SB1 Assessed Practical 2

Candidate number: 1035161

January 4, 2021

1 Data

The data set provides information about private health insurance ownership (*privins*) in the US. It consists of 3103 observations and has 11 explanatory variables. Five of them are binary (the sex of the individual, whether they are retired, white, hispanic or married), three are discrete (the numbers of chronic conditions, the limitations on activities of daily living (*adl*) and the years of education) and two are continuous (the household income and its natural logarithm).

From Figure (1) we can see that the biggest effect in the proportion of *privins* due to a binary variable is caused by *hisp* (non-hispanic are 23.5% more likely to own private insurance than hispanic). On the other hand, being male, married, retired or white leads to an increase in the percentage of privately insured individuals.

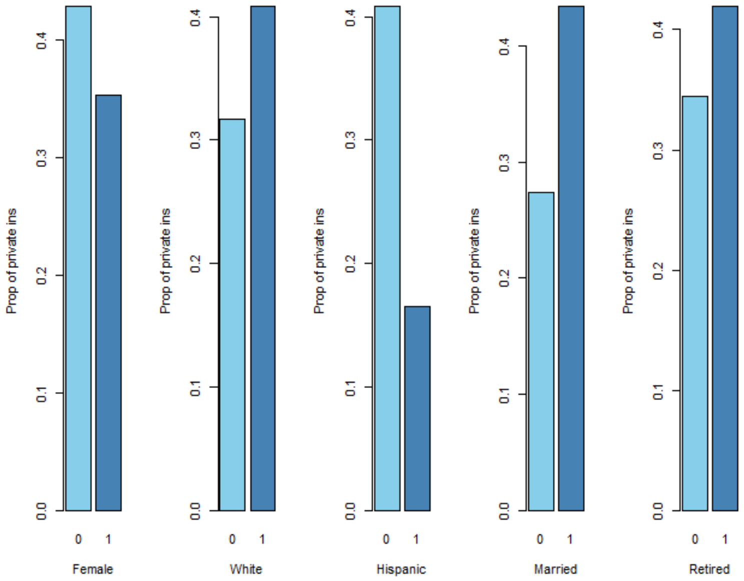

Figure (1) Barcharts for explanatory binary variables

From the table below, it can be observed that the great majority of the people analysed have zero limitations on daily activities (we will later make the decision of treating *adl* as a binary variable, based on similar considerations). From Figure (2) we can see that *adl* seems to have a negative effect on the proportion of $privins$, with a 10.3% drop occurring from $adl = 0$ to $adl = 1$.

adl	0	1	2	3	4	5
$privins = 0$	1534	175	87	51	25	15
$privins = 1$	1090	79	20	16	7	4

For what concerns the number of chronic diseases, we can see from the table below that most of the observations have between 0 and 3 chronic conditions, with only 1.6% having a value for *chronic* of 6 or above (hence these percentages might be unreliable). From Figure (2), it can be observed that $\%privins$ is lower when $chronic = 0$ than when *chronic* is between 1 and 3. For $chronic \geq 4$, as anticipated, $\%privins$ varies greatly (due to the small sample size).

chronic	0	1	2	3	4	5	6	7	8
$privins = 0$	237	507	472	359	179	95	30	7	1
$privins = 1$	136	347	342	247	93	39	8	3	1

In the table below we can see that both the mean and median number of education years in the sample is 12 (almost certainly corresponding to high school education). From Figure (2) it is clear that education years and *privins* have quite a strong positive correlation.

Min	1st Quart	Median	Mean	3rd Quart	Max
3.0	10.5	12	12.0	14	17

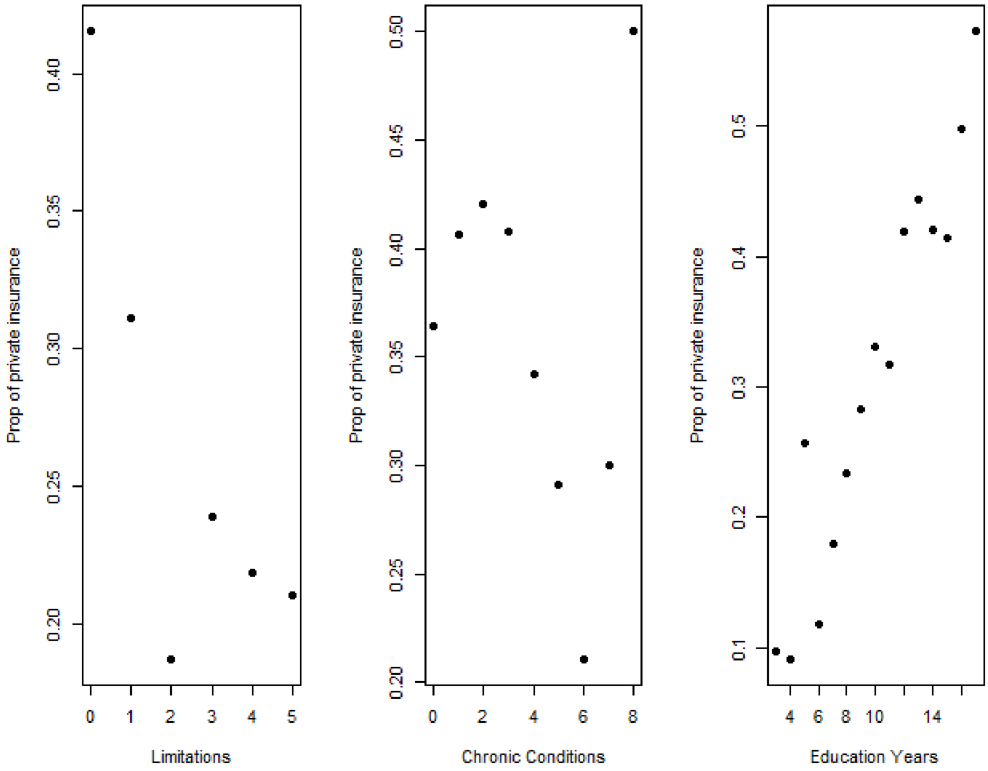

Figure (2) Proportion of private insurance owners for number of limitations in daily activities, chronic conditions and education years

The mean age for the sample is 66.9, ranging from 52 to 86. From the boxplot in Figure (3), the mean age for privately insured observations does not differ significantly from that of non-privately insured. This leads us to think that age might not be a significant explanatory variable for our analysis.

Min	1st Quart	Median	Mean	3rd Quart	Max
52	65.0	67.0	66.9	69.0	86

Household income ranges from \$1300 to \$285400 with a mean of \$41500. The variable seems to have quite a significant effect on *privins* (Figure (3)). The boxplot highlights that there are many observations with high *hhincome* which lie at least at a distance of 1.5 times the standard deviation from the 3rd quartile (i. e. are outliers in the plot). This might lead to problems when fitting a GLM. On the other hand, the boxplot for *lincome* is well-behaved (we are in fact taking a concave transformation of *hhincome*, so high values of *hhincome* are brought down significantly).

Min	1st Quart	Median	Mean	3rd Quart	Max
1.3	17.6	31.5	41.5	52.6	285.4

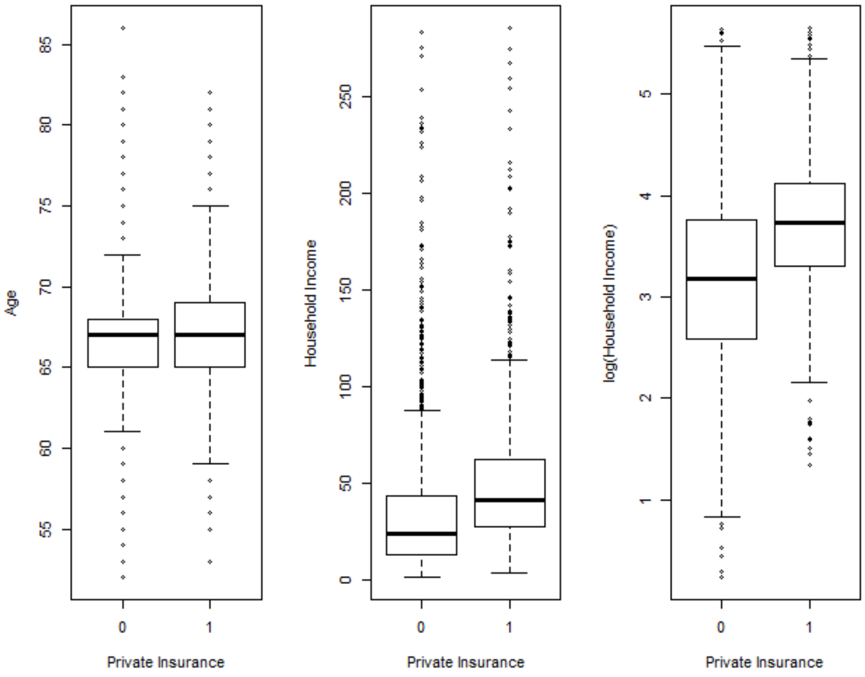

Figure (3) Boxplots for *age*, *hhincome* and *lincome*

2 Modelling

Since our response variable *privins* is a $Bernoulli(\pi)$, we can adopt a GLM with canonical logit link function where:

$$\eta = x^T\beta = \log(\frac{\pi}{1-\pi}) \qquad \pi = \frac{exp(\eta)}{1+exp(\eta)} = \frac{exp(x^T\beta)}{1+exp(x^T\beta)}$$

Firstly, we have to address the possible issues noticed in the initial data exploration. How should we treat *adl* and *chronic*? A full model M_0 (with 5 parameters for 6 levels of adl and 8 parameters for chronic) is likely to be overparametrized, perhaps a model M_1 where we treat *adl* and *chronic* as continuous variables is better.

$$\begin{aligned}
M_0: \quad & \eta = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 white + \beta_4 hisp + \beta_5 married + \\
& \beta_6 educyrs + \beta_7 hhincome + \beta_8 lincome + \beta_9 retired + \\
& \sum_{i=1}^{5} \beta_{9+i} I(adl = i) + \sum_{i=1}^{8} \beta_{14+i} I(chronic = i) \\
M_1: \quad & \eta = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 white + \beta_4 hisp + \beta_5 married + \\
& \beta_6 educyrs + \beta_7 hhincome + \beta_8 lincome + \beta_9 retired + \\
& \beta_{10} adl + \beta_{11} chronic
\end{aligned}$$

Since the two models are nested (they have parameter spaces of dimension 22 and 11 respectively), we can perform a likelihood ratio test to test the hypothesis:

$H_0: \quad \beta_{10} = 2\beta_{11} = 3\beta_{12} = 4\beta_{13} = 5\beta_{14} \quad \beta_{15} = 2\beta_{16} = ... = 8\beta_{22}$
$H_1: \quad \beta_{10}, .., \beta_{22} \quad unconstrained$

$$\Lambda(Y) = D^{M_1}(Y) - D^{M_0}(Y) \sim \chi^2_{11} \quad under \;\; H_0$$
$$\Lambda(y) = 9.8 \implies p \;\; value = P(\chi^2_{11} > 9.8) \approx 0.54$$

We have no evidence to reject H_0, therefore we opt for model M_1. However, by looking at the initial plots in Section 1, we might wonder whether treating *adl* and *chronic* as binary variables with indicator functions $I(adl \geq 1)$, $I(chronic \geq 1)$ leads to better results. In this case we get a lower scaled deviance that in M_1 and we have the same dimension of parameter space as in M_1, so M_2 is to be preferred:

$$\begin{gathered} M_2: \qquad \eta = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 white + \beta_4 hisp + \beta_5 married + \\ \beta_6 educyrs + \beta_7 hhincome + \beta_8 lincome + \beta_9 retired + \\ \beta_{10} I(adl \geq 1) + \beta_{11} I(chronic \geq 1) \end{gathered}$$

Coefficients for M_2:

```
              Estimate  Std.  Error  z value  Pr(>|z|)
 (Intercept)  -4.738    0.880        - 5.38   7.4e-08 ***
 age          -0.02304  0.01201      -1.92    0.05504 .
 female       -0.08067  0.08661      -0.93    0.35164
 white        -0.14251  0.11443      -1.25    0.21301
 hisp         -0.57914  0.20662      -2.80    0.00506
 married      -0.08801  0.11222      -0.78    0.43291
 educyrs      0.05293   0.01560      3.39     0.00069 ***
 hhincome     -0.02077  0.00242      -8.58    < 2e-16 ***
 lincome      1.71918   0.13685      12.56    < 2e-16 ***
 retired      0.13159   0.08949      1.47     0.14144
 adl          -0.25045  0.12370      -2.02    0.04290 *
 chronic      0.34613   0.12269      2.82     0.00479 **
```

Going back to our second problem, the potential outliers with high *hhincome*, a quick plot of the fitted values of M_2 (sorted by *hhincome*) clearly highlights the issue (Figure (4)). The negative coefficient of *hhincome*, greatly outweighs the positive coefficient of *lincome*, when observations have really high income. This leads to a substantial underestimation of $\hat{\pi}$ in such cases, and to very high leverage (Figure(5)).

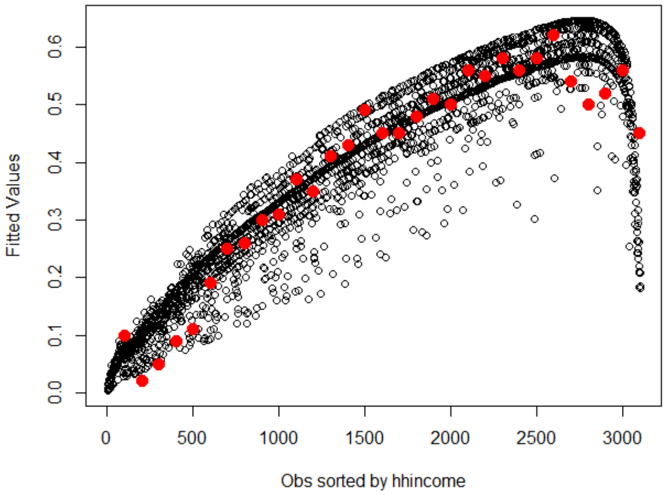

Figure (4) Fitted values for M_2 against observations sorted by income. The red dots are the proportion of private insurances for every 100 data points

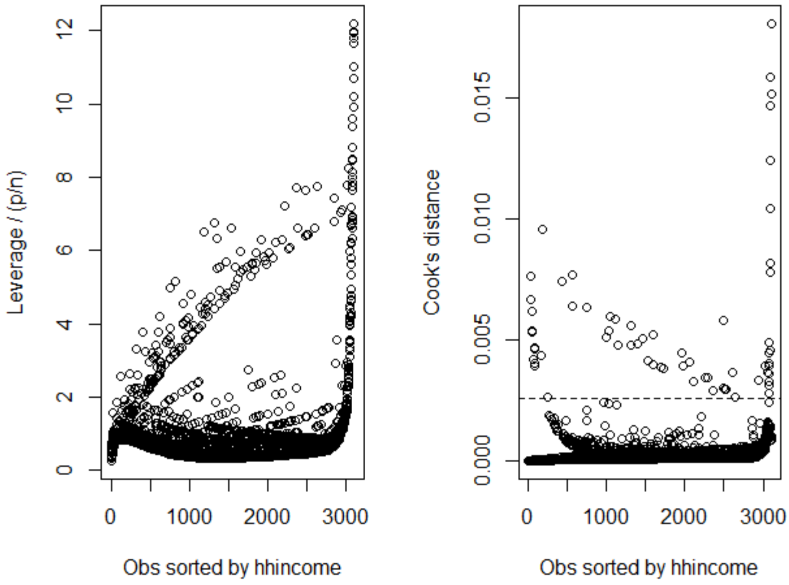

Figure (5) Diagnostics for M_2; leverage and influence plots

We have two options, either delete *hhincome* from our model, or modify the observations that were outliers in the box plot for *hhincome*.

In M_2 we can see that *hhincome* is highly significant, and this remains unchanged even when we further simplify the model (the model which minimizes BIC has explanatory variables *hisp*, *educyrs*, *hincome*, *lincome*, all highly significant). This is certainly due to the pattern that can be seen from the red dots (percentage of *privins* for every 100 observations) in Figure (4): the percentage of private insurances increases until observation 2600 approximately, and then starts decreasing. Such behaviour cannot be described by *lincome* alone because this would result in a "monotonic" plot of fitted values (suppose the coefficient of *lincome* is positive, then the higher the income the higher $\hat{\pi}$).

This suggests that our second option might work better. We create two modified versions of *lincome* and *hhincome* (called *new lincome* and *new hhincome*) such that all the observations $1.5 \cdot std(hhincome)$ above the mean are brought down to such threshold ($97400). More formally:

```
$threshold = mean(hhincome) + 1.5 \cdot std(hhincome)$
if $hhicome[i] > threshold$ then $new\ hhincome = threshold$
else $new\ hhincome[i] = hhincome[i]$
$new\ lincome[i] = log(new\ hhincome[i])$
```

If we fit a new model M_3 with *new hhincome*, *new lincome* and refit M_2 on the new dataframe (so that the deviance is calculated using the same saturated model), we see that the deviance of M_3 is considerably smaller (3691.8 against 3676.9) and since they have the same number of parameters, M_3 is preferable.

Coefficients for M_3:

```
               Estimate   Std.  Error  z value  Pr(>|z|)
 (Intercept    ) -5.94722  0.93948      -6.33    2.4e-10 ***
 age           -0.02558   0.01205      -2.12    0.03381 *
 female        -0.07518   0.08657      -0.87    0.38518
 white         -0.15893   0.11541      -1.38    0.16848
 hisp          -0.56345   0.20799      -2.71    0.00675 **
 married       -0.11854   0.11290      -1.05    0.29374
 educyrs       0.05462    0.01565      3.49     0.00048 ***
 retired       0.13292    0.08940      1.49     0.13705
 adl           -0.23327   0.12457      -1.87    0.06113 .
 chronic       0.33589    0.12230      2.75     0.00603 **
 new lincome   2.31840    0.20177      11.49    < 2e-16 ***
 new hhincome  -0.03867   0.00476      -8.13    4.3e-16 ***
```

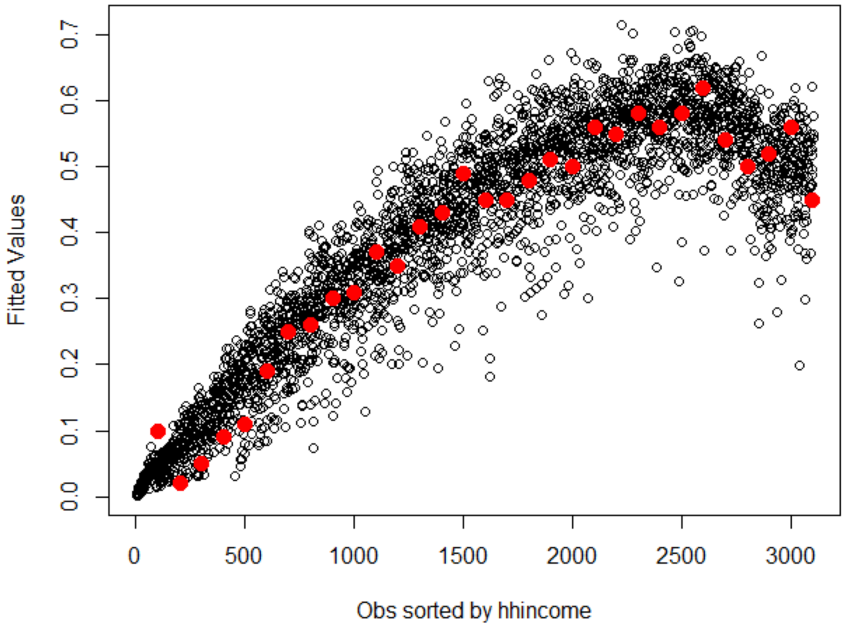

Figure (6) Fitted values for M_3 against observations sorted by income. The red dots are the proportion of private insurances for every 100 data points

A quick look at the fitted values forM_3 (Figure (6)), tells us that $\hat{\pi}$ for high income observations are much more realistic now. From the summary table above, we can see that there are 5 coefficient which are not significant. We can do a likelihood ratio test to test such hypothesis.

$$M_4: \qquad \eta = \beta_0 + \beta_1 age + \beta_4 hisp + \beta_6 educyrs + \beta_7 new\, hhincome +$$
$$\beta_8 new\, lincome + \beta_{10} I(adl \geq 1) + \beta_{11} I(chronic \geq 1)$$

$H_0: \quad \beta_3 = \beta_5 = \beta_7 = \beta_8 = 0$
$H_1: \quad \beta_3 = \beta_5 = \beta_7 = \beta_8 \;\; unconstrained$

$$\Lambda(Y) = D^{M_4}(Y) - D^{M_3}(Y) \sim \chi^2_4 \quad under \;\; H_0$$
$$\Lambda(y) = 3683 - 3676.8 = 6.2 \implies p \;\; value = P(\chi^2_4 > 6.2) \approx 0.19$$

So we do not have enough evidence to reject H_0.

Coefficients for M_4:

```
               Estimate  Std.  Error  z value  Pr(>|z|)
 (Intercept)   -6.30362  0.88492      -7.12    1.1e-12 ***
 age           -0.02028  0.01136      -1.78    0.07435 .
 hisp          -0.58139  0.20734      -2.80    0.00505 **
 educyrs        0.05643  0.01533       3.68    0.00023 ***
 adl           -0.24307  0.12399      -1.96    0.05995 .
 chronic        0.33917  0.12200       2.78    0.00543 **
 new lincome    2.25598  0.19159      11.77    < 2e-16 ***
 new hhincome  -0.03811  0.00468      -8.15    3.6e-16 ***
```

From the summary above, we can see that two Wald tests suggest us to drop *age* and *adl*. If we carry out a likelihood ratio test (which is considered to be more precise for small degrees of freedom) we reach the same conclusion.

$$M_5: \qquad \eta = \beta_0 + \beta_1 hisp + \beta_2 educyrs + \beta_3 new\, hhincome +$$
$$\beta_4 new\, lincome + \beta_5 I(chronic \geq 1)$$

$H_0: \quad \beta_{age} = \beta_{adl} = 0 \qquad H_1: \quad \beta_{age}, \beta_{adl} \;\; unconstrained$

$$\Lambda(Y) = D^{M_5}(Y) - D^{M_4}(Y) \sim \chi^2_2 \quad \textit{under} \;\; H_0$$
$$\Lambda(y) = 5.4 \implies p \;\; \textit{value} = P(\chi^2_2 > 5.4) \approx 0.665$$

Hence we do not reject H_0. Moreover, we can see in the table below that all coefficients of M_5 are highly significant.

Coefficients for M_5:

```
               Estimate  Std.  Error  z value  Pr(>|z|)
 (Intercept)   -7.70857  0.52146      -14.78   < 2e-16 ***
 hisp          -0.58353  0.20682      -2.82    0.00478 **
 educyrs       0.05756   0.01526      3.77     0.00016 ***
 chronic       0.32045   0.12148      2.64     0.00834 **
 new lincome   2.25810   0.18982      11.90    < 2e-16 ***
 new hhincome  -0.03790  0.00465      -8.15    3.6e-16 ***
```

The standardised residuals (as we are modelling Bernoulli random variables) are not very helpful, but still they have approximately unit variance which is good (Figure (7)). We see from Figure (8) that there are a few observations with low *hhincome* and high Cook's distance. This has a very natural explanation: $\hat{\pi}$, when *hhincome* is low, is around 0.1 hence, the few observations with $privins = 1$ and low *hhincome*, when removed will change the parameters significantly. Indeed all the points with Cook's distance above 0.006 have income less that $2500 but have purchased private insurance nonetheless. So there is no reason to delete them.

On the other hand, as *hhincome* grows, $\hat{\pi}$ approaches 0.5 and so both residuals and Cook's distances are not so extreme (intuitively, in this case, $\hat{\pi}$ is around 0.5, so it is at distance ± 0.5 from 1 and 0 so we get symmetric standardized residuals with values ± 1, while for low *hhincome* $\hat{\pi}$ is much closer to zero so we get residuals with approimate values of 2 and 0).

The patterns in the leverage plot also have natural explanation. Since most of our variables are binary, the effect on influence (if we fix those) depends only on income and education. Hence, as our observations in the plot are sorted by income, the patterns emerge. As a matter of fact, all observations with leverage at least four times bigger than the average are hispanic (this is the "line" in the leverage plot).

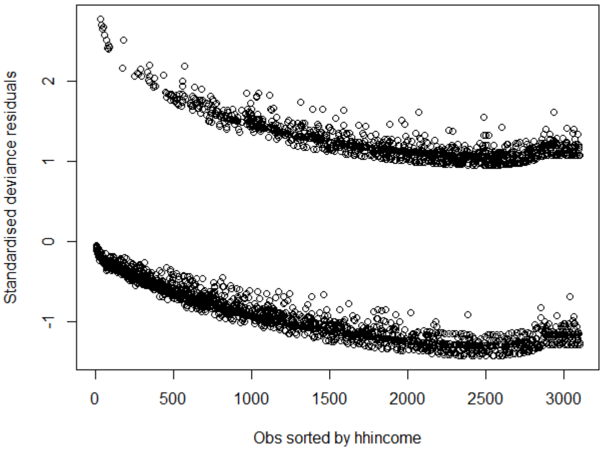

Figure (7) Standardised deviance residuals for M_5

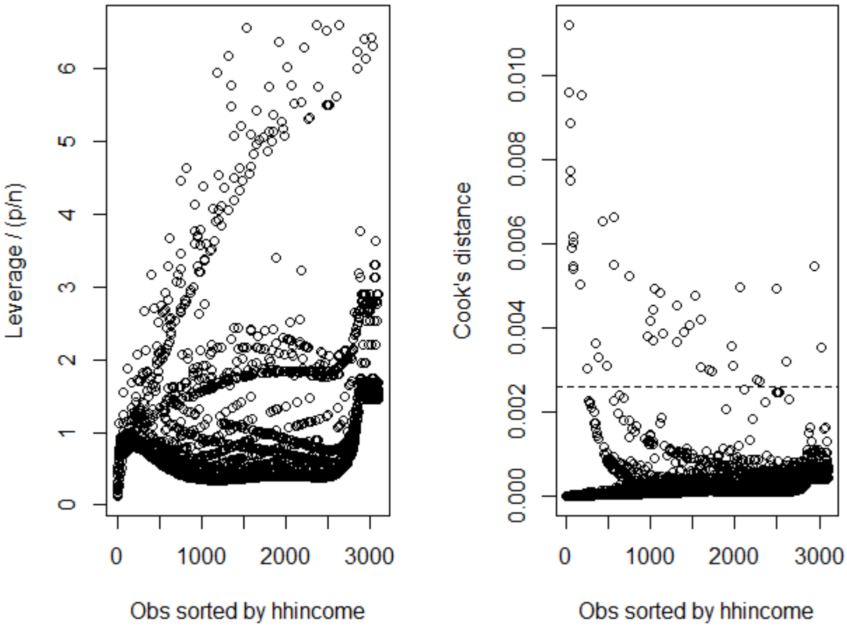

Figure (8) Diagnostics for M_5; leverage and influence plots

We just saw how some variables could have very easily been included or not in the model depending on our discretion (for instance, p-values were very close to the arbitrary 5% threshold). In addition, stepwise model selection is order-dependent. Therefore, depending on our criteria there might be better GLMs. We can find by enumeration (using *leaps* library) the optimal models according to AIC and BIC . From the Figure(9) we can see that the best model according to BIC has only 3 parameters, while that for AIC has 9. Our model with 5 explanatory variables is a good compromise between the two (we are not favouring interpretability over prediction power or viceversa).

```
Best Model for AIC:
privins ∼ 1 + age + white + hisp + educyrs + retired + adl + chronic
+ new lincome + new hhincome

Best Model for BIC:
privins ∼ 1 + educyrs + new lincome + new hhincome
```

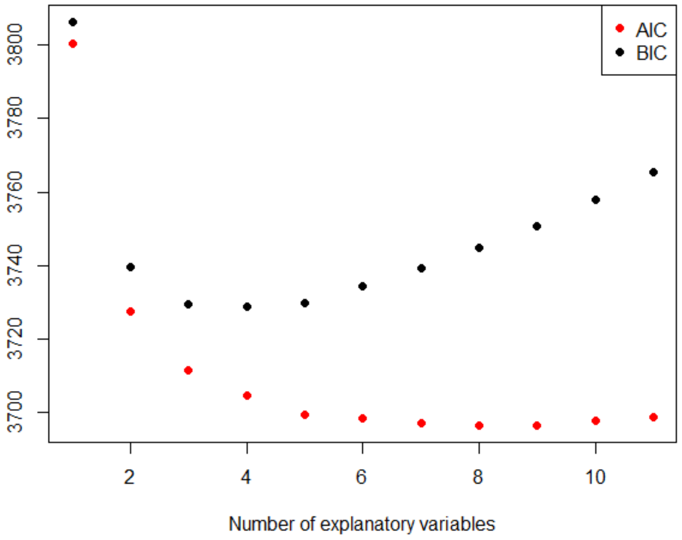

Figure (9) Minimum AIC and BIC values for every number of explanatory variables

3 Interpretation

Our final model is:

$$M_5: \qquad \eta = \beta_0 + \beta_1 hisp + \beta_2 educyrs + \beta_3 new\, hhincome +$$

$$\beta_4 new\, lincome + \beta_5 I(chronic \geq 1) \quad (1)$$

$$\eta = x^T\beta = \log(\frac{\pi}{1-\pi}) \quad (2) \qquad \pi = \frac{exp(\eta)}{1+exp(\eta)} = \frac{exp(x^T\beta)}{1+exp(x^T\beta)} \quad (3)$$

Let x, x' be two vectors of the explanatory values which share the same values but $x_{hisp} = 0$, $x'_{hisp} = 1$. Then from (2) we can see that the estimated change in the log odds (where a success corresponds to the purchase of private insurance) is -0.58, with 95% confidence interval $(-0.989, -0.178)$. Similarly, having at least 1 chronic disease causes an estimated increase in the log odds of 0.32 and every extra 1 year of education increases the estimated log odds by 0.058. The confidence intervals are given in the table below.

Estimated 95% confidence intervals and coefficients for M_5:

```
                lower  est beta    upper
 hisp         -0.9889   -0.5835  -0.1782
 educyrs       0.0277    0.0576   0.0875
 chronic       0.0823    0.3204   0.5586
 new lincome   1.8861    2.2581   2.6302
 new hhincome -0.0470   -0.0379  -0.0288
```

Suppose that we have two vectors x, x' such that $x_{hisp} = x'_{hisp}$, $x_{educyrs} = x'_{educyrs}$, $x_{chronic} = x'_{chronic}$
Let $x_{hhincome} = \alpha \leq 97.4 - 1 = 96.4$, $x'_{hhincome} = x_{hhincome} + 1$.
That is, there is a \$1000 difference in income and the values are below the \$97.4k threshold (so that $\beta_{hhincome}$, $\beta_{lincome}$ come into play). Then:

$$\begin{aligned} Log(\hat{O}')–Log(\hat{O}) &= 2.258(log(\alpha+1)–log(\alpha)) - 0.0379 \\ &= 2.258 \cdot log(1+1/\alpha) - 0.0379 \;\; (*) \end{aligned} \quad (1)$$

Hence the log odds increase by $(*)$. Note that this value depends on income α because we have $log(income)$ as explanatory variable (and so terms do not cancel out).

In addition, since we decided to exclude interaction terms in the first place, the effects of other variables do not depend on *hisp*.

4 Appendix

```
df = hins

#BARCHARTS FOR BINARY VARIABLES
par(mfrow=c(1,5))
barplot(prop.table(table(df$female, df$privins), 1)[,2], beside=TRUE,
        col=c("skyblue","steelblue"), xlab=’Female’,
        ylab=’Prop of private ins’)
barplot(prop.table(table(df$white, df$privins), 1)[,2], beside=TRUE,
        col=c("skyblue","steelblue"), xlab=’White’,
        ylab=’Prop of private ins’)
barplot(prop.table(table(df$hisp, df$privins), 1)[,2], beside=TRUE,
        col=c("skyblue","steelblue"), xlab=’Hispanic’,
         ylab=’Prop of private ins’)
barplot(prop.table(table(df$married, df$privins), 1)[,2], beside=TRUE,
        col=c("skyblue","steelblue"), xlab=’Married’,
        ylab=’Prop of private ins’)
barplot(prop.table(table(df$retired, df$privins), 1)[,2], beside=TRUE,
        col=c("skyblue","steelblue"), xlab=’Retired’,
         ylab=’Prop of private ins’)

options(digits = 3)
prop.table(table(df$adl, df$privins), 1)
table(df$adl, df$privins)

#TABLES FOR DISCRETE VARIABLES
prop.table(table(df$chronic, df$privins), 1)
table(df$chronic, df$privins)
table(df$educyrs, df$privins)
prop.table(table(df$educyrs, df$privins),1)
summary(df$ed)

#PLOTS FOR DISCRETE VARIABLES
par(mfrow = c(1,3))
plot(prop.table(table(df$adl, df$privins), 1)[,2] ˜ c(0:5), pch = 16,
     xlab = ’Limitations’, ylab = ’Prop of private insurance’)
plot(prop.table(table(df$chronic, df$privins), 1)[,2] ˜ c(0:8),
 pch = 16,
     xlab = ’Chronic Conditions’, ylab = ’Prop of private insurance’)
plot(prop.table(table(df$educyrs, df$privins),1)[,2] ˜ c(3:17),
     pch = 16, xlab = ’Education Years’,
     ylab = ’Prop of private insurance’)
```

```
#BOXPLOTS FOR CONTINUOUS VARIABLES
summary(df$hhincome)
summary(df$age)
par(mfrow = c(1,3))
boxplot(df$age ~ df$privins  ,xlab = "Private Insurance", ylab = 'Age')
boxplot(df$hhincome ~ df$privins  ,xlab = "Private Insurance",
 ylab = 'Household Income')
boxplot(df$lincome ~ df$privins  ,xlab = "Private Insurance",
ylab = 'log(Household Income)')

#ADD NEW COLS TO DATASET
df2 = within(hins, {
  adl_f = factor(adl)
  chronic_f = factor(chronic)
  adl_b = 1*(adl >= 1)
  chronic_b = 1*(chronic >= 1)
  adl_bb = 1*(adl>=2)

  adl2 = 1*(adl >=2)
  chronic2 = 1*(chronic >= 2)
  chronic3 = 1*(chronic >= 3)
  chronic4 = 1*(chronic >= 4)})

#SECTION 2
#INITIAL GLMS
m0 = glm(privins ~ age + female + white + hisp + married + educyrs +
hhincome + lincome + retired + adl_f +
 chronic_f, data = df2, family = "binomial")
summary(m0)
m1 = glm(privins ~ age + female + white + hisp + married + educyrs +
 hhincome + lincome + retired + adl +
chronic, data = df2, family = "binomial")
summary(m1)
m2 = glm(privins ~ age + female + white + hisp + married + educyrs +
 hhincome + lincome + retired + adl_b +
 chronic_b, data = df2, family = "binomial")
summary(m2)

#MODIFYING DATASET:
df3 <-within(df, {
  adl <- 1*(adl>=1)
  chronic <- 1*(chronic >=1)})
```

```
#BRUTE FORCE OPTIMAL BIC, AIC
mod_df3 <- within(df3, {
  y <- privins
  privins <- NULL
})
library(leaps)
library(bestglm)
bic_best_glm <-
  bestglm(Xy = mod_df3,
          family = binomial,
          IC = "BIC",
          method = "exhaustive")
aic_best_glm <-
  bestglm(Xy = mod_df3,
          family = binomial,
          IC = "AIC",
          method = "exhaustive")
bic_best_glm$BestModels
aic_best_glm$BestModels

#NEW SIMPLIFIED MODELS
m7 = glm(privins ˜ hisp + educyrs  + lincome+ hhincome,
 data = df3,
family = "binomial" )
summary(m7)
m8 = glm(privins ˜ ., data = df3, family = "binomial" )
summary(m8)

vec = numeric(31)
x = numeric(31)
for (i in 1:31){
  t = sum(df$privins[(100*(i-1)):(100*i - 1)])/100
  vec[i] = t
  x[i] = 100*i
}

#DIAGNOSTICS FOR M7
par(mfrow = c(1,1))
plot(fitted(m7), xlab = ’Obs sorted by hhincome’,
ylab=’Fitted Values’)
points(x,vec, pch = 19, col = ’red’, cex = 1.5)
par(mfrow = c(1,2))
```

```
p <- m7$rank
n <- nrow(model.frame(m7))
plot(influence(m7)$hat/(p/n), ylab='Leverage / (p/n)',
xlab = 'Obs sorted by hhincome')
plot(cooks.distance(m7), xlab = 'Obs sorted by hhincome',
ylab = "Cook's distance")
abline(h=8/(n-2*p),lty=2)

#MODIFIYING INCOME OBSERVATIONS
thold = mean(df3$hhincome)+ 1.5*sqrt(var(df3$hhincome))
df4 <-within(df3, {
  n_hhincome = hhincome*(hhincome <= thold)+
  thold*(hhincome > thold)
  n_lincome = log(n_hhincome)})

m9 <- glm(privins~ .-n_hhincome - n_lincome , data = df4,
 family = "binomial")
summary(m9)
m10 = glm(privins~ . -hhincome - lincome, data = df4,
family = "binomial"  )
summary(m10)

par(mfrow = c(1,1))
plot(fitted(m10), xlab = 'Obs sorted by hhincome',
ylab='Fitted Values')
points(x,vec, pch = 19, col = 'red', cex = 1.5)

df5 <-within(df4, {
  n_hhincome = hhincome*(hhincome <= thold)+
  thold*(hhincome > thold)
  n_lincome = log(n_hhincome)
  hhincome = NULL
  lincome = NULL})

mod_df5 <- within(df5, {
  y <- privins
  privins <- NULL
})

#BRUTE FORCE BEST BIC, AIC
bbic_best_glm <-
  bestglm(Xy = mod_df5,
          family = binomial,
          IC = "BIC",
          method = "exhaustive")
aaic_best_glm <-
  bestglm(Xy = mod_df5,
          family = binomial,
          IC = "AIC",
          method = "exhaustive")
bbic_best_glm$BestModels
aaic_best_glm$BestModels

#CHISQ TESTS
m11 = glm(privins˜ ., data = df5, family = "binomial"  )
summary(m11) #res dev 3676.8
m12 = glm(privins˜ . - female - white - married - retired,
data = df5, family = "binomial"  )
summary(m12) #res dev 3683.0
1 - pchisq(6.2, df = 4)

#DIAGNOSTICS PLOTS
par(mfrow = c(1,1))
bic_vec = bbic_best_glm$Subsets[,14]
aic_vec = aaic_best_glm$Subsets[,14]
bic_vec = bic_vec[2:12]
aic_vec = aic_vec[2:12]
y = c(bic_vec, aic_vec)
x = rep(c(1:11),2)
plot(x,y, pch = 16, col = c(rep(c(1),11), rep(c(2),11)), ylab = '',
     xlab = 'Number of explanatory variables')
legend("topright", c( "AIC", "BIC") , col = c(2,1) , pch = 19)

par(mfrow = c(1,1))
points(x,vec, pch = 19, col = 'red', cex = 1.5)
par(mfrow = c(1,2))
p <- m14$rank
n <- nrow(model.frame(m14))
plot(influence(m14)$hat/(p/n), ylab='Leverage / (p/n)',
 xlab = 'Obs sorted by hhincome')
plot(cooks.distance(m14),
 xlab = 'Obs sorted by hhincome', ylab = "Cook's distance")
abline(h=8/(n-2*p),lty=2)
plot(rstandard(m14), ylab = "Standardised deviance residuals",
     xlab = "Obs sorted by hhincome")
```

```
v = which(cooks.distance(m14) > 0.006)
df5[v,]
w = which(influence(m14)$hat/(p/n) > 4)
df5[w,]

#ESTIMATED CONFIDENCE INTERVALS
options(digits = 2)
beta <- summary(m14)$coef[2:6,1]
se <- summary(m14)$coef[2:6,2]
cval <- qnorm(0.975)
lower <- beta-cval*se
upper <- beta+cval*se
ci95 <- cbind(lower,beta,upper)
```