SB1 Assessed Practical 2

Candidate number: 1035161

March 14, 2021

1 Data

The data set provides information about the number of visits to a doctor (*docvis*) in 2001 for a sample of 1084 Americans. It has 6 binary explanatory variables (the sex of the individual, whether they are married, white, hispanic, have a chronic condition or own a private health insurance plan), two discrete variables (the age of the patient divided by 10 and their years of education) and a continuous explanatory variable *lincome* defined as $log(\frac{income\ (in\ \$)}{1000})$.

As can be seen in Tables (1a) and (1b), income for the observations ranges from \$9703 to \$185209 per year, while the most popular number of schooling years is 12 (likely corresponding to high school education).

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
\$9703	\$19999	\$31000	\$31414	\$46513	\$185209

Table (1a): Quartiles and mean for income of the observations

Education yrs	6	7	8	9	10	11	12	13	14	15	16	17
Frequency	24	5	16	39	42	58	352	81	163	34	177	93

Table (1b):; Frequency table for education years

In Figure (2) we can see that *chronic* seems to be very strongly correlated with *docvis*. In fact, an individual with no chronic conditions is over 3 times more likely to never visit the doctor than a person with a chronic condition. On the other hand, over 20% of those with a chronic condition go to the doctor at least 10 times a year.

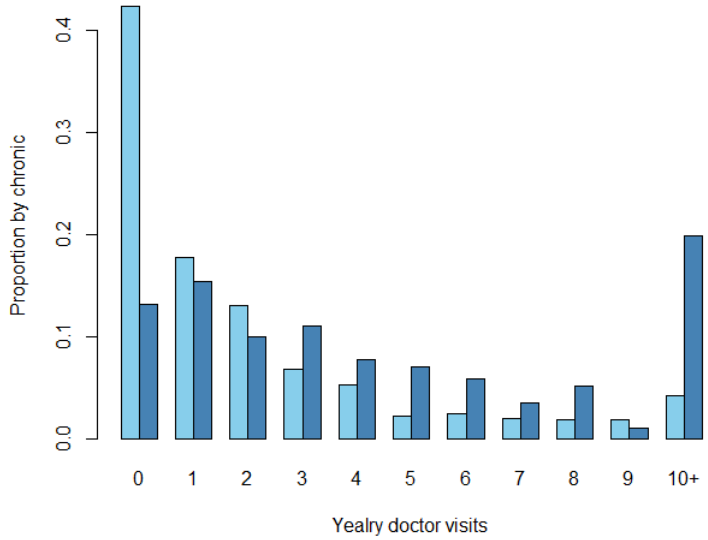

Figure (2): Barchart of the proportion of individuals with (dark blue) and without (light blue) chronic conditions, for number of yearly visits to a GP

We might expect that older people tend to go to the doctor more often. In Table (3) we can see that as *docvis* increases the descriptive statistics for *age* tend to increase.

doc visits	1st Qu.	Median	Mean	3rd Qu.
0-2	32	39	40.71	47
3-9	35	42	42.56	49
10+	34	43	43.9	53

Table (3): Quartiles and mean for *age* for three different ranges of *docvis*

For what concerns ethnicity, Table (4) shows that the majority of the individuals in the sample are white.

	non hisp	hisp
non white	135	202
white	747	0

Table (4): Cross table for *white* and *hispanic*

From Figure (5) it can be seen that gender and *docvis* have a good correlation, with males going to fewer doctor's appointments than females.

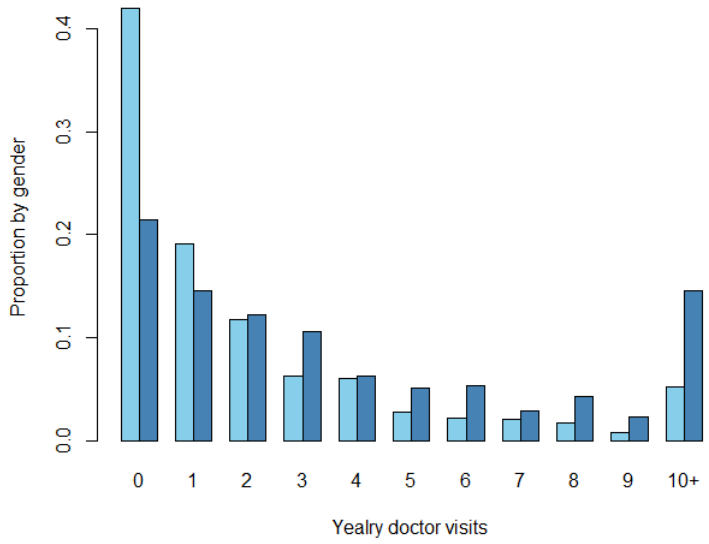

Figure (5): Barchart of the proportion of males (light blue) and females (dark blue) for number of yearly visits to a GP

2 Modelling

Modelling our response variable *docvis* as a $Poisson(\lambda)$ seems reasonable from the exploratory data analysis, we can therefore adopt a GLM with canonical link function where:

$$\eta = x^T\beta = log(\lambda) \qquad \lambda = exp(x^T\beta) = exp(\eta) \qquad (1)$$

We can start with a full Poisson GLM which we shall call M_0. In Table (6) we can see that *hispanic* does not seem to be significant (as a matter of fact we saw in our initial analysis that *white* already encoded most of the information about ethnicity).

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-0.95300	0.15770	-6.04	1.5e-09 ***
age	0.07288	0.01714	4.25	2.1e-05 ***
educ	0.02845	0.00827	3.44	0.00058 ***
lincome	0.05846	0.03074	1.90	0.05716 .
female	0.58299	0.03527	16.53	< 2e-16 ***
married	0.09476	0.03706	2.56	0.01056 *
white	0.20261	0.05447	3.72	0.00020 ***
private	0.43375	0.07249	5.98	2.2e-09 ***
chronic	0.79083	0.03417	23.15	< 2e-16 ***
hispanic	-0.08659	0.07395	-1.17	0.24166

Table (6): Summary for M_0

Formally, let:

$$M_0: \qquad \eta = \beta_0 + \beta_1 age + \beta_2 educ + \beta_3 lincome + \beta_4 hispanic + \beta_5 married + \beta_6 white + \beta_7 private + \beta_8 chronic + \beta_9 hispanic$$

$$M_1: \qquad \eta = \beta_0 + \beta_1 age + \beta_2 educ + \beta_3 lincome + \beta_4 hispanic + \beta_5 married + \beta_6 white + \beta_7 private + \beta_8 chronic$$

Since the two models are nested (they have parameter spaces of dimension 10 and 9 respectively), we can perform a likelihood ratio test to test the hypothesis:

$H_0: \quad \beta_9 = 0$
$H_1: \quad \beta_9 \quad unconstrained$

$\Lambda(Y) = D^{M_1}(Y) - D^{M_0}(Y) \sim \chi^2_1 \quad under \;\; H_0$

$\Lambda(y) = 4681 - 4679.6 = 1.4 \implies p \;\; value = P(\chi^2_1 > 1.4) \approx 0.23$

We have no evidence to reject H_0 and therefore opt for M_1, dropping the variable *hispanic*.

In the Table (7) we can see that all the Wald tests carried out to test the null $H_0: \quad \beta_i = 0$ have p-values considerably below our 0.1 significance level (the highest is 0.06). Hence we can conclude that we cannot further simplify the model as long as we are estimating the standard errors using:

$Var(\hat{\beta}_j) = (X^T W X)^{-1}_{jj} = (X^T diag(\hat{\lambda}_1, \ldots, \hat{\lambda_n}) X)^{-1}_{jj} \quad (*)$

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-1.02621	0.14481	-7.09	1.4e-12 ***
age	0.074401	0.01708	4.36	1.3e-05 ***
educ	0.03029	0.00811	3.74	0.00019 ***
lincome	0.05619	0.03068	1.83	0.06704 .
female	0.58431	0.03525	16.58	< 2e-16 ***
married	0.08956	0.03678	2.44	0.01488 *
white	0.24357	0.04239	5.75	9.2e-09 ***
private	0.44428	0.07195	6.17	6.6e-10 ***
chronic	0.79302	0.03412	23.24	< 2e-16 ***

Table (7): Summary for M_1

Using non-parametric paired bootstrap we can obtain an estimate $\Sigma_{bootstrap}$ for the covariance matrix of $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_8)^T$.

Taking the square root of the diagonal entries we can see in Table (8) that the bootstrap estimates for the standard errors of β_i are about 2 times higher than the standard errors computed using (*).

Beta	0	1	2	3	4	5
standard error	0.145	0.0171	0.00811	0.0307	0.0353	0.0368
bootstrap st.err.	0.400	0.043	0.020	0.076	0.088	0.099

Beta	6	7	8
standard error	0.0424	0.0720	0.0341
bootstrap st.err.	0.105	0.177	0.086

Table (8): Standard errors computed using (*) and using non-parametric paired bootstrap

If we assume that:

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{Var(\hat{\beta}_i)}} \approx N(0,1) \; for \; large \; n$$

Then by estimating $Var(\hat{\beta}_i)$ using $(\Sigma_{bootstrap})_{ii}$ we have that the following is a $1-\alpha$ confidence interval for β_i

$$(\hat{\beta}_i - z_{\alpha/2} \cdot \sqrt{(\Sigma_{bootstrap})_{ii}}, \hat{\beta}_i + z_{\alpha/2} \cdot \sqrt{(\Sigma_{bootstrap})_{ii}})$$

Beta	Lower	upper
0 - intercept	-1.6837	-0.369
1 - age	0.0036	0.145
2 - educ	-0.0030	0.064
3 - lincome	-0.0692	0.182
4 - female	0.4394	0.729
5 - married	-0.0727	0.252
6 - white	0.0706	0.417
7 - private	0.1539	0.735
8 - chronic	0.6516	0.934

Table (9): 90% normal bootstrap confidence intervals for β_i

Initially, when computing $\hat{\beta}$ we worked under the assumption that $Y_i \sim Poisson(\lambda_i)\ for\ all\ i \leq n$ and $x_i^T\beta = log(\lambda_i)$ $(*)$, and we still kept such assumption when computing the standard errors $se(\beta_j)\ for\ j \leq n$. So we had:
$Var(\hat{\beta}_j) = (X^T W X)^{-1}_{jj} = (X^T diag(\hat{\lambda}_1, \ldots, \hat{\lambda}_n)X)^{-1}_{jj}$

On the other hand, after having found $\hat{\beta}$ (under the Poisson GLM assumption), when computing the bootstrap standard errors we dropped any assumption about the distribution of $\hat{\beta}$ and simply approximated it with the empirical distribution of the bootstrap samples.

Therefore if $(*)$ does not accurately describe the relationship between Y and X, our initial estimates for $Var(\hat{\beta})$ might be off and consequently in our case small estimated standard errors resulted in unrealistically low p values for the Wald tests.

Since now our estimates for variances and covariances have increased significantly, we might want to see whether we can simplify our model. In Table (9) it can be seen that 90% normal bootstrap confidence intervals for $\beta_{2-educ},\ \beta_{3-lincome}$ and $\beta_{5-married}$ contain zero; therefore it might be sensible to test if we can drop these variables.

$H_0: \quad \beta_{2-educ}, = \beta_{3-lincome} = \beta_{5-married} = 0$

$H_1: \quad \beta_{2-educ},\ \beta_{3-lincome},\ \beta_{5-married}\ unconstrained$

$$W^B = \begin{pmatrix}\hat{\beta}_2 & \hat{\beta}_3 & \hat{\beta}_5\end{pmatrix} \left(\hat{Var} \begin{pmatrix}\hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_5\end{pmatrix} \right)^{-1} \begin{pmatrix}\hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_5\end{pmatrix} \approx \chi^2_3 \quad under\ H_0 \qquad (1)$$

We can estimate $Var \begin{pmatrix}\hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_5\end{pmatrix}$ using a submatrix of the bootstrap covariance matrix Σ^B that we already calculated.

$$\hat{Var} \begin{pmatrix}\hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_5\end{pmatrix} = \begin{pmatrix}\Sigma^B_{2,2} & \Sigma^B_{2,3} & \Sigma^B_{2,5} \\ \Sigma^B_{3,2} & \Sigma^B_{3,3} & \Sigma^B_{3,5} \\ \Sigma^B_{5,2} & \Sigma^B_{5,3} & \Sigma^B_{5,5}\end{pmatrix}$$

$$W^B_{obs} = 4.36 \implies p\ value = P(\chi^2_3 > 4.36) \approx 0.23$$

Hence we have no evidence to reject H_0, so we can drop *educ*, *lincome* and *married*.

$$M_2: \quad \eta = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 white + \beta_4 private + \beta_5 chronic$$

Similarly to what done before, we can use non parametric paired bootstrap to estimate the covariance matrix of the coefficients $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_6)$ of the new model M_2.

Similarly to what R does automatically (i.e. it carries out Wald tests for $H_0 : \beta_i = 0$ for each i), we can do the same thing using bootstrap estimates for the standard errors. Although the p-values have increased (you can see it in Table (10)) they are still considerably below our 0.1 significance level. Hence we cannot further simplify the model.

	Estimate	Std.Err.	p value	S.E.(btstr)	p val(btstr)
(Intercept)	-0.4721	0.0978	1.4e-06	0.246	5.5e-02
age	0.0717	0.0168	2.0e-05	0.040	7.2e-02
female	0.5715	0.0348	< 2e-16	0.086	2.9e-11
white	0.3032	0.0411	1.6e-13	0.093	1.1e-03
private	0.5337	0.0700	2.5e-14	0.185	3.9e-03
chronic	0.7907	0.0341	< 2e-16	0.085	0.0e+00

Table (10): Coefficients estimates and p-values for Wald tests using 'classic' standard errors and bootstrap standard errors

3 Interpretation

Our final model is:

$Y \sim Poisson(\lambda)$

$\lambda = exp(x^T\beta) = exp(\eta)$

$M_2: \quad \eta = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 white + \beta_4 private + \beta_5 chronic$

$\hat{\beta} = (-0.472,\ 0.0717,\ 0.572,\ 0.303,\ 0.534,\ 0.791)^T$

A 10 year increase in age (because *age* is expressed in 10 yrs units) makes $\hat{\lambda}$ grow by a factor of $e^{0.0717} \approx 1.07$ so causes a 7% increase in the estimated Poisson parameter.

Let x_1 and x_2 be two observations that differ only in gender (so all other entries are identical), and suppose x_1 is female and x_2 is male. Then $\frac{\hat{\lambda_1}}{\hat{\lambda_2}} = exp(0.572) \approx 1.77$.

Similarly, being white makes $\hat{\lambda}$ grow by a factor of $e^{0.303} \approx 1.35$, owning private insurance by a factor of $e^{0.534} \approx 1.70$ and having a chronic condition by a factor of of $e^{0.791} \approx 1.70$.

4 Appendix

```
#Summary for continuous variables
summary(df$lincome)
summary(df$educ)

#Barchart for chronic
barplot(prop.table(table(df$chronic,
        reduced_docvis), 1), beside=TRUE,
        col=c("skyblue","steelblue"), xlab='Yealry doctor visits',
        ylab='Proportion by chronic',
        names.arg = c(0,1,2,3,4,5,6,7,8,9,"10+"))

#Table for ethnicities
table(df$white, df$hispanic)

#Subsetting dataframe based on number of visits
df_temporary1 = subset(df, docvis <= 2)
df_temporary2 = subset(df, docvis > 2 & docvis < 10)
df_temporary3 = subset(df, docvis >= 10)
summary(df_temporary1$age)
summary(df_temporary2$age)
summary(df_temporary3$age)

barplot(prop.table(table(df$female, reduced_docvis), 1), beside=TRUE,
        col=c("skyblue","steelblue"), xlab='Yealry doctor visits',
        ylab= 'Proportion by gender',
        names.arg = c(0,1,2,3,4,5,6,7,8,9,"10+"))

set.seed(540)
n = length(df$docvis)

#Fitting full GLM
m0 = glm(docvis ~ ., data = df, family = "poisson")
summary(m0)

#Dropping hispanic
m1 = glm(docvis ~ . -hispanic, data = df, family = "poisson")

#Chisq test
1 - pchisq(1.4, 1) #0.23 so we have no evidence against H0
summary(m1)
```

```
#Bootstrap variance algorithm
B = 1000

beta_bootstrap = data.frame(matrix(0,B,9))
for (i in 1:B){
  sam = sample(1:n, replace = TRUE)
  df_temp = df[sam,]
  model = glm(docvis ~ . -hispanic, data = df_temp, family = "poisson")
  beta_bootstrap[i,] = model$coefficients
}
variance_matrix_bootstrap = var(beta_bootstrap)

options(digits = 3)
variance_matrix_bootstrap

#Taking the square root of the diagonal entries
#to get standard errors
std_devs = numeric(9)
for(i in 1:9){
  std_devs[i] = sqrt(variance_matrix_bootstrap[i,i])
}

std_devs

#90% confidence intervals for beta_{i}
conf_int = matrix(0,9,2)
for (i in 1:9){
  conf_int[i, 1] = m1$coefficients[i] - std_devs[i]*qnorm(0.95)
  conf_int[i, 2] = m1$coefficients[i] + std_devs[i]*qnorm(0.95)
}


#testing if we can simplify our model.
#let h_0: beduc = b_lincome = b_married = 0 h_1:
ml_cov = matrix(c(variance_matrix_bootstrap[3,3],
                  variance_matrix_bootstrap[3,4],
                  variance_matrix_bootstrap[3,6],
                  variance_matrix_bootstrap[3,4],
  variance_matrix_bootstrap[4,4], variance_matrix_bootstrap[4,6],
  variance_matrix_bootstrap[3,6], variance_matrix_bootstrap[4,6],
  variance_matrix_bootstrap[6,6]), 3,3)
```

```
v1 = c(m1$coefficients[3], m1$coefficients[4], m1$coefficients[6])
t(v1)%*% solve(ml_cov) %*% v1

1-pchisq(4.36, df = 3) #so can delete married and lincome and educ

mm2 = glm(docvis ~ . -hispanic - educ -
            married - lincome, data = df, family = "poisson")
summary(mm2)

#Calculating bootstrap covariance for simplified model
B = 1000

beta_bootstrap2 = data.frame(matrix(0,B,6))
for (i in 1:B){
  sam = sample(1:n, replace = TRUE)
  df_temp = df[sam,]
  model = glm(docvis ~ . -hispanic- educ -
                married - lincome, data = df_temp, family = "poisson")
  beta_bootstrap2[i,] = model$coefficients
}
variance_matrix_bootstrap2 = var(beta_bootstrap2)

std_devs2 = numeric(6)
for(i in 1:6){
  std_devs2[i] = sqrt(variance_matrix_bootstrap2[i,i])
}

std_devs2

options(digits = 3)
variance_matrix_bootstrap2
summary(mm2)

options(digits = 5)
tests = numeric(6)
tests[1] = 2*(1- pnorm(-mm2$coefficients[1]/std_devs2[1]))
for (i in 2:6){
  tests[i] = 2*(1 - pnorm(mm2$coefficients[i]/std_devs2[i]))
}

#Wald tests with bootstra st. err.
test
```