HRS

Jean Clark

November 30, 2020

Contents

1 Descriptives statistics

This database contains information on the individuals that have subscribed to a private insurance (1) or a public insurance (0). The age, ethnicity, sex, level of education (years spent in school in total), status (married or not), health status, intensity of a possible chronic disease, the number of limitations medicaly imposed on daily activities, the retirement status, the retirement status of their spouse and the household income are represented.

Table 1: Data

age	hisp	white	female	educyear	married	excel	vegood	good	fair	poor
62	0	0	1	12	0	0	0	0	1	0
59	0	1	1	12	0	0	0	0	1	0
60	0	0	0	13	0	0	1	0	0	0
62	0	1	1	10	0	0	0	0	1	0
54	0	1	1	9	0	0	0	0	0	1
62	0	1	1	12	1	0	1	0	0	0

chronic	adl	retire	sretire	hhincome	ins	hstatusg
3	0	0	0	0	0	0
1	3	0	0	0	0	0
2	0	1	0	0	0	1
4	3	0	0	0	0	0
6	0	0	0	0	0	0
0	0	1	1	0	0	1

We can already make a few comments on that database:

- For the variable *hhincome*, we can't precisely estimate the income of the individual because of other potential incomes in the household. Therefore, we have to be careful when we will make assumptions with that variable.
- *hstatusg* is nearly the same variable as all the binary variables showing the health status, when *hstatusg* is equal to 1, either $excel, vegood$ or *good* are equal to 1, when *hstatusg* is equal to 0, either $fair$ or *poor* are equal to 1.
- We may have some issues with the *sretire* variable as it doesn't make any difference between someone that has a working spouse and someone who is single.

1.1 Quantitative variables

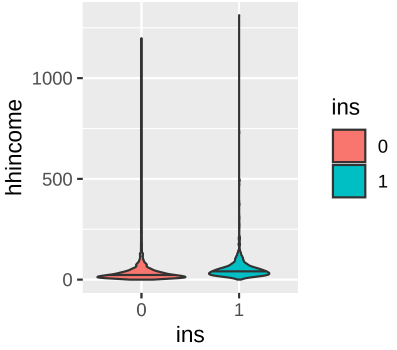
Table 2: Summary: Quantitative variables

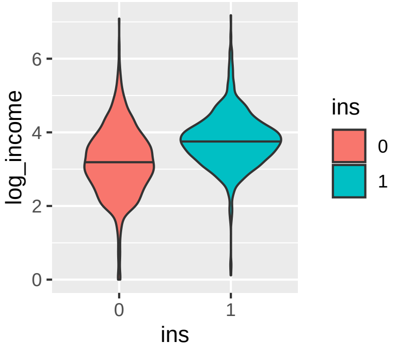
	age	adl	chronic	educyear	hhincome
Mean	66.91	0.30	2.06	11.9	45.26
Sd	3.68	0.83	1.42	3.3	64.34
Minimum	52.00	0.00	0.00	0.0	0.00
Q1	65.00	0.00	1.00	10.0	17.00
Median	67.00	0.00	2.00	12.0	31.10
Q3	69.00	0.00	3.00	14.0	52.80
Maximum	86.00	5.00	8.00	17.0	1312.12

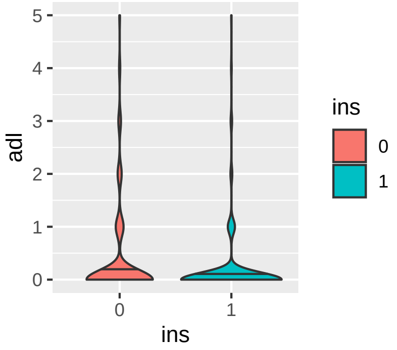
We can see that the variable “age” doesn’t vary to much, all the people represented in that database are quite old. The minimum being 52 and the maximum being 86 we can safely say that they are near retirement or retired.

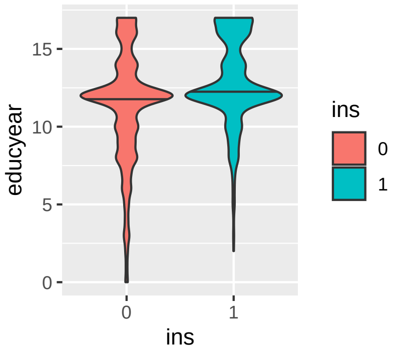
On the other hand the household income varies a lot from individual to individual. 75% of individuals have a household income below 52.8 thousand dollars, the top 25% have a household income between 52.8 thousand dollars and 1312.12 thousand dollars! To linearise this relationship, we have added a variable which is the Neperian Logarithm of the household income.

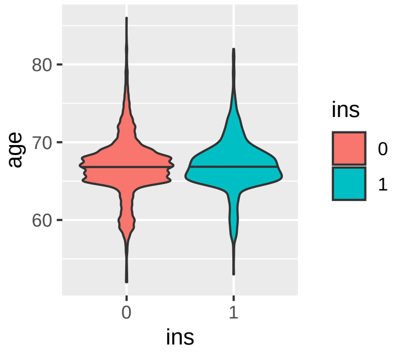
1.1.1 Comparison (ins=1 or ins=0)

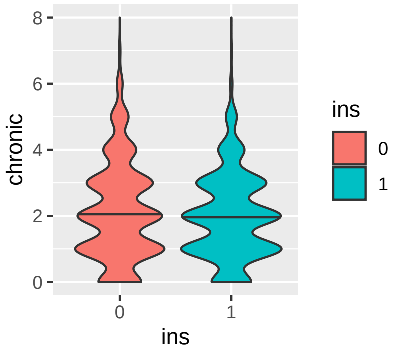












We can see that differences between the two groups are minimal. The household income and the number of years spent in school are higher for the insured with private insurance than the others.

1.2 Qualitative variables

The majority of the individuals on this database (variables taken individualy) are white (82%), male (52%), married (73%), retired (62%), not insured with private insurance (61%) and healthy (70%).

Table 3: Summary: Binary variables

	hisp	white	female	married	retire	sretire	ins	hstatusg
0	2973.00	575.00	1674.00	856.00	1203.00	1961.00	1965.00	947.0
1	233.00	2631.00	1532.00	2350.00	2003.00	1245.00	1241.00	2259.0
Frequency 0	0.93	0.18	0.52	0.27	0.38	0.61	0.61	0.3
Frequency 1	0.07	0.82	0.48	0.73	0.62	0.39	0.39	0.7

	excel	vegood	good	fair	poor
0	2844.00	2304.00	2211.00	2549.0	2916.00
1	362.00	902.00	995.00	657.0	290.00
Frequency 0	0.89	0.72	0.69	0.8	0.91
Frequency 1	0.11	0.28	0.31	0.2	0.09

1.2.1 Comparison

We are going to compare the differences between the two groups by comparing how the proportions might vary depending on its affiliation.

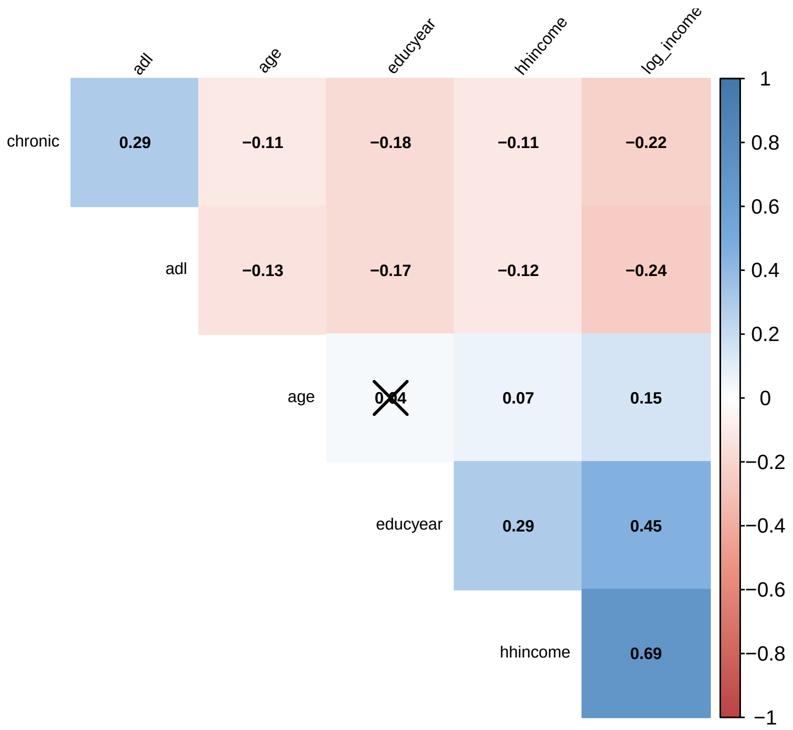
	hisp	white	female	married	retire	sretire	hstatusg
Is not ins 0	0.90	0.20	0.49	0.32	0.41	0.64	0.35
Is not ins 1	0.10	0.80	0.51	0.68	0.59	0.36	0.65
Is ins 0	0.97	0.14	0.57	0.19	0.33	0.56	0.22
Is ins 1	0.03	0.86	0.43	0.81	0.67	0.44	0.78

There are even less hispanics represented in the insured group than in the other group, there are more whites represented in the insured group than in the other group. There are less women represented in the insured group than the other group which means that men are more insured than women. There are relatively more married individuals insured than singles. Those with private insurance are relatively healthier than those without.

	excel	vegood	good	fair	poor
Is not ins 0	0.90	0.75	0.70	0.77	0.89
Is not ins 1	0.10	0.25	0.30	0.23	0.11
Is ins 0	0.87	0.67	0.68	0.84	0.95
Is ins 1	0.13	0.33	0.32	0.16	0.05

1.3 Correlation

Let's check if there is any correlation between our quantitative variables.



All the correlation coefficients are significant except for *age* and *educyear*. The number of years spent in school and the household income are positively correlated. These variables are both negatively correlated with the intensity of a possible chronic disease and the number of limitations on daily activities. By the way, *chronic* and *adl* are strongly positively correlated.

2 Model : general model and log-likelihood function

As we want to predict the probability of an individual being subscribed to a private insurance or not, we may use two types of general models and see which one works best. $Ins_i = 1$ with probability Pi, 0 with probability $1 - Pi$.

In most applications of binary response models, the primary goal is to explain the effects of the exogenous variables on the response probability $Prob(Ins_i = 1|Xi)$.

We let $Pi = Prob(Ins_i = 1|Xi) = F(Xi)$.

The logit and probit models are some of the most effective to predict a binary response.

- The logit model: $F(Xi\beta) = \Lambda(Xi\beta) = \frac{exp(Xi\beta)}{1+exp(Xi\beta)}$
- The probit model: $F(Xi\beta) = \phi(Xi\beta) = \int_{-\infty}^{Xi\beta} \frac{1}{\sqrt{2\pi}} exp(\frac{-z^2}{2})$
- With a log likelihood function equal to : $lnL(Y, \beta) = \sum_{i:Y_i=1} Y_i ln[F(X_i\beta)] + \sum_{i:Y_i=0} ln[1 - F(X_i\beta)]$.

We want an estimation of the log-likelihood that is as close as possible to zero, the closer to 0 the better the prediction as the parameters are correctly estimated.

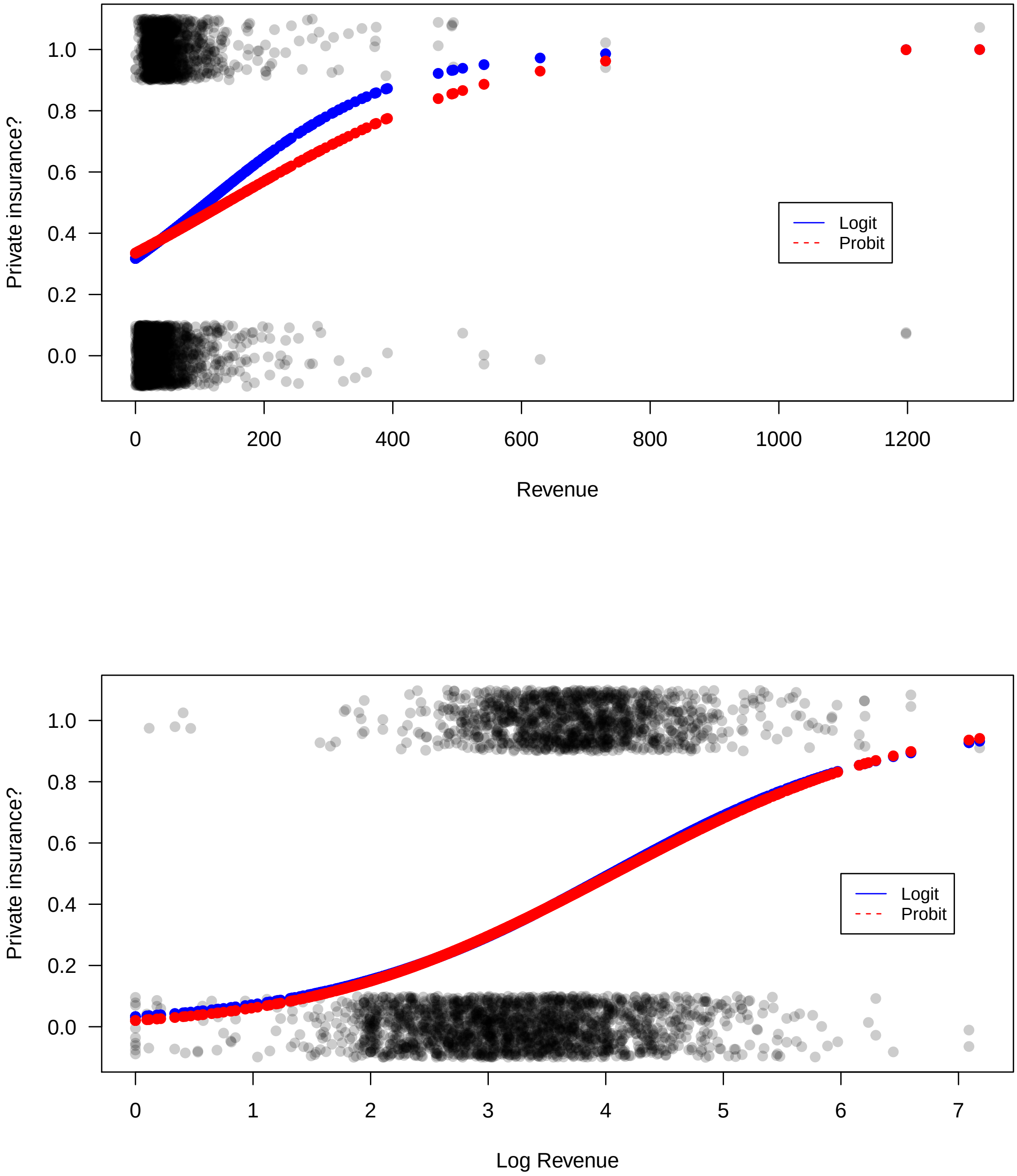
We already have a few ideas on what kind of variables we want to use in our models. First we will try with a full model with all possible variables and then take out one by one any variable that is insignificant until we have the best model:

- $logit(ins_i) = \alpha + \beta_1 age_i + \beta_2 hisp_i + \beta_3 white_i + \beta_4 female_i + \beta_5 educyear_i + \beta_6 married_i + \beta_7 excel_i + \beta_1 vegood_i + \beta_1 good_i + \beta_1 fair_i + \beta_1 poor_i + \beta_1 chronic_i + \beta_1 adl_i + \beta_1 retir_i + \beta_1 sretire_i + \beta_1 hhincome_i + \beta_1 hstatusg_i + \mu_i$

$hstatus$ is pretty much the same variable as $excel, vegood, good, fair, poor$ as when $hstatus$ is equal to 1 it means that the individual is healthy, the opposite being when it's equal to 0. We can maybe have a second model where we take $hstatus$ or $excel, vegood, good, fair, poor$ out as we might have some colinearity issues.

We also might want to consider a model with the log_income variable instead of $hhincome$ for reasons previously mentionned. Let's take a look at the difference between the two models if we decide to use only the household income and the log_ household income as exogenous variables.

Household income vs log household income



% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 20:19:27

Table 4: Models

	Dependent variable:			
	ins			
	logistic	*probit*	*logistic*	*probit*
	(1)	(2)	(3)	(4)
hhincome	0.007^{***} (0.001)	0.003^{***} (0.0005)		
log_income			0.836^{***} (0.050)	0.503^{***} (0.029)
Constant	-0.765^{***} (0.052)	-0.426^{***} (0.030)	-3.381^{***} (0.182)	-2.047^{***} (0.105)
Observations	3,206	3,206	3,206	3,206
Log Likelihood	−2,096.994	−2,104.209	−1,971.580	−1,971.056
Akaike Inf. Crit.	4,197.988	4,212.417	3,947.160	3,946.112
Note:				*p<0.1; **p<0.05; ***p<0.01

The variable log_income gives some much better results, we will use it from now on.

Let’s start looking at these first models.

3 Estimation ((Logit, Probit)

We will choose our models by using three measures. The Log-likelihood estimate, the Akaike Information Criterion (AIC) and the MacFadden R-Squared.

- The AIC takes into account the number of variables used for our models which makes it a good measure to compare models ($AIC = 2lnL + 2k$).
- The MacFadden R-squared doesn't give much information, but we can compare them between models as a MacFadden R-squared equal to one means that the log-likelihood is equal to 0. $R^2 = 1 - \frac{lnL(Y,B_{est})}{lnL(Y_{tot})}$.

For better results we will use the adjusted MacFadden so we can take into account the number of variables. For the Log-likelihood estimate and the AIC, the closer these values are to zero, the better the model.

Estimated coefficients do not quantify the influence of the variables on the probability that the *ins* variable takes on the value one. Estimated coefficients are parameters of the latent model.

3.1 Logit

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 20:19:27

Table 5: Models

	Dependent variable:		
	ins		
	(1)	(2)	(3)
age	-0.022^{*}	-0.019^{*}	-0.022^{*}
	(0.012)	(0.012)	(0.012)
hisp	-0.652^{***}	-0.653^{***}	-0.650^{***}
	(0.200)	(0.199)	(0.199)
white	-0.075	-0.066	-0.075
	(0.110)	(0.110)	(0.110)
female	-0.086	-0.076	-0.086
	(0.090)	(0.089)	(0.085)
educyear	0.069^{***}	0.073^{***}	0.069^{***}
	(0.015)	(0.015)	(0.015)
married	0.150	0.144	0.146
	(0.122)	(0.122)	(0.106)
excel	0.219		
	(0.218)		
vegood	0.300		
	(0.192)		
good	0.262		
	(0.182)		
fair	0.113		
	(0.181)		
poor			
chronic	0.077^{**}	0.060^{**}	0.076^{**}
	(0.032)	(0.030)	(0.031)
adl	-0.144^{**}	-0.181^{***}	-0.153^{**}
	(0.063)	(0.060)	(0.061)
retire	0.201^{**}	0.204^{**}	0.201^{**}
	(0.088)	(0.088)	(0.087)
sretire	-0.007	0.001	
	(0.094)	(0.094)	
log_income	0.636^{***}	0.646^{***}	0.637^{***}
	(0.062)	(0.062)	(0.062)
hstatusg			0.180^{*}
			(0.105)
Constant	-2.465^{***}	-2.488^{***}	-2.393^{***}
	(0.829)	(0.815)	(0.810)

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 20:19:27

Table 6: Models

	Dependent variable:		
	ins		
	(1)	(2)	(3)
age	-0.019^{*}	-0.017	
	(0.012)	(0.012)	
hisp	-0.644^{***}	-0.647^{***}	-0.651^{***}
	(0.199)	(0.199)	(0.198)
educyear	0.065^{***}	0.068^{***}	0.069^{***}
	(0.015)	(0.015)	(0.015)
chronic	0.077^{**}	0.062^{**}	0.064^{**}
	(0.031)	(0.030)	(0.030)
adl	-0.152^{**}	-0.177^{***}	-0.171^{***}
	(0.061)	(0.059)	(0.059)
retire	0.227^{***}	0.228^{***}	0.192^{**}
	(0.085)	(0.085)	(0.082)
log_income	0.678^{***}	0.686^{***}	0.679^{***}
	(0.056)	(0.056)	(0.055)
hstatusg	0.165		
	(0.105)		
Constant	-2.681^{***}	-2.745^{***}	-3.842^{***}
	(0.787)	(0.785)	(0.245)
Observations	3,206	3,206	3,206
Log Likelihood	−1,935.603	−1,936.848	−1,937.928
Akaike Inf. Crit.	3,889.206	3,889.696	3,889.856
Note:			*p<0.1; **p<0.05; ***p<0.01

The last five models give some similar log likelihood and AIC results. They are the best models so far. Only the last one have all its variables significant.

Let's think about these models, *hisp* is always a significant variable. It would mean that the probability of being insured decreases if you are hispanic, in my opinion it is a too simple to assume that your ethnical affiliation would impact that much on the fact that you have a private insurance or not. On the other hand, it seems a bit odd that the variable *married* isn't significant in all the best models so far as the household income may directly depend on the fact that the individual is married or not.

For both variables we can try to put them in interaction with log_income, it seems more logical that *hisp* would have so much impact on the model if *hisp* and the household income would be linked, meaning that hispanic people have a smaller household income as white people and that it affects the probability of being insured or not.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 20:19:28

Table 7: Models

	Dependent variable:		
	ins		
	(1)	(2)	(3)
age	−0.020*	−0.021*	−0.020*
	(0.012)	(0.012)	(0.012)
hisp	−4.590***	−0.664***	−4.649***
	(1.197)	(0.199)	(1.205)
married	0.171*	1.152***	1.159***
	(0.103)	(0.429)	(0.430)
educyear	0.068***	0.069***	0.069***
	(0.015)	(0.015)	(0.015)
chronic	0.075**	0.078**	0.078**
	(0.031)	(0.031)	(0.031)
adl	−0.150**	−0.150**	−0.147**
	(0.062)	(0.062)	(0.062)
retire	0.219**	0.209**	0.211**
	(0.086)	(0.086)	(0.086)
log_income	0.600***	0.867***	0.829***
	(0.062)	(0.117)	(0.117)
hstatusg	0.164	0.164	0.158
	(0.105)	(0.105)	(0.105)
hisp:log_income	1.213***		1.224***
	(0.353)		(0.354)
married:log_income		−0.308**	−0.308**
		(0.130)	(0.130)
Constant	−2.538***	−3.248***	−3.191***
	(0.788)	(0.836)	(0.838)
Observations	3,206	3,206	3,206
Log Likelihood	−1,926.836	−1,931.380	−1,923.927
Akaike Inf. Crit.	3,875.671	3,884.759	3,871.854
Note:		*p<0.1; **p<0.05; ***p<0.01	

The last model with both variables in interaction gives the best results. The *hisp* variable in interaction with the *logincome* gives a positive coefficient. The only issue with that is that the variable *hstatusg* isn't significant. Its p-value is equal to 0.14 which is quite high. I would suggest testing our model without that variable.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 20:19:28

Table 8: Models

	Dependent variable:	
	ins	
	(1)	(2)
age	−0.018	
	(0.012)	
hisp	−4.679***	−4.713***
	(1.206)	(1.205)
married	1.171***	1.140***
	(0.429)	(0.430)
educyear	0.073***	0.073***
	(0.015)	(0.015)
chronic	0.063**	0.066**
	(0.030)	(0.030)
adl	−0.171***	−0.165***
	(0.060)	(0.060)
retire	0.212**	0.175**
	(0.086)	(0.082)
log_income	0.841***	0.831***
	(0.117)	(0.117)
married:log_income	−0.313**	−0.307**
	(0.130)	(0.130)
hisp:log_income	1.232***	1.242***
	(0.355)	(0.354)
Constant	−3.268***	−4.427***
	(0.836)	(0.395)
Observations	3,206	3,206
Log Likelihood	−1,925.046	−1,926.284
Akaike Inf. Crit.	3,872.091	3,872.568
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 9: MacFadden R-squared adjusted for the 5 last models

x
0.094
0.092
0.095
0.095
0.095

For the last 5 models, the MacFadden R-squared ajusted are very similar. When you take *hstatusg* out, *age* becomes insignificant. Our final logit model is then :

- $logit(ins_i) = \alpha + \beta_1 hisp_i + \beta_2 educyear_i + \beta_3 married_i + \beta_4 chronic_i + \beta_5 adl_i + \beta_6 retir_i + \beta_7 logincome_i + \beta_8(logincome_i * married_i) + \beta_9(logincome_i * hisp_i) + \mu_i$

3.2 Probit

We build the same modelsas before and it gives some different results, the best probit model contains the following variables: *age*, *hisp*,*educyear*, *chronic*, *adl*, *retire*, *logincome* and *hstatusg*.

We can try to put the same variables as before in interaction.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 20:19:28

Table 10: Models

	Dependent variable:					
	ins					
	(1)	(2)	(3)	(4)	(5)	(6)
age	-0.012^{*}	-0.010		-0.013^{*}	-0.012^{*}	-0.012^{*}
	(0.007)	(0.007)		(0.007)	(0.007)	(0.007)
hisp	-0.388^{***}	-0.389^{***}	-0.390^{***}	-0.401^{***}	-2.457^{***}	-2.511^{***}
	(0.113)	(0.113)	(0.113)	(0.113)	(0.632)	(0.639)
married				0.597^{**}	0.107^{*}	0.613^{**}
				(0.244)	(0.062)	(0.246)
educyear	0.040^{***}	0.043^{***}	0.043^{***}	0.043^{***}	0.043^{***}	0.044^{***}
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
chronic	0.046^{**}	0.037^{**}	0.038^{**}	0.047^{**}	0.046^{**}	0.047^{**}
	(0.019)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)
adl	-0.093^{***}	-0.109^{***}	-0.106^{***}	-0.092^{**}	-0.091^{**}	-0.090^{**}
	(0.036)	(0.034)	(0.034)	(0.036)	(0.036)	(0.036)
retire	0.137^{***}	0.137^{***}	0.116^{**}	0.125^{**}	0.132^{**}	0.128^{**}
	(0.052)	(0.052)	(0.050)	(0.052)	(0.052)	(0.052)
log_income	0.410^{***}	0.414^{***}	0.410^{***}	0.496^{***}	0.363^{***}	0.477^{***}
	(0.033)	(0.033)	(0.033)	(0.067)	(0.037)	(0.067)
hstatusg	0.106^{*}			0.107^{*}	0.106^{*}	0.103
	(0.063)			(0.063)	(0.064)	(0.064)
married:log_income				-0.154^{**}		-0.158^{**}
				(0.075)		(0.075)
hisp:log_income					0.648^{***}	0.661^{***}
					(0.191)	(0.193)
Constant	-1.649^{***}	-1.689^{***}	-2.344^{***}	-1.906^{***}	-1.571^{***}	-1.887^{***}
	(0.474)	(0.473)	(0.144)	(0.499)	(0.475)	(0.500)
Observations	3,206	3,206	3,206	3,206	3,206	3,206
Log Likelihood	−1,932.906	−1,934.294	−1,935.334	−1,929.183	−1,924.620	−1,922.248
Akaike Inf. Crit.	3,883.812	3,884.588	3,884.669	3,880.366	3,871.240	3,868.496
Note:						*p<0.1; **p<0.05; ***p<0.01

The models with variables interacting are much better than those without, but, same as the logit model, *hstatusg* becomes insignificant.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 20:19:29

Table 11: Models

	Dependent variable:	
	ins	
	(1)	(2)
age	-0.011	
	(0.007)	
hisp	-2.528^{***}	-2.539^{***}
	(0.640)	(0.638)
married	0.616^{**}	0.597^{**}
	(0.246)	(0.246)
educyear	0.046^{***}	0.046^{***}
	(0.009)	(0.009)
chronic	0.037^{**}	0.039^{**}
	(0.018)	(0.018)
adl	-0.105^{***}	-0.102^{***}
	(0.035)	(0.035)
retire	0.128^{**}	0.106^{**}
	(0.052)	(0.050)
log_income	0.483^{***}	0.477^{***}
	(0.067)	(0.067)
married:log_income	-0.160^{**}	-0.156^{**}
	(0.075)	(0.075)
hisp:log_income	0.665^{***}	0.668^{***}
	(0.193)	(0.192)
Constant	-1.932^{***}	-2.629^{***}
	(0.499)	(0.224)
Observations	3,206	3,206
Log Likelihood	−1,923.553	−1,924.753
Akaike Inf. Crit.	3,869.106	3,869.506
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 12: MacFadden R-squared adjusted for the 5 last models

x
0.095
0.093
0.096
0.096
0.096

For the last 5 models, the MacFadden R-squared ajusted are very similar, the best though is the last model's one. When you take *hstatusg* out, *age* becomes insignificant. Our final probit model with variables interacting is then :

- $probit(ins_i) = \alpha + \beta_1 hisp_i + \beta_2 educyear_i + \beta_3 married_i + \beta_4 chronic_i + \beta_5 adl_i + \beta_6 retir_i + \beta_7 logincome_i + \beta_8 (logincome_i * married_i) + \beta_9 (logincome_i * hisp_i) + \mu_i$

Same model than the logit model, quite logical in fact. The probit model though gives some better results than its logit counterpart ($AIC_{probit} < AIC_{logit}$).

4 Model comparison

We will use the best model without variables in interaction and the best model with variables in interaction and compare them with the probit and the logit estimations.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 20:19:29

Table 13: Models

	Dependent variable:			
	ins			
	logistic		*probit*	
	(1)	(2)	(3)	(4)
hisp	-0.651^{***}	-4.713^{***}	-0.390^{***}	-2.539^{***}
	(0.198)	(1.205)	(0.113)	(0.638)
married		1.140^{***}		0.597^{**}
		(0.430)		(0.246)
educyear	0.069^{***}	0.073^{***}	0.043^{***}	0.046^{***}
	(0.015)	(0.015)	(0.009)	(0.009)
chronic	0.064^{**}	0.066^{**}	0.038^{**}	0.039^{**}
	(0.030)	(0.030)	(0.018)	(0.018)
adl	-0.171^{***}	-0.165^{***}	-0.106^{***}	-0.102^{***}
	(0.059)	(0.060)	(0.034)	(0.035)
retire	0.192^{**}	0.175^{**}	0.116^{**}	0.106^{**}
	(0.082)	(0.082)	(0.050)	(0.050)
log_income	0.679^{***}	0.831^{***}	0.410^{***}	0.477^{***}
	(0.055)	(0.117)	(0.033)	(0.067)
married:log_income		-0.307^{**}		-0.156^{**}
		(0.130)		(0.075)
hisp:log_income		1.242^{***}		0.668^{***}
		(0.354)		(0.192)
Constant	-3.842^{***}	-4.427^{***}	-2.344^{***}	-2.629^{***}
	(0.245)	(0.395)	(0.144)	(0.224)
Observations	3,206	3,206	3,206	3,206
Log Likelihood	−1,937.928	−1,926.284	−1,935.334	−1,924.753
Akaike Inf. Crit.	3,889.856	3,872.568	3,884.669	3,869.506
Note:			*p<0.1; **p<0.05; ***p<0.01	

The models with some variables in interaction are the best models. The best models are definitely the two with variables in interaction.

5 Hypothesis and specifications tests

We want to check both models to see if the coefficients are individually significant, simultaneously significant. We might take a look at the extreme observations, high-leverage points…

Wald tests are used to check if coefficients are individually significant, they are implicitly made within the estimation of the coefficients of our models. As in linear regression, the stars show the significance level of each coefficient.

We will check for both models the simultaneous significance of the coefficients by applying an ANOVA test comparing our models with a model without any coefficients.

5.1 Logit

ANOVA test

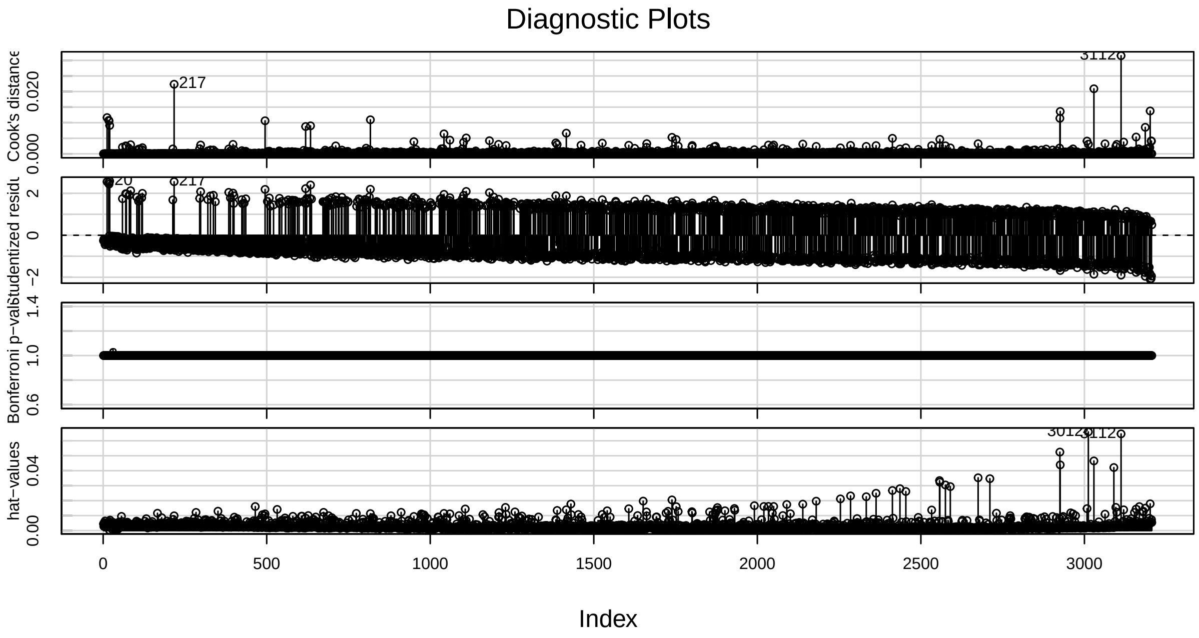
```
##
## Call:
## glm(formula = ins ~ 1, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9895  -0.9895  -0.9895   1.3778   1.3778
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.45957    0.03626  -12.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4279.5  on 3205  degrees of freedom
## Residual deviance: 4279.5  on 3205  degrees of freedom
## AIC: 4281.5
##
## Number of Fisher Scoring iterations: 4
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
3205	4279.542	NA	NA	NA
3196	3852.568	9	426.9745	0

```
## [1] 0
```

We can reject the simultaneous nullity of the exogenous variable coefficients (at the 5% level).

Extreme observations and high-leverage points


Diagnostic Plots
Cook's distance
0.000
0.020
217
3112
Studentized residuals
-2
0
2
20
217
Bonferroni p-value
0.6
1.0
1.4
hat-values
0.00
0.04
3012
3112
0
500
1000
1500
2000
2500
3000
Index

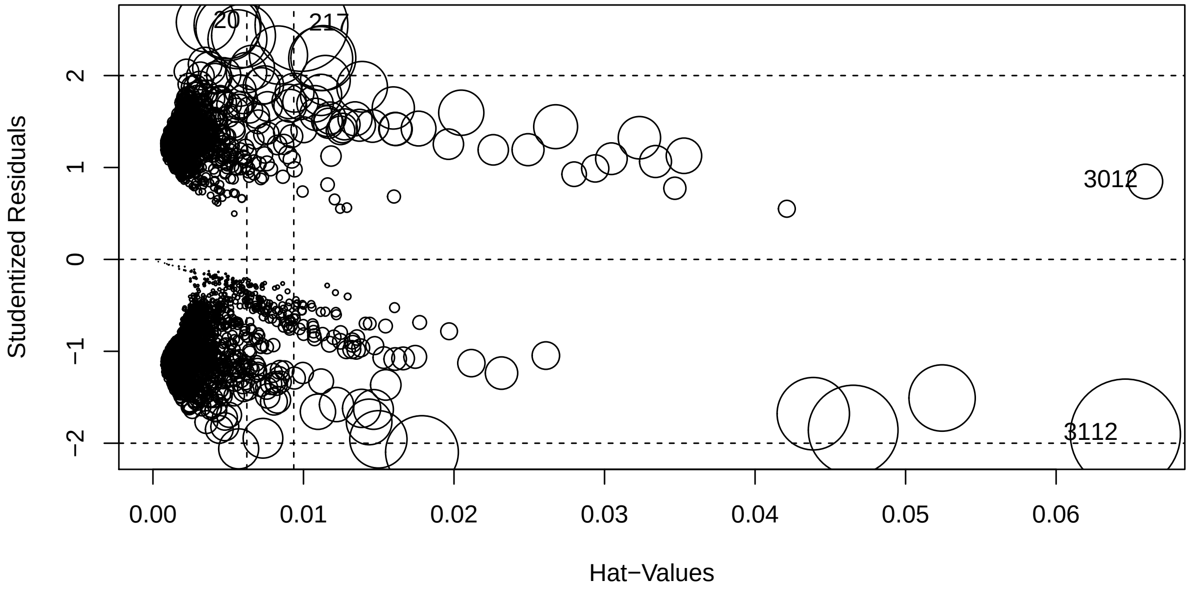

Studentized Residuals
-2
-1
0
1
2
20
217
3012
3112
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Hat-Values

Table 14: Extreme observations

	StudRes	Hat	CookD
20	2.58	0.00	0.01
217	2.55	0.01	0.02
3012	0.85	0.07	0.00
3112	-1.90	0.06	0.03

```
## Calls:
## 1: glm(formula = ins ~ hisp + married + educyear + chronic + adl + retire +
##   log_income + log_income * married + log_income * hisp, family =
##   binomial(link = logit), data = df)
## 2: glm(formula = ins ~ hisp + married + educyear + chronic + adl + retire +
##   log_income + log_income * married + log_income * hisp, family =
##   binomial(link = logit), data = dflog)
##
##                    Model 1 Model 2
## (Intercept)         -4.427  -4.500
## SE                   0.395   0.400
##
## hisp                 -4.71   -5.81
## SE                    1.20    1.41
##
## married              1.140   1.210
## SE                   0.430   0.435
##
## educyear            0.0730  0.0707
## SE                  0.0149  0.0149
##
## chronic             0.0659  0.0661
## SE                  0.0297  0.0298
##
## adl                -0.1655 -0.1629
## SE                  0.0597  0.0602
##
## retire              0.1749  0.1804
## SE                  0.0825  0.0827
##
## log_income           0.831   0.862
## SE                   0.117   0.118
##
## married:log_income  -0.307  -0.330
## SE                   0.130   0.132
##
## hisp:log_income      1.242   1.564
## SE                   0.354   0.414
##
```

We took these observations out. The coefficients have changed a bit.

Let's run a Hosmer-Lemeshow goodness of fit test to see if our model is correctly specified.

- Test statistic: $HL = \sum_{g=1}^{G} \frac{(P_g - Y_g)^2}{Y_g(1-Y_g)}$

Hosmer-Lemeshow goodness of fit test

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  glm_final_2$data$ins, fitted(glm_final_2)
## X-squared = 54.766, df = 8, p-value = 4.902e-09
```

The p-value is inferior to 0.01, it strenghens the belief that our model is of good fit.

```
##
## Call:
## glm(formula = ins ~ hisp + married + educyear + chronic + adl +
##     retire + log_income + log_income * married + log_income *
##     hisp, family = binomial(link = logit), data = dflog)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1065  -0.9990  -0.5851   1.1621   2.5602
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.50038    0.40042 -11.239  < 2e-16 ***
## hisp               -5.81191    1.41091  -4.119 3.80e-05 ***
## married             1.21030    0.43464   2.785  0.00536 **
## educyear            0.07069    0.01492   4.737 2.17e-06 ***
## chronic             0.06606    0.02976   2.220  0.02644 *
## adl                -0.16286    0.06019  -2.706  0.00682 **
## retire              0.18037    0.08273   2.180  0.02924 *
## log_income          0.86220    0.11842   7.281 3.32e-13 ***
## married:log_income -0.33020    0.13152  -2.511  0.01205 *
## hisp:log_income     1.56399    0.41438   3.774  0.00016 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4272.9  on 3201  degrees of freedom
## Residual deviance: 3834.9  on 3192  degrees of freedom
## AIC: 3854.9
##
## Number of Fisher Scoring iterations: 6
```

Our final model for the logit estimation gives a much better AIC and log-likelihood.

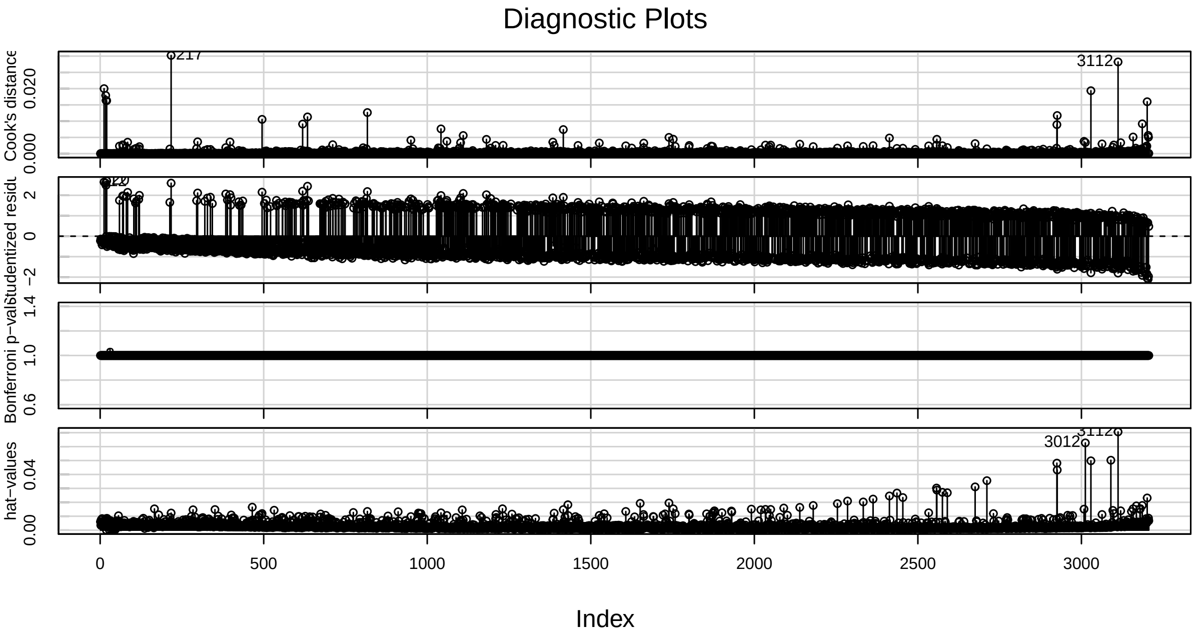
5.2 Probit

```
##
## Call:
## glm(formula = ins ~ 1, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9895  -0.9895  -0.9895   1.3778   1.3778
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.45957    0.03626  -12.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4279.5  on 3205  degrees of freedom
## Residual deviance: 4279.5  on 3205  degrees of freedom
## AIC: 4281.5
##
## Number of Fisher Scoring iterations: 4
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
3205	4279.542	NA	NA	NA
3196	3849.506	9	430.0366	0

```
## [1] 0
```



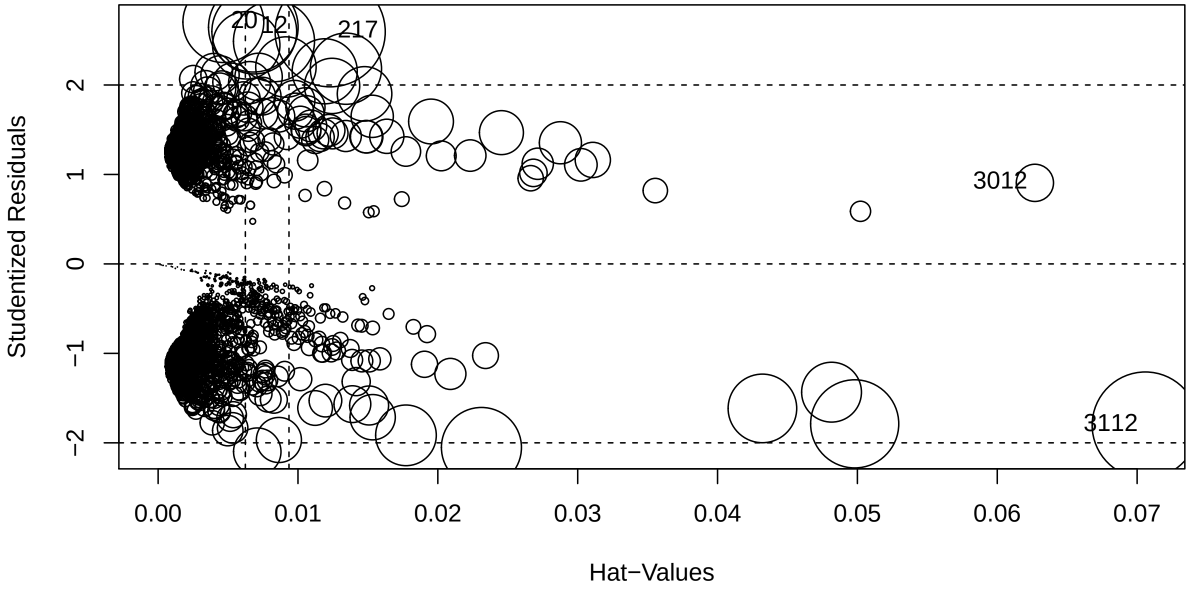


Table 15: Extreme observations

	StudRes	Hat	CookD
12	2.65	0.01	0.02
20	2.70	0.00	0.02
217	2.59	0.01	0.03
3012	0.91	0.06	0.00
3112	-1.80	0.07	0.03

```
## Calls:
## 1: glm(formula = ins ~ hisp + married + educyear + chronic + adl + retire +
##   log_income + log_income * married + log_income * hisp, family =
##   binomial(link = probit), data = df)
## 2: glm(formula = ins ~ hisp + married + educyear + chronic + adl + retire +
##   log_income + log_income * married + log_income * hisp, family =
##   binomial(link = probit), data = dfprob)
##
##                     Model 1 Model 2
## (Intercept)          -2.629  -2.773
## SE                    0.224   0.231
##
## hisp                 -2.539  -3.118
## SE                    0.638   0.742
##
## married               0.597   0.738
## SE                    0.246   0.252
##
## educyear            0.04632 0.04425
## SE                  0.00893 0.00899
```

```
##
## chronic             0.0388  0.0397
## SE                  0.0180  0.0181
##
## adl                -0.1023 -0.0982
## SE                  0.0348  0.0351
##
## retire              0.1056  0.1070
## SE                  0.0501  0.0503
##
## log_income          0.4770  0.5291
## SE                  0.0668  0.0689
##
## married:log_income -0.1560 -0.2015
## SE                  0.0750  0.0769
##
## hisp:log_income      0.668   0.839
## SE                   0.192   0.223
##
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  p_final_2$data$ins, fitted(p_final_2)
## X-squared = 52.196, df = 8, p-value = 1.543e-08
```

The p-value is inferior to 0.01, it strenghens the belief that our model is of good fit.

```
##
## Call:
## glm(formula = ins ~ hisp + married + educyear + chronic + adl +
##     retire + log_income + log_income * married + log_income *
##     hisp, family = binomial(link = probit), data = dfprob)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1510  -0.9975  -0.5729   1.1649   2.6955
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.773203   0.230899 -12.010  < 2e-16 ***
## hisp               -3.117534   0.741975  -4.202 2.65e-05 ***
## married             0.738415   0.251998   2.930  0.00339 **
## educyear            0.044254   0.008989   4.923 8.53e-07 ***
## chronic             0.039726   0.018077   2.198  0.02798 *
## adl                -0.098174   0.035129  -2.795  0.00519 **
## retire              0.106981   0.050300   2.127  0.03343 *
## log_income          0.529137   0.068929   7.677 1.63e-14 ***
## married:log_income -0.201497   0.076924  -2.619  0.00881 **
## hisp:log_income     0.838665   0.223019   3.761  0.00017 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 4271.0  on 3200  degrees of freedom
## Residual deviance: 3823.9  on 3191  degrees of freedom
## AIC: 3843.9
##
## Number of Fisher Scoring iterations: 6
```

Same as the logit estimation, this final model gives a better AIC and log-likelihood.

6 Robustness and prediction

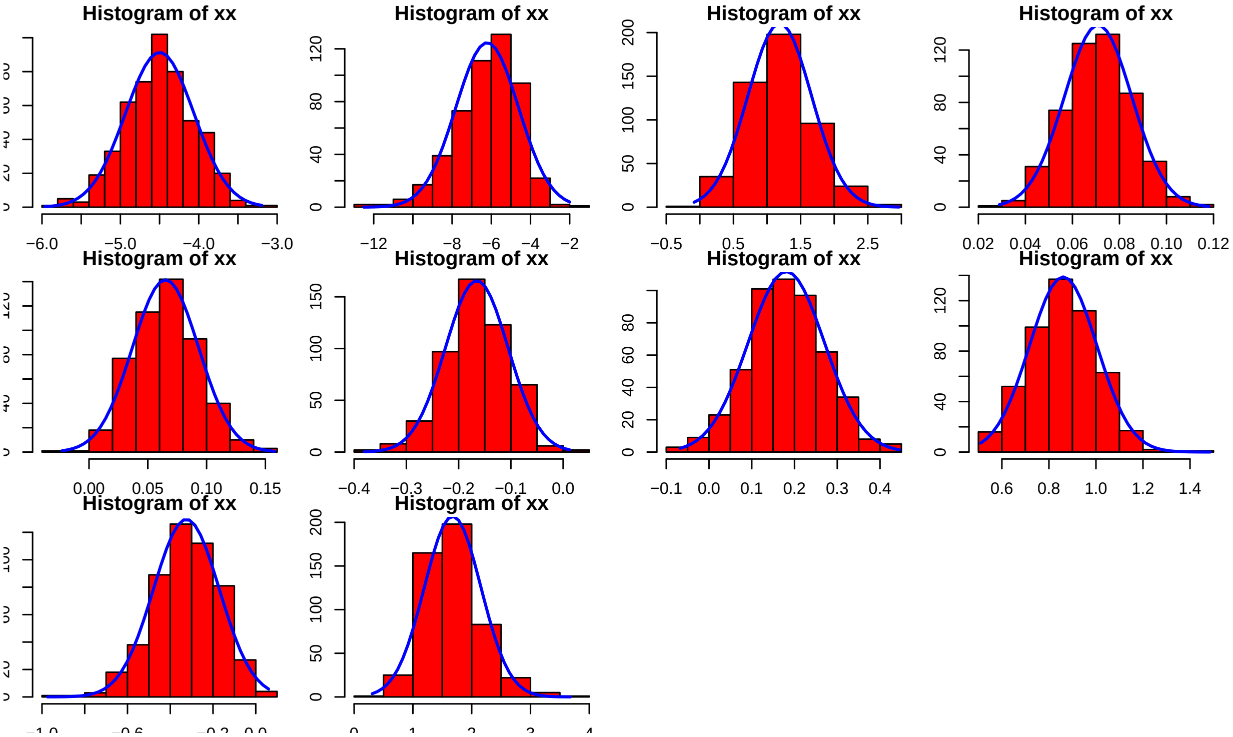
We are going to test the robustness of our models by generating 500 bootstrap samples and testing our latent model on them.

6.1 Logit

Table 16: Confidence intervals

	2.5 %	97.5 %
(Intercept)	-5.30	-3.73
hisp	-8.84	-3.27
married	0.37	2.08
educyear	0.04	0.10
chronic	0.01	0.12
adl	-0.28	-0.05
retire	0.02	0.34
log_income	0.64	1.10
married:log_income	-0.59	-0.08
hisp:log_income	0.81	2.44

For the logit estimation for 95% of the samples the coefficients vary around the originally estimated coefficients and they never change signs, which means that our model is quite stable.

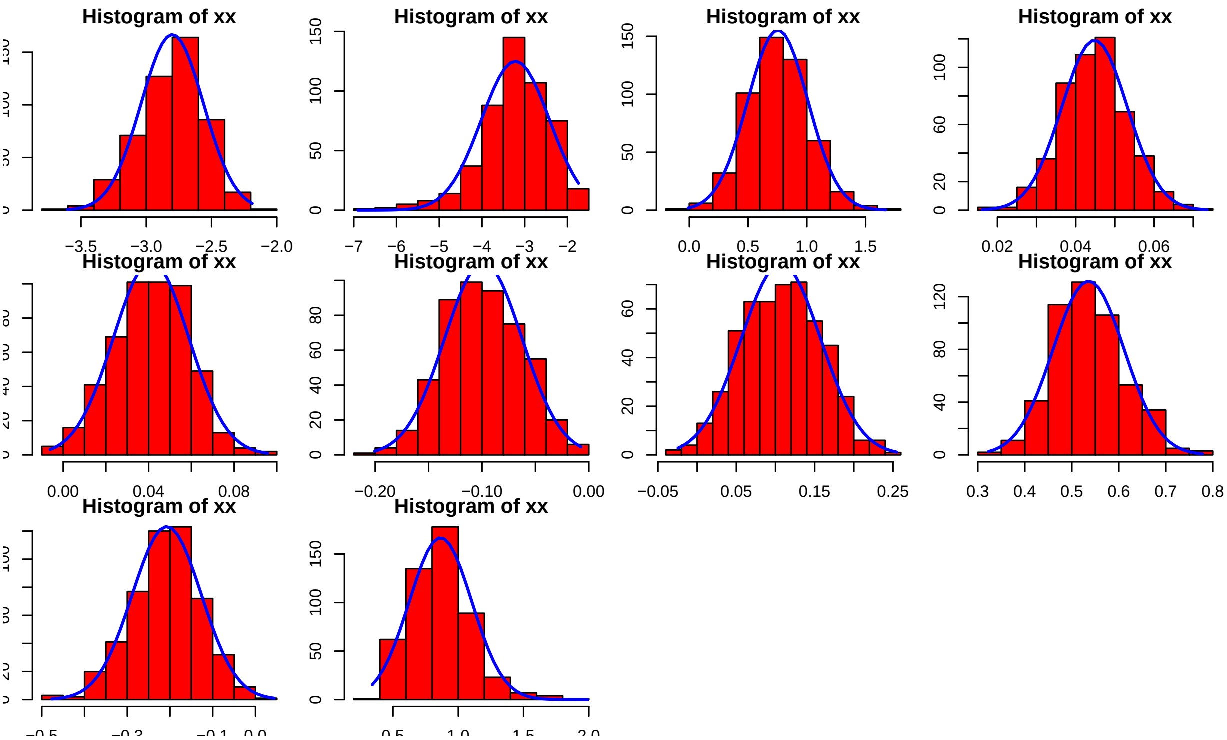


6.2 Probit

Table 17: Confidence intervals

	2.5 %	97.5 %
(Intercept)	-3.23	-2.33
hisp	-4.64	-1.75
married	0.25	1.23
educyear	0.03	0.06
chronic	0.00	0.08
adl	-0.17	-0.03
retire	0.01	0.21
log_income	0.40	0.66
married:log_income	-0.35	-0.05
hisp:log_income	0.43	1.29

For the probit estimation for 95% of the samples the coefficients vary around the originally estimated coefficients and they never change signs, which means that our model is quite stable.



6.3 Predictions

We will take a look at the confusion matrices to see the quality of the prediction at a threshold value equal to 0.5 .

Table 18: Confusion matrix : Logit

Reality	Prediction		
	0	1	Sum
0	1587	377	1964
1	764	474	1238
Sum	2351	851	3202

Table 19: Confusion matrix, proportions : Logit

Reality	Prediction	
	0	1
0	80.8	19.2
1	61.7	38.3

The uninsured group is quite well (specificity: 80.8%) predicted with only 19.2% predicted as insured. For the insured though, the prediction isn't as good; only 38.3% (sensitivity) is correctly predicted. We may have an issue with the size of the group, as twice as small as the not insured group. The global error stands at 35.63% which isn't that good, but not that bad either.

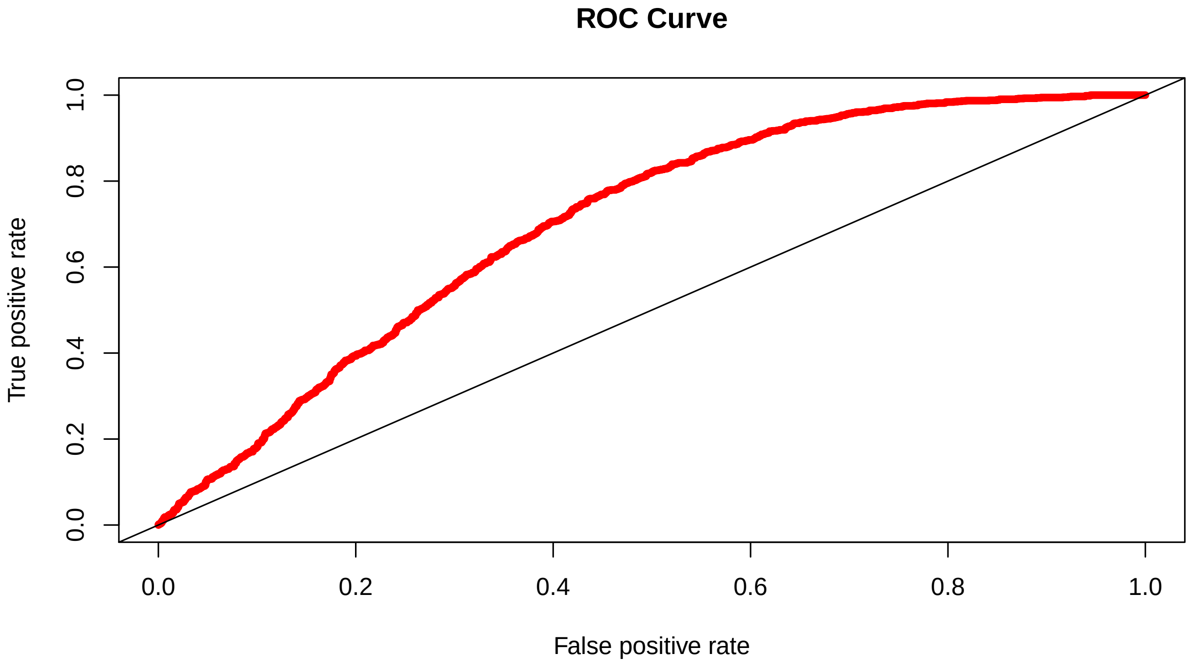


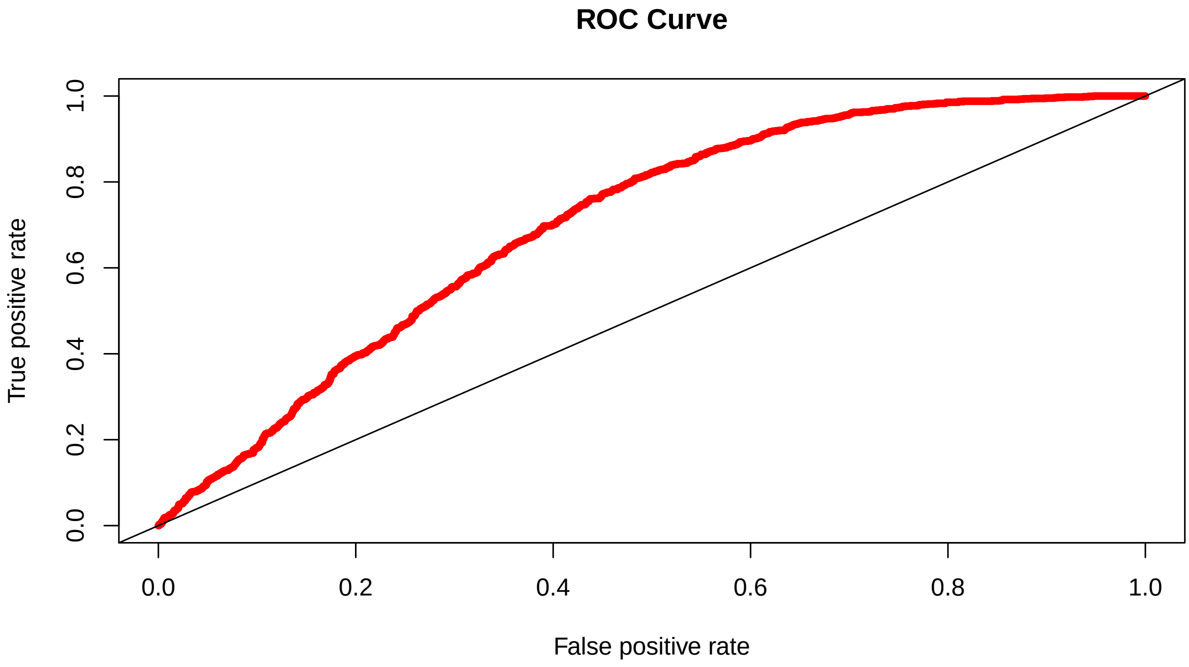
Table 20: Confusion matrix : Probit

Reality	Prediction		
	0	1	Sum
0	1594	370	1964
1	770	467	1237
Sum	2364	837	3201

Table 21: Confusion matrix, proportions : Probit

Reality	Prediction	
	0	1
0	81.2	18.8
1	62.2	37.8

The uninsured group is quite well predicted (specificity: 81.2%) with only 18.8% predicted as insured. For the insured though, the prediction isn't as good; only 37.8% is rightly predicted. The global error stands at 35.61% which isn't that good, but not that bad either. It's nearly the same prediction as the logit estimation, the uninsured are slightly more accuratly predicted and the insured are slightly less accuratly predicted. The overall prediction is slightly better with the probit estimation.

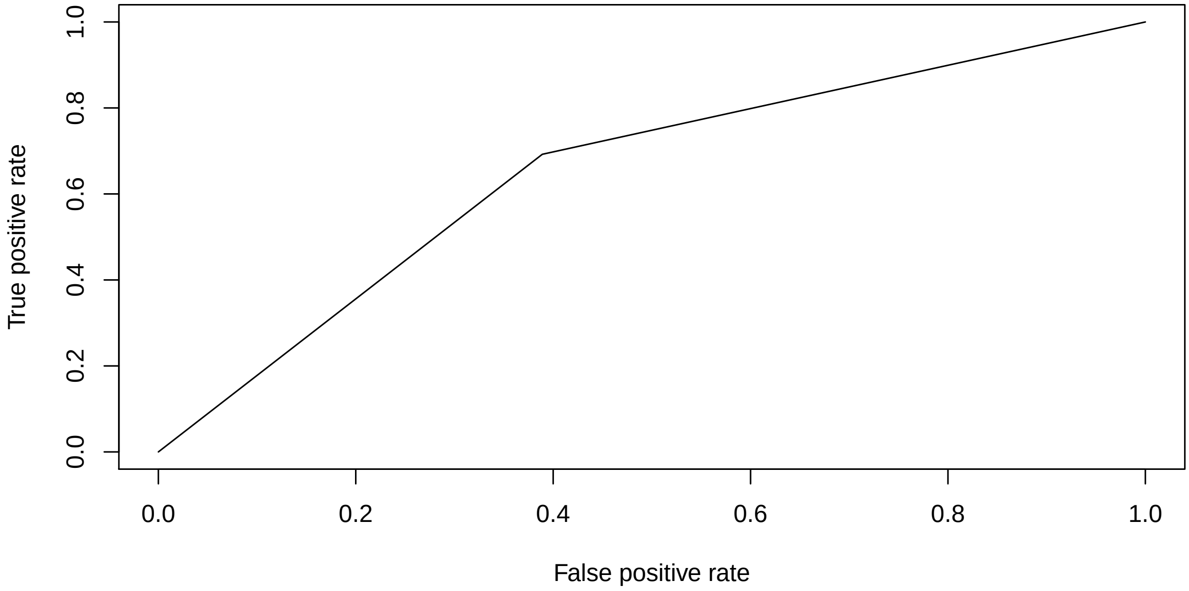


6.4 Improving the prediction

We can play with the decision probability c (if p predicted > c then we categorize the individual as insured) and lets try to maximize that area under the curv ROC.

```
## [1] 0.5954602
```

The area under the curv ROC with c=0.5 is 0.595.



```
## [1] 0.6516218
```

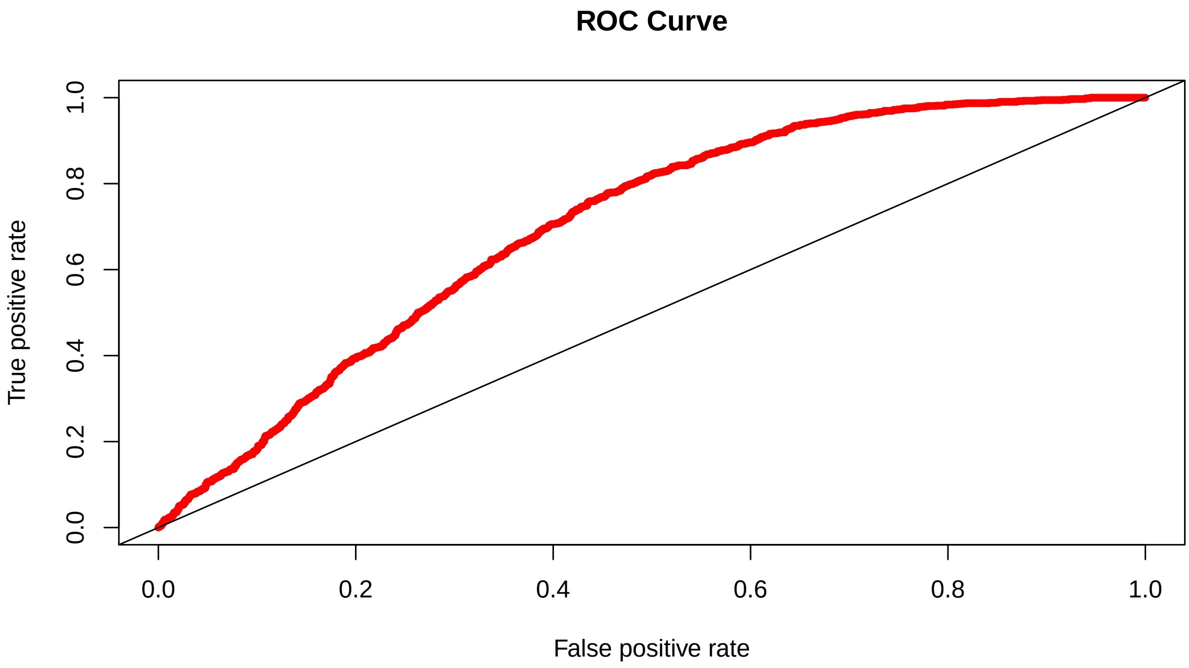
The area under the curv is now 0.652. Let's build the new confusion matrix.

Table 22: New confusion matrix

Reality	Prediction	
	FALSE	TRUE
0	1200	764
1	381	857

Table 23: New confusion matrix

Reality	Prediction	
	FALSE	TRUE
0	61.10	38.90
1	30.78	69.22



The global error is a bit better, we predict a lot more insured individuals but far less not insured. We can work with the categorizing probability to try to improve our predictions.

7 Odd ratios, z-values and average marginal effects

As the model contains variables in interaction, both in the logit and probit model, we have to be careful with how we interpret some results. Marginal effects can't be interpreted for variables in interaction in non-linear models it makes no-sense.

7.1 Logit

Table 24: Odd ratios : Logit

	x
(Intercept)	0.011
hisp	0.003
married	3.354
educyear	1.073
chronic	1.068
adl	0.850
retire	1.198
log_income	2.368

Table 25: Average marginal effects : Logit

	x
(Intercept)	-0.937
hisp	-1.210
married	0.252
educyear	0.015
chronic	0.014
adl	-0.034
retire	0.038
log_income	0.179

- If the individual isn't married and isn't hispanic, if the logincome goes up by one it increases the chances of being insured by 25%.
- If the number of years spent in school goes up by one, it increases the chances of being insured by 1.5%
- If intensity of the chronic disease goes up by one, it increases the chances of being insured by 1.4%%
- If number of restricted activities goes up by one, it diminishes the chances of being insured by 3.4%
- If the individual is retired, it increases the chances of being insured by 3.8%

Predicted probabilities

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0001072 0.2751976 0.4016832 0.3866334 0.5070740 0.9104751
```

50% of the individuals of this database are predicted to have a probability of being insured over 40%.

7.2 Probit

Table 26: Z-score increase : Probit

	x
(Intercept)	-2.773
hisp	-3.118
married	0.738
educyear	0.044
chronic	0.040
adl	-0.098
retire	0.107
log_income	0.529

Table 27: Average marginal effects : Probit

	x
(Intercept)	-1.531
hisp	-1.977
married	0.412
educyear	0.024
chronic	0.022
adl	-0.055
retire	0.061
log_income	0.293

- If the individual isn't married and isn't hispanic, if the logincome goes up by one it increases the chances of being insured by 29.3%.
- If the number of years spent in school goes up by one, it increases the chances of being insured by 2.4%
- If intensity of the chronic disease goes up by one, it increases the chances of being insured by 2.2%%
- If number of restricted activities goes up by one, it diminishes the chances of being insured by 5.5%
- If the individual is retired, it increases the chances of being insured by 6.1%

Predicted probabilities

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000001 0.2741402 0.4022680 0.3848023 0.5049755 0.9010721
```

50% of the individuals of this database are predicted to have a probability of being insured over 40%.

8 Interpretation

8.1 Logit

As we have two interactions within these models, one continuous variable interacting with two discrete variables, we have two models that can be written in four expressions.

8.1.1 Variables in interaction

When you have an $hisp$=0 and $married$=0, $hisp$=0 and $married$=1, $hisp$=1 and $married$=0, $hisp$=1 and $married$=1:

- $logit(ins_i) = \alpha + \beta_2 educyear_i + \beta_4 chronic_i + \beta_5 adl_i + \beta_6 retir_i + \beta_7 logincome_i + \mu_i$
- $logit(ins_i) = \alpha + \beta_1 hisp_i + \beta_2 educyear_i + \beta_4 chronic_i + \beta_5 adl_i + \beta_6 retir_i (\beta_7 + \beta_9)(logincome_i) + \mu_i$
- $logit(ins_i) = \alpha + \beta_2 educyear_i + \beta_3 married_i + \beta_4 chronic_i + \beta_5 adl_i + \beta_6 retir_i + (\beta_7 + \beta_8)(logincome_i) + \mu_i$
- $logit(ins_i) = \alpha + \beta_1 hisp_i + \beta_2 educyear_i + \beta_3 married_i + \beta_4 chronic_i + \beta_5 adl_i + \beta_6 retir_i + (\beta_7 + \beta_8 + \beta_9)(logincome_i) + \mu_i$

Now we can map the logistic regression output to these two equations. So we can say that the coefficient for logincome is the effect of $logincome$ when $hisp$ = 0 and $married$ = 0. More explicitly, we can say that for single non-hispanics, a one-unit increase in logincome score yields a change in log odds of 0.862. For a single hispanic, a one-unit increase in logincome yields a change in log odds of $(0.862 + 1.564) = 2.426$.
In terms of odds ratios, we can say that for single non-hispanics, the odds ratio is $exp(0.862) = 2.368$ for a one-unit increase in $logincome$ and the odds ratio for single hispanics is $exp(2.426) = 11.314$ for a one-unit increase in logincome. The ratio of these two odds ratios (single hispanic over single non-hispanic) turns out to be the exponentiated coefficient for the interaction term of single hispanic by $logincome$: $11.314/2.426 = exp(1.564) = 4.7$.

Same process for married non-hispanics, a one-unit increase in logincome yields a change in log odds of (0.862 - 0.330) = 0.532. The odds ratio for married non-hispanics is exp(0.532) = 1.702 for a one-unit increase in logincome. The ratio of these two odds ratios (maried non-hispanic over single non-hispanic) turns out to be the exponentiated coefficient for the interaction term of married non-hispanics by $logincome$: 1.702/2.426 = exp(-0.330) = 0.71.

At last, for married hispanics, a one-unit increase in logincome yields a change in log odds of (0.862 - 0.330 + 1.564) = 2.096. The odds ratio for married hispanics is exp(2.096) = 8.134 for a one-unit increase in logincome. The ratio of these two odds ratios (maried hispanic over single non-hispanic) turns out to be the exponentiated coefficient for the interaction term of married hispanics by $logincome$: 8.134/2.426 =exp(-0.330+1.564) = 3.40.

8.1.2 Other variables

- This fitted model says that, holding all other exogenous variables at a fixed value, the odds of being insured for retired individuals (retire = 1)over the odds of being insured for not-retired individuals (retire = 0) is exp(0.180) = 1.198. In terms of percent change, we can say that the odds for retired are 19.8% higher than the odds for working individuals.
- we will see 6.8% increase in the odds of being insured for a one-degree increase in intensity of a chronic disease since exp(0.066) = 1.068.
- we will see 15% decrease in the odds of being insured for a one restriction added in adl since exp(-0.163) = 0.850.
- we will see 7.3% increase in the odds of being insured for one more year of studying since exp(0.071) = 1.073.

8.2 Probit

The probit estimation coefficients are a bit more complicated to interpret. We say that the coefficients represent for a one-unit increase the change in the z-score. Again we have to be careful with the variables in interaction.

8.2.1 Variables in interaction

We won't write the equations again but they are the same as the logit model.

```
## [1] 1.167
```

So we can say that the coefficient for logincome is the effect of *logincome* when $hisp = 0$ and $married = 0$. More explicitly, we can say that for $singlenon - hispanics$, a one-unit increase in logincome yields an increase in z-score of 0.529. For a *singlehispanic*, a one-unit increase in logincome yields an increase in z-score of $(0.529 + 0.839) = 1.368$.

Same process for married non-hispanics, a one-unit increase in logincome yields an increase in z-score of (0.529-0.201) = 0.328.

At last, for married hispanics, a one-unit increase in logincome increases the z-score of (0.529 - 0.201 + 0.839) = 1.167.

8.2.2 Other variables

- This fitted model says that, holding all other exogenous variables at a fixed value, the z-score for retired individuals (retire = 1)over the z-score for not-retired individuals (retire = 0) increases by 0.107.
  - for a one added degree of intensity of the chronic disease the z-score increases by 0.040
  - for one more restriction on activities the z-score decreases by 0.098
  - for one more year of studying the z-score increases by 0.044.

8.3 Conclusion

The probability of being insured, not suprisingly, relies heavily on the household income, as the income goes up, the probability of being insured goes up too. Many other variables interact with the household income: being married probably increases the household income as the spouse potentially adds another source of revenue, being hispanic may be related to lower incomes and so the overcompensated coefficient may mean that for the same income as non- hispanics the probability of being insured may be the same.

Having a more intense chronic disease slightly increases your chances of being insured too. It sounds quite logical as a chronic disease happens to be a source of expenses in medical bills, foreseeable expenses.

I don't really know how to interpret how the restrictions on daily activities have a negative impact on the probability of being insured or not. Maybe restrictions on daily activities prevent you for getting an insurance, or maybe they are in related to a lower wage as seen in the correlation matrices.

The positive impact of a higher education level on the probability of being insured isn't surprinsing at all. Firstly, the number of years of education are strongly correlated with a higher income. Then, it may have an impact on the knowledge of the perks of being insured versus not being insured.

And now the last variable, *retire*. Being retired have a positive impact on the probability on being insured. It may be correlated to the fact that retired individuals may think about the increasing risk of having some upcoming heavy medical bills to pay, as getting older means having a bigger chance to fall seriously ill.

9 Discusion and limitations

9.1 The data

- In my opinion, this database doesn't provide sufficient information to predict accuratly the probability for an individual to be insured with a private company. Firstly, I am not sure that someone would insure himself when he is sick, and by that I mean that insurance is useful only to a healthy person as a prevention in case he becomes sick and then has to pay some medical bills that might be covered by his insurance. So all the variables on the health status are insignificant. Insurance is a long term investment, in this case on your health and your financial ability to pay for healthcare. This database lacks of temporality. Let's say we would be able to get some additionnal data to improve our model and predictions, I would search for the number of health incidents that have occured during the lifetime of the individual. Or having an estimation on the amount payed in medical bills in the last 5 years…
- Also, let's talk about the household income which doesn't really give enough information about the individual. Obviously it has a great positive impact on the probability of being insured, but, it lacks details like individual income, number of children…
- I would also add some information about how much money does the insured spend on insurance, we could then create some measures with the income like the percentage of the income spent on insurance…
- Sample size: Both probit and logit models require more cases than OLS regression because they use maximum likelihood estimation techniques. It is sometimes possible to estimate models for binary outcomes in datasets with only a small number of cases using exact logistic regression. It is also important to keep in mind that when the outcome is rare, even if the overall dataset is large, it can be difficult to estimate a probit model. In our case the data wasn't balanced, but, it wasn't too unbalanced. It may have caused a little biase but not that much.

9.2 Pros and cons of the logistic regression

Like any method, it has its pros and cons. Perhaps the biggest pro is that the gradient and Hessian, which are typically used for optimization, are functions of the logit probabilities themselves, so require no additional computation. Most variables are very easy to interpret in the end even though odds ratios aren't the easiest concept to communicate. We can communicate in terms of predicted probabilities, average marginal effect which all have a direct interpretation quite easy to understand.

Now onto the cons. First, it was very hard to use variables in interaction. Even if it made sense in to use them and it improved the overall performance I don't really know if I am allowed to use variables in interaction in that non-linear model.

Then the binary (black or white - no grey) nature of logistic regression.

Suppose you are monitoring sales. If the person buys, a 1 is scored - if no sale, a 0 is scored. That does not tell you that the person did buy the article but not from you. All the near misses are 0 - all the total failures are 0.

A similar example is a rugby national team wining the world cup. The winner is awarded 1, the losing finalist 0. All the other tournament competitors get 0. But the losing finalist is a far better team(a 0.9 say) than a team knocked out in the first round - but they both score 0. You could do an LR on just the 2 finalists in every world cup but then you would be losing all the information from the majority of teams in the tournament who did not reach the finals but may have won or lost rounds or even got to the quarter or semi-finals.

The values of 0 (fail) and 1 (success) are arbitrary. The important part (that tends to get lost) is not to predict the numerical value of Y, but the probability that success or failure occurs, and the extent to which that probability depends on the predictor variables. If you throw out or corrupt most of the potential informatory data how can you be sure you have the right probabilities of success or failure?

When we come back at our problematic, we can ask ourselves, is the person predicted as insured has a full insurance? for how many years has he been insured? Will he stay insured? There is no grey, just black or white.

And finally one of the biggest problem is that logit (and probit) all have very specific tail assumptions. That is, extreme cases are presumed to become progressively rare at a very specific rate. While this isn't so severe where you have a lot of data, by definition the tails are where data are quite rare, so you often will have the most of your function, where you want to make accurate predictions for most data, influenced strongly by a bit of data in the tails. We should be very careful in extrapolating probabilities from the logit model, or any model that makes these predictions only on functional form assumptions.