docvis \sim (.)^2, data = dt, family = poisson) \tag{1}$$

To assess the quality of our models we use four key metrics: R^2, Akaike Information Criterion (AIC), interpretability, and complexity. This first model, Equation 1, has AIC of 4346, and R^2 of 10%. Since this equation has too many terms, we pursue term elimination by using AIC criterion via step function in R language. Since elimination via AIC criterion depends on the order of coefficients, we randomly shuffle the columns of the data 8 times and obtain the best model according to AIC criterion each time. As a result, we receive the following equation:

$$\begin{aligned} &glm(docvis \sim employed + hhkids + addins + hhninc + married + age + privateins + female + \\ &eductype + employed : hhkids + employed : addins + employed : married + employed : age + \\ &hhkids : hhninc + addins : hhninc + addins : age + hhninc : married + married : eductype + \\ &age : female + privateins : eductype, family = poisson, data = sample(dt)) \end{aligned} \tag{2}$$

The regression table corresponding to Equation 2, is given in Table 1. After AIC elimination, we inspect the variables and interactions terms. We suspect 3 of the terms may not be statistically significant. These are namely:

1. $married \times eductype$
2. $privateins \times eductype$
3. $hhkids \times hhninc$

We use Likelihood Ratio Test to assess whether these variables are significant. This test is given below:

$$\Delta = D^{(R)}(y) - D^{(P)}(y) \tag{3}$$

where $D^{(R)}$ corresponds to a model with dimension r and $D^{(P)}$ corresponds to a model with dimension p and $r < p$. We approximate this statistic by $\chi^2(p - r)$.

The p-values for these LRT statistics are respectively .02%, 14%, and 8%. As a result, we cannot reject that the interaction terms between privateins and eductype are not statistically significant at 10% confidence level. So, we drop these interaction terms. We choose to keep the interaction terms between *hhkids* and *hhninc* as we can reject the the respective null hypothesis under 10% confidence level.

The resulting model has an R^2 of 7.1% and an AIC of 4295 as can be seen in the second column of Table 1.

Table 1: Regression Results

	Dependent variable:	
	docvis	
	(1)	(2)
employed1	0.256	0.223
	(0.266)	(0.266)
age	0.017***	0.017***
	(0.005)	(0.005)
hhninc	−0.157*	−0.148*
	(0.088)	(0.088)
married1	1.113	1.126
	(0.981)	(0.975)
female1	0.685***	0.705***
	(0.205)	(0.204)
privateins1	0.432**	−0.036
	(0.186)	(0.084)
hhkids1	−1.541	−1.534
	(0.958)	(0.947)
addins1	−3.755	−4.492
	(2.842)	(2.792)
eductype1	−0.323	−0.332
	(0.364)	(0.364)
eductype2	−0.086	−0.095
	(0.370)	(0.370)
eductype3	−0.754*	−0.781*
	(0.457)	(0.452)
eductype4	−0.425	−0.445
	(0.385)	(0.384)
eductype5	−1.062**	−0.842**
	(0.412)	(0.402)
employed1:age	−0.009*	−0.009
	(0.005)	(0.005)
hhkids1:hhninc	0.196	0.194
	(0.120)	(0.118)
addins1:hhninc	0.788**	0.702**
	(0.344)	(0.349)
addins1:age	−0.031**	−0.031**
	(0.014)	(0.014)
employed1:married1	0.254*	0.255*
	(0.133)	(0.133)
employed1:hhkids1	−0.278**	−0.259**
	(0.131)	(0.130)
employed1:addins1	−0.529*	−0.580**
	(0.289)	(0.288)
age:female1	−0.008*	−0.009**
	(0.004)	(0.004)
hhninc:married1	−0.191	−0.193*
	(0.116)	(0.115)
married1:eductype1	0.307	0.301
	(0.444)	(0.444)
married1:eductype2	−0.020	−0.023
	(0.454)	(0.454)
married1:eductype3	0.559	0.552
	(0.539)	(0.539)
married1:eductype4	0.025	0.008
	(0.491)	(0.490)
married1:eductype5	0.889*	0.893*
	(0.488)	(0.487)
privateins1:eductype1	−0.617***	
	(0.231)	
privateins1:eductype2	−0.529**	
	(0.256)	
privateins1:eductype3	−0.625	
	(0.410)	
privateins1:eductype4	−0.657*	
	(0.374)	
Constant	1.219	1.137
	(0.806)	(0.802)
Observations	1,198	1,198
Log Likelihood	-2,115.872	-2,119.989
Akaike Inf. Crit.	4,295.745	4,295.977
Note:	*p<0.1; **p<0.05; ***p<0.01	

3 Quality of Model Fit

To assess the quality of the model fit, first, we look at Cook's distance and Leverage at Figure 7.

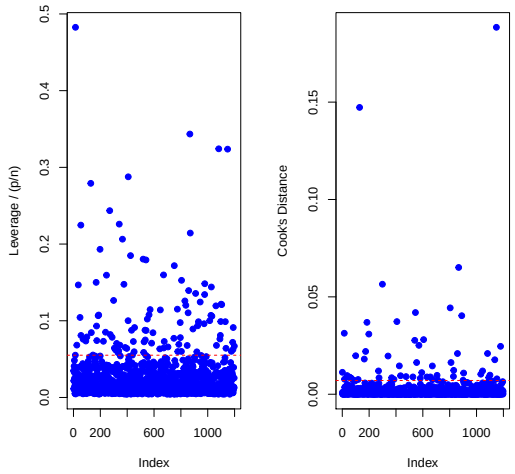

Figure 7: Leverage and Influence

As can be seen, we have 108 observations above the critical threshold of leverage which is $\frac{2p}{n}$ and 57 observations above the critical threshold of influence which is $\frac{8}{n-2p}$. The observations corresponding to the highest influence are given below:

	female	age	hhninc	hhkids	married	employed	docvis	addins	eductype	privateins
1	0	62	8.52	0	0	0	7	1	1	0
2	1	29	7.31	1	0	0	7	0	0	0

We notice that both of these observations have 7 doctor visits. Referring to Figure 2, we see that this an extremely rare occurrence. However, since we do not have any other reason to exclude these two observations other than their high influence and rarity, we choose to leave them in the model.

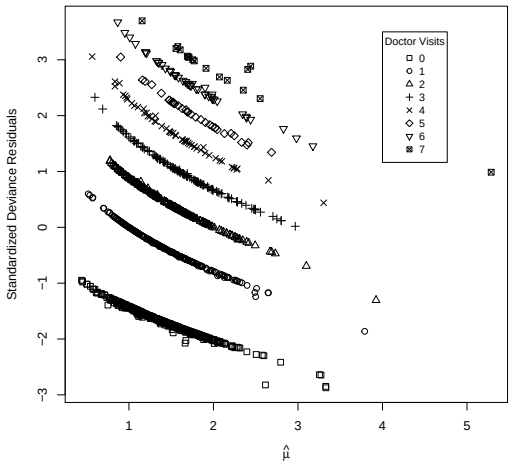

Figure 8: Standardized Deviance Residuals Colored by Number of Doctor Visits

Secondly, we look at the standardized deviance residuals. We realize that this plot has levels and decide to color the standardized deviance residuals with number of doctor visits. This leads us to discover that the levels occur because of different number of doctor visits. Our model seems to do well for observations with 0 doctor visits, which is the most common number of doctor visits in our data. So, this is expected. However, the fit of our model is not great for higher number of doctor visits. Since 0 doctor visits dominate our data, this is justified.

Lastly, we observe that all of the variables, and interaction terms remaining in our model are statistically significant. We also reject that our model is indifferent from the Null Model with approximately 100% confidence. The ANOVA table with chi-square test is also given below to provide evidence of a good fit:

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			1197	2539.51	
age	1	35.43	1196	2504.07	0.0000
married	1	1.74	1195	2502.33	0.1869
privateins	1	5.65	1194	2496.68	0.0174
employed	1	18.78	1193	2477.90	0.0000
hhkids	1	4.44	1192	2473.46	0.0351
addins	1	5.97	1191	2467.49	0.0146
eductype	5	19.10	1186	2448.40	0.0018
hhninc	1	6.16	1185	2442.24	0.0131
female	1	36.18	1184	2406.06	0.0000
age:employed	1	0.09	1183	2405.97	0.7643
age:addins	1	4.40	1182	2401.56	0.0358
age:female	1	3.80	1181	2397.76	0.0514
married:employed	1	0.63	1180	2397.14	0.4275
married:eductype	5	19.13	1175	2378.00	0.0018
married:hhninc	1	1.59	1174	2376.41	0.2068
privateins:eductype	4	9.12	1170	2367.29	0.0583
employed:hhkids	1	3.59	1169	2363.70	0.0581
employed:addins	1	1.04	1168	2362.66	0.3070
hhkids:hhninc	1	3.20	1167	2359.46	0.0738
addins:hhninc	1	4.17	1166	2355.29	0.0411

Since the p-values are quite high for *age:employed*, and *married:employed*, we run the corresponding Likelihood Ratio Tests and conclude that we can reject that these coefficients are not 0 at 10% confidence level.

4 Model Interpretation

To interpret our model, we begin by introducing some equations. In a poisson GLM, $log(\mu_i) = \eta_i = \beta x$ where in our case since the link function is *log* thus:

$$log(\hat{y_i}) = \beta_0 + \beta_1 x_i \tag{4}$$

which implies:

$$y_i = exp(\beta_0 + \beta_1 x_i) \tag{5}$$

and thus assume $x_i = x_j$ initially and x_j goes up by 1

$$y_j = exp(\beta_0 + \beta_1 (x_i + 1)) \tag{6}$$

thus,

$$\frac{y_j}{y_i} = exp(\beta_1) \tag{7}$$

Thus, given a unit change in x_j, the fitted $\hat{y}$ changes by $\hat{y}\left(e^{\beta_1} - 1\right)$.This is valid for all single terms, including numerical and categorical variables. Now onto explaining interaction terms:

$$y_i = exp(\beta_0 + \beta_1 x_i + \beta_2 x_i z_i + \epsilon_i) \tag{8}$$

Since all of our interaction terms are either a multiplication of two binary terms or one binary and one numerical value, we will assume z_i is a binary variable. Then, if x_i increases by 1, using a similar logic to above:

$$y_{i_{new}} = exp(\beta_1 + \beta_2 z_i) y_i \tag{9}$$

Thus, given a unit change in x_j for an observation with $z_i = 1$, the fitted $\hat{y}$ changes by $\hat{y}\left(e^{\beta_1 + \beta_2} - 1\right)$. We define our baseline as an *unemployed and unmarried man without kids, without additional or private insurance with 0 income and 0 age and no schooling.* Below, we omit interpreting the same coefficient twice for the different terms in the interaction. **Below when we say "Doctor visit changes by x", our scale is in terms of the response variable, as we showed in Equation 7 and Equation 9**.

1. **employed:** The respective coefficient is 0.22, which is statistically insignificant at 10%. There are also interactions term involving *employed* with *hhkids, addins, married*, and *age* with coefficients -0.258, -.58, .26,-.008 respectively. The first two are significant at 5%, and the third coefficient is significant at 10%. Thus, in expectation, becoming employed increases doctor visits for a baseline individual by $e^{0.22} - 1 = 0.25$. The differential impact of becoming employed is $e^{0.22-0.258} - 1 = -0.04$ for the baseline individual with kids, $e^{0.22-0.58} - 1 = -0.3$ for the baseline individual who has additional health insurance, and $e^{0.22+0.26} - 1 = 0.62$ for a baseline individual who is married, and lastly $e^{0.22-0.008} - 1 = .24$ for a baseline individual with a unit change in age.

2. **hhkids:** The respective coefficient is -1.53. There is also an interaction term with *hhninc* which has coefficient 0.19. So, having kids changes the number of doctor visits by $e^{-1.53} - 1 = -.78$. The differential impact is *hhninc* $e^{-1.53+0.19} - 1 = -.73$. Both of the terms are insignificant at 10% confidence level.

3. **addins:** The respective coefficient is -4.5. The are also interaction terms with *hhninc,age* with coefficients 0.79, and -.03 respectively. So, having additional insurance reduces doctor visits by .99. However, referring to Table 1 we know we have very limited number of observations and this is statistically insignificant at 10% confidence level. The differential impact is $e^{-4.5+0.79} - 1 = -.97$ for a unit change in log monthly income (e), and .99 for a unit change in age for the baseline individual. The first coefficient is insignificant at 10% confidence level and the coefficients for interactions terms are significant at 5% level.

4. **hhninc:** For a unit change in log(income) which would be e increase in income, doctor visits go down $e^{-.14} - 1 = -.13$. The differential impact is $e^{-.14-.19} - 1 = -.28$ for a married individual versus the baseline characteristics. The first coefficient is significant at 5%, and the interaction term is significant at 10%.

5. **married:** Becoming married, increases doctor visits by $e^{1.1} - 1 = 2$. The differential impact is $e^{1.1+0.3} - 1$ for someone who attained high school. The other levels of schooling can be interpreted in a similar fashion. All interaction terms and the single term are insignificant at 10% except the interaction term corresponding to $eductype = 5$.

6. **age:** Every unit increase in age increases the doctor visits for a baseline individual by $e^{.02} - 1 = .02$. The differential impact is $e^{.02-0.008} - 1 = .01$ for females for every unit increase in age. Both of the terms are significant at 5%.

7. **privateins:** Having private insurance reduces doctor visits by -0.04 at the baseline. This is insignificant at the 5% level.

8. **female:** Being female increases doctor visits by 1 at the baseline. This is significant at %1 level.

9. **eductype:** Going to high school reduces doctor visits by $e^{-.33} - 1 = -.28$. The interpretation can be done in a similar fashion for other categories. The coefficients corresponding to 3rd and 5th *eductype* are significant at 10% and 5% respectively.

5 Dispersion Parameter and Implications

We estimate $\hat{\phi}$ using

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \tag{10}$$

where $\hat{\mu}_i = g^{-1}(x_i^T \hat{\beta})$ and $V(x) = x$ since we have an underlying poisson distribution. The estimated $\hat{phi}$ is 1.945.

The estimated value implies that there is an underlying issue with the assumption that the sample is distributed with a poisson distribution. We can fix this issue by choosing another underlying distribution. Another way of fixing it is by multiplying the standard errors by $\sqrt{\hat{\phi}}$. The resulting regression table is provided below:

Table 2: Regression Results (Dispersion Adjusted)

	Dependent variable:
	docvis
employed1	0.256
	(0.371)
age	0.017**
	(0.007)
hhninc	−0.157
	(0.123)
married1	1.113
	(1.368)
female1	0.685**
	(0.285)
privateins1	0.432*
	(0.260)
hhkids1	−1.541
	(1.337)
addins1	−3.755
	(3.963)
eductype1	−0.323
	(0.508)
eductype2	−0.086
	(0.516)
eductype3	−0.754
	(0.637)
eductype4	−0.425
	(0.537)
eductype5	−1.062*
	(0.575)
employed1:age	−0.009
	(0.007)
employed1:married1	0.254
	(0.186)
employed1:hhkids1	−0.278
	(0.182)
employed1:addins1	−0.529
	(0.403)
age:female1	−0.008
	(0.006)
age:addins1	−0.031
	(0.020)
hhninc:married1	−0.191
	(0.162)
hhninc:hhkids1	0.196
	(0.167)
hhninc:addins1	0.702
	(0.487)
married1:eductype1	0.307
	(0.619)
married1:eductype2	−0.020
	(0.633)
married1:eductype3	0.559
	(0.752)
married1:eductype4	0.025
	(0.684)
married1:eductype5	0.889
	(0.680)
privateins1:eductype1	−0.617*
	(0.322)
privateins1:eductype2	−0.529
	(0.358)
privateins1:eductype3	−0.625
	(0.571)
privateins1:eductype4	−0.657
	(0.522)
Constant	1.219
	(1.124)
Observations	1,198
Log Likelihood	-2,115.872
Akaike Inf. Crit.	4,295.745
Note:	*p<0.1; **p<0.05; ***p<0.01

From Table 2, We can observe that most coefficients lost their significance because of the change in standard errors. Thus, we can conclude that we have a poor fit because of our wrong assumption about $\hat{\phi}$. Some of the consequences of having over-dispersion are:

1. The variances of β_j are too small
2. CIs will be too narrow
3. p-values are too optimistic

These theoretical consequences are in line with our observations using Table 2.

6 Conclusion

In this analysis, first we analyzed the frequencies of different exploratory variables and commented on how imbalances may skew results for some of the variables. After, we used box-and-whiskers plots and confidence intervals to explore the relationships between different variables and number of doctor visits. We used AIC, ANOVA, and R^2 to obtain the regression in the second column of Table 1. We also checked that all of our coefficients are significant using LRT statistics. In the third section, we discovered that we do not have a great fit through high cook's distance and leverage values. Also, the standardized deviance residuals plot helped us identify that we have a very poor fit as the number of doctor visits increase. Our model however was shown to be robust to various tests and it has a decent R^2. In Section 4 we derived the equations to interpret the coefficients and provided the interpretations for the coefficients in our model. In section 5, we estimated the dispersion parameter and talked about how over dispersion may impact our results. Lastly, we provided the regression table accounting for overdispersion by adjusting standard errors. This provided empirical evidence for the theoretical implications.

7 R Code

```
library(data.table)
library(ggplot2)
library(stargazer)
library(xtable)
library(dplyr)
library(rsq)
require(plotrix)
library('plyr')

dt = fread("dvis.csv")

pdf("educyrs_vs_eductype.pdf")
plot(unique(dt[,list(eductype, educyrs)]))
dev.off()

# Clean eductype, educyrs
dt = dt[educyrs %% 0.5 == 0,]

# Drop educyrs. Leave eductype because more informative. More school doesnt mean
# better quality or more education. It could be repetition etc
cor(dt)

dt[,educyrs := NULL]
dt[,hhninc:=log(hhninc*1000)]

# Convert variables to factor
dt[,female:=as.factor(female)]
dt[,hhkids:=as.factor(hhkids)]
dt[,married:=as.factor(married)]
dt[,employed:=as.factor(employed)]
dt[,addins:=as.factor(addins)]
dt[,privateins:=as.factor(privateins)]
```

```
dt[,eductype:=as.factor(eductype)]

# dt[,age:=log(age)]
# dt[,educyrs:= log(educyrs)]

summary(dt[,-c("age","hhninc","docvis")])

# 4.1

# Get frequencies
pdf("Binary Variables freq.pdf")
par(mfrow = c(2,3))
barplot(table(dt$female), main = "Sex",
names.arg=c("Male","Female"), ylab="Frequency",
ylim=c(0,800))
barplot(table(dt$hhkids), main = "Child under 16",
names.arg = c("No", "Yes"),
ylab="Frequency",ylim=c(0,800))
barplot(table(dt$married), main ="Married",
names.arg=c("Other", "Married"),
ylab="Frequency",ylim=c(0,1000))
barplot(table(dt$employed), main = "Employed",
names.arg=c("No", "Yes"),
ylab="Frequency",ylim=c(0,1000))
barplot(table(dt$addins), main = "Addt. Insurance",
names.arg=c("No", "Yes"),
ylab="Frequency",ylim=c(0,1200))
barplot(table(dt$privateins), main= "Private
Insurance", names.arg = c("No", "Yes"),
ylab="Frequency",ylim=c(0,1200))
dev.off()

pdf("Non-binary variables.pdf")
par(mfrow = c(2,2))
barplot(table(dt$eductype), main= "Highest Degree",
ylab="Frequency",ylim=c(0,800))
hist(exp(dt$hhninc), main="Monthly Income",
ylab="Frequency", xlab="")
barplot(count(dt$age)$freq ~ count(dt$age)$x,
main="Age", ylab="Frequency", xlab="",ylim=c(0,50))
barplot(table(dt$docvis), main="Doctor Visits",
ylab="Frequency",ylim=c(0,600))
dev.off()

dt[,agegroup:=20]
dt[age < 40 & age >= 30, agegroup:=30]
dt[age < 50 & age >= 40, agegroup:=40]
dt[age < 60 & age >= 50, agegroup:=50]
dt[age < 70 & age >= 60, agegroup:=60]

pdf("Docvis against age and Income.pdf")
par(mfrow = c(1, 2))
m = aggregate(dt$docvis, list(dt$agegroup), FUN=
"mean")$x
s = aggregate(dt$docvis, list(dt$agegroup), FUN=
"sd")$x

plotCI(sort(as.integer(unique(dt$agegroup))),m,
ui=m+s/2, li=m-s/2, xlab= "Age", ylab= "Doctor
Visits",
       main="Doctor Visits against Age Groups")
dt[,agegroup:=NULL]

m = aggregate(dt$hhninc, list(dt$docvis), FUN=
"mean")$x
s = aggregate(dt$hhninc, list(dt$docvis), FUN= "sd")$x

plotCI(sort(unique(dt$docvis)),m, ui=m+s/2, li=m-s/2,
xlab= "Doctor Visits",
       ylab= "Log Monthly Household Income",
       main="Doctor Visits against Log Income")
```

```
dev.off()

pdf("Doctor Visits against Categorical Variables.pdf")
par(mfrow=c(1,4))
boxplot(docvis ~ eductype, data=dt)
boxplot(docvis ~ female, data = dt)
boxplot(docvis ~ hhkids, data = dt)
boxplot(docvis ~ addins, data = dt)
dev.off()
pdf("Doctor Visits against Categorical
Variables 2.pdf")
par(mfrow=c(1,3))
boxplot(docvis ~ employed, data =dt)
boxplot(docvis ~ privateins, data =dt)
boxplot(docvis ~ married, data =dt)
mtext("Doctor Visits against Categorical Variables",
side = 3, line = -3, outer = TRUE)
dev.off()

# 4.2 Poisson GLM & Model Selection
glm1 <- glm(docvis ~ (.)^2, data= dt, family=poisson)

summary(glm1)
rsq(glm1)

bl = glm(docvis ~ (.)^2, data= dt, family=poisson)
best = 100000
# Iterate over to ensure we get the best model because AIC depends on order
# Scope undefined defaults to backwards
for (i in range(1,8)) {
  temp = step(glm(docvis ~ (.)^2, data= sample(dt), family=poisson),direction="both")
  print(temp$aic)
  if (temp$aic < best) {
    bl = temp
    best = temp$aic
  }
}

1 - pchisq(bl$deviance - glm1$deviance, length(coef(glm1)) - length(coef(bl)))

pdf("Deviance Residuals.pdf")
par(mfrow=c(1,1))
plot(predict(bl,type="response"), rstandard(bl),
     xlab=expression(hat(mu)), ylab="Standardized Deviance Residuals",
     pch=16, col=dt$docvis)
legend(4.5,3.5, sort(unique(dt$docvis)),lty=1:2, cex=0.8, pch=16,col=1:7)
dev.off()

rsq.kl(bl)
summary(bl)
anova(bl)

s <- stargazer(bl, title="Regression Results", align=TRUE,column.sep.width = "0.4pt",
               font.size="tiny")
fileConn<-file("stargazer.txt")
writeLines(s, fileConn)
close(fileConn)

# Check if we can drop any variable
for (i in 1:length(dt)){
  if (i == which(colnames(dt) == "docvis")) {
    next
  }
  glm2 = glm(docvis ~ (.)^2, data=subset(dt, select= -c(i)))
  print(1 - pchisq(glm2$deviance - glm1$deviance, length(coef(glm1)) - length(coef(glm2))))
}
# Check if we can drop interaction terms from the regression
# married-eductype
test2 <- glm(formula = docvis ~ employed + hhkids + addins + hhninc +
               married + age + privateins + female + eductype + employed:hhkids +
               employed:addins + employed:married + employed:age + hhkids:hhninc +
```

```
                addins:hhninc + addins:age + hhninc:married +
                age:female + privateins:eductype, family = poisson, data = sample(dt))

1 - pchisq(test2$deviance - bl$deviance,  length(coef(bl)) - length(coef(test2)))

rsq(test2)

# privateins-eductype
test3 <- glm(docvis ~ employed + hhkids + addins + hhninc +
                married + age + privateins + female + eductype + employed:hhkids +
                employed:addins + employed:married + employed:age + hhkids:hhninc +
                addins:hhninc + addins:age + hhninc:married + married:eductype +
                age:female, family = poisson, data = sample(dt))

1 - pchisq(test3$deviance - bl$deviance,  length(coef(bl)) - length(coef(test3)))

rsq(test3)

# hhkids1:hhninc
test4 <- glm(docvis ~ employed + hhkids + addins + hhninc +
                married + age + privateins + female + eductype + employed:hhkids +
                employed:addins + employed:married + employed:age +
                addins:hhninc + addins:age + hhninc:married + married:eductype +
                age:female, family = poisson, data = sample(dt))

1 - pchisq(test4$deviance - bl$deviance,  length(coef(bl)) - length(coef(test4)))
rsq(test4)

# age:employed
test5 <- glm(docvis ~ employed + hhkids + addins + hhninc +
                married + age + privateins + female + eductype + employed:hhkids +
                employed:addins + employed:married + hhkids:hhninc +
                addins:hhninc + addins:age + hhninc:married + married:eductype +
                age:female, family = poisson, data = sample(dt))

1 - pchisq(test5$deviance - bl$deviance,  length(coef(bl)) - length(coef(test5)))

s <- stargazer(bl, test3, title="Regression Results", align=TRUE, column.sep.width = "0.4pt",
                font.size="tiny")
fileConn<-file("stargazer.txt")
writeLines(s, fileConn)
close(fileConn)


# 4.3 Model Fit

p <- length(coef(bl) - 1)
n <- nrow(dt)

pdf("leverage and influence.pdf")
par(mfrow=c(1,2))
plot(hatvalues(bl),
     pch=19, col='blue', ylab='Leverage / (p/n)')

abline(2*p/n, 0,lty=2, col="red")

plot(cooks.distance(bl),
     pch=19, col='blue', ylab="Cook's Distance")

abline(8/(n - 2*p), 0, lty = 2, col = "red")
dev.off()

dt[which(hatvalues(bl) > 0.3,)]
dt[which(cooks.distance(bl) > 0.1),]

dt[which(hatvalues(bl) > 2*p/n),]
dt[which(cooks.distance(bl) > 8/(n-2*p)),]

qqnorm(rstandard(glm1), pch=19, main="")
qqline(rstandard(glm1))
```

```
lambda <- bl$null.deviance - bl$deviance

1 - pchisq(lambda, p)

rsq.kl(bl)

xtable(anova(bl, test="Chisq"))
# 4.4 Interpretation


# 4.5 Estimating Dispersion Parameter
mu = predict(bl, type="response")
phi = 1/(n-p)*sum((dt$docvis- mu)^2/mu)

# Fix the standard errors using phi
s <- stargazer(bl, title="Regression Results
(Dispersion Adjusted)", align=TRUE, column.sep.width =
"0.4pt",
                font.size="tiny", apply.se =
                function(x) {x*phi^0.5})
fileConn<-file("regression with dispersion fixed.txt")
writeLines(s, fileConn)
close(fileConn)
```