Uppsala University

Applied Statistical Methods

Homework Assignment 1

Vote for Social Democrats

Claes Kock

Mayara Latrech

Yuchong Wu

December 17, 2019

Abstract

In this study, we tried to figure out what variables impact the choice of voting for the Social Democratic Party of Sweden as well as their effects. The study is based on a dataset from the European Social Survey from 2006. We selected 4 Likert variables. Then we established Logit and Probit models and found that most of our variables were significant and constant scale, except for one.

The accuracy of our logit model was 61.1%, while the accuracy of our probit model was slightly higher, at 61.3%. Thus, we conclude that the probit model is better than the logit model, but the difference is very small.

Contents

1 Introduction

Sweden has many political parties. The purpose of this project is to use statistical tools to predict voting behavior in Sweden. To do that, we try to predict whether an individual would vote for The Social Democrats based on known variables. We get a data-set that contains information about 1927 such as, their political interested if they are Swedish born or even the time they spend reading newspapers. Throughout our work, we apply some transformation to this data-set and choose what we think are important variables. After that, we will feed those variables into a model: we will use probit and logit models. We will also apply the likelihood ratio test to determine if our variables are Lickert variables or not. Finally, we will predict if the individuals would vote for the Social Democrats or not. The model that has the best predictions would be the best.

2 Theory

In this assingnment, a linear regression model will **not** be suitable for the reason that the independant variable we use is binary.

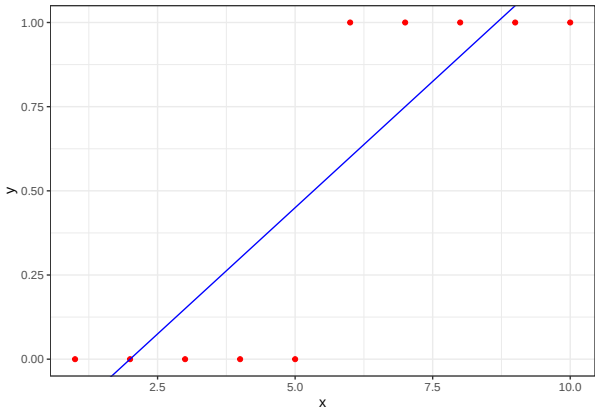


Figure 1: Linear Regression Model

Thus we need to apply both a logit and a probit model to the same dataset. We start by transforming our data, removing all individuals who answered "I don't know" and those who did not answer. We decided to keep those who refused to answer. The logic behind that is refusing to answer is a valid statement and not a case of ignorance. After applying this transformation we apply a logit model.

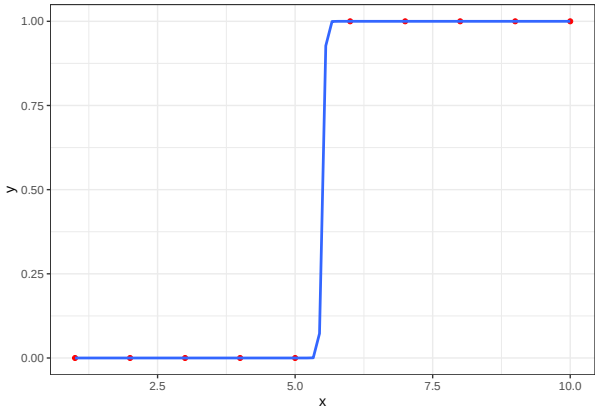


Figure 2: Logit Model

A Logit model is a regression model where the dependent variable is a categorical, binary (zero or one) variable.

$$P(X = 1|X_1, X_2, ..., X_k) = F(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)$$

Where,

$$F(x) = \frac{1}{1 + e^{-x}}$$

$\beta_0, \beta_1, .., \beta_k$ are the parameters of the model to be estimated. $X_1, X_2, ..., X_k$ are the variables and the dependant variable X is "social democrats" in our dataset. The variable we are trying to explain is wether an individual voted for the Social Democratic party or not.

The probit model is similar to the logit model, but uses the CDF form the normal distribution instead of the logistic distribution.

$$P(X = 1|X_1, X_2, ..., X_k) = F(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k)$$

Where, $F(*)$ is the CDF from the normal distribution.

In order to evaluate the models, we draw a confusion matrix, which is putting the predictions of the model wersus their true value.

Table 1: Confusion Matrix

Obs_Pred	Obs_0	Obs_1
Pred 0	True Negative(TN)	False Negative(FN)
Pred 1	False Positive(FP)	True Negative(TN)

The goodness of fit measure that we chose is the accuracy: it is the measure of all the correctly predicted values. The formula to calculate it is:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

3 Data

Our data is from the European Social Survey, which is a biannual multicountry survey. This survey measures attitudes, beliefs and behaviours across the European Union. Our dataset is the survey from 2006 and contains a total of 487 variables from 1927 respondents. However, we are only going to use the variables that are relevant to swedish politics.

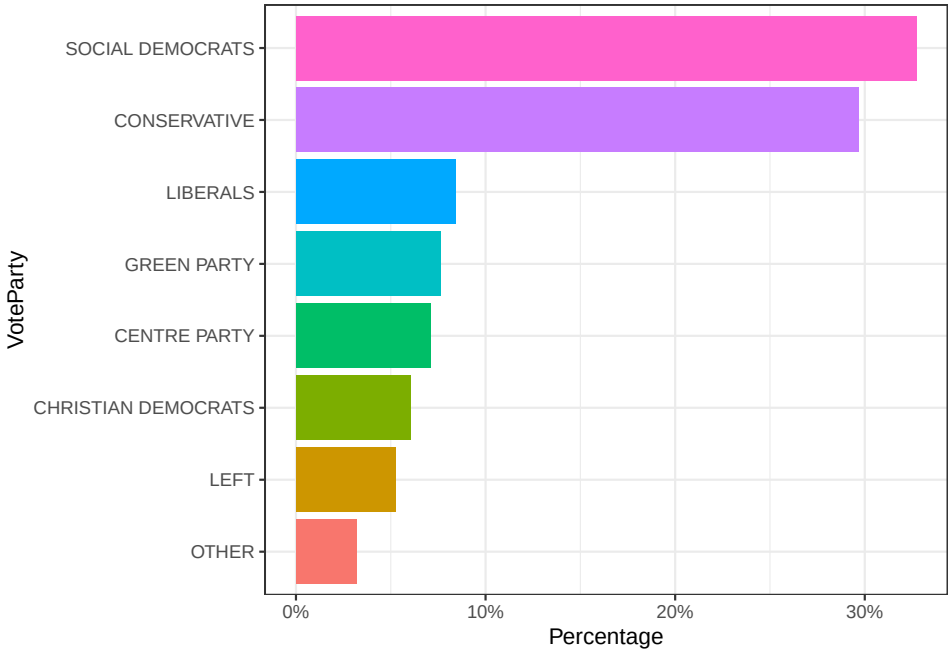
Thus we use a total of 24 variables from this dataset, plus some dummy variables.

Table 2: All the variables in the dataset

Variables	Names
1	PaperTime
2	PoInt
3	VoteLast
4	VoteParty
5	HapScale
6	HealthLvl
7	SweBorn
8	FSweBorn
9	MSweBorn
10	Crime5year
11	centre
12	christian
13	conserv
14	Spouse3year
15	Married
16	TotalChildren
17	OptLvl
18	PosLvl
19	SatLvl
20	FamSize
21	Gender
22	FormUnemp
23	Bdate
24	TradeUn
25	IncLvl
26	Area
27	green
28	happy
29	leftparty
30	liberal
31	socialdemocrats

4 Descriptive Statistics

1)In order to know how the voters voted in the last general election in Sweden, we draw a barplot of how many votes each party got. Since we are intrested in voters only this task, we do not take into account people who are not applicable, who refused to answer the survey, who do not know and who did not give an answer.



We notice that the Social Democrats received the largest share of the votes, 32%. The left party got the least amount of votes with 6% of the total votes. It seems like most people either vote for the right wing or the left wing. The Centre party does not seem that popular. We also notice that people prefer parties that are not in either extreme; the left party, liberal party and the christian democrats are not popular compared to the social democrates and the conservative party. The Green party does not seem that popular as well, it got only 8% of the total votes. Overall, it seems like voters are more left leaning when voting to various parties

4.1 Logit model

We have too many variables and will have to restrict the number of variables we use in our logit-model. We have selected a couple of variables which we think will have a high impact

1) We expect trade union members to be more likely on average to vote SocDem, due to the symbiotic relationship between labor and SocDem in Sweden, especially in 2006. However, due to how the variables are coded, we expect this to be a negative influence on voting SocDem since the higher values are for former members and non-members.

2) The variable political interest could be a positive correlation, since the social democrats have been a major force in Swedish politics for decades, and a high political interest could increase the chance to vote overall, instead of the individual staying at home due to having no interest in politics. However, high political interest is a low value for this variable, so we expect this variable to be negative.

3) High family size is a socio-demographic factor which is often correlated with working-class households - we expect larger households to have a higher chance to vote for SocDem for this reason.

4) Finally, the type of area that the individual is living in could also be a positive impact whether they vote SocDem since a lot of industry is spread out over the smaller cities and countryside, whereas cities might be expected to have a higher chance to vote for liberal parties.

Table 3: Chosen variables

VarName	Measure	ExpEffect
TradeUn	Trade Union membership	Neg.
PoInt	Political interest	Neg.
FamSize	Size of family	Pos.
Area	Urbanisation	Pos.

Table 4: Logit Model

Coefficients	Estimate	Significant
Intercept	-0.51	-
PoInt	0.17	Yes
TradeUn	-0.53	Yes
FamSize	-0.11	Yes
Area	0.07	No

4.2 Intepretation of parameters

After estimating our model, we obtain $\hat{\beta}_0, ..., \hat{\beta}_4$ estimates of the parameters of the model. $\hat{\beta}_0, ..., \hat{\beta}_3$ are significant and $\hat{\beta}_4$ is not significant. - For the person who is one level less interested in politics, the log odds of voting for the Social democrates increases by 0.1710897, holding other variables constant. - For the person who is a member of the Trade Union compared to the one who was a member or the person who was a member of the Trade Union compared to the person who was not a member, the log odds of voting for the Social democrates decreases by -0.5333775, holding other variables constant. - For every one additional member increase in family size, the log odds of voting for the Social democrates decreases by -0.1087834, holding other variables constant. - The parameter for the variable Area is not significant so it does not contribute any significant explanatory power.

4.3 Change one sample point

Let us now try to choose one sample point and calculate the probability of the chosen individual to vote for the Social democrats, if we choose the 18^{th} individual for example, the probability stemmed from the model would be 0.2902279.

We want to see what happens when we manipulate the regressors of a single individual. For that purpose, we have picked out individual #18 from all the observations. This individual is "quite interested" in politics. We decided to change this value to observe how this affects the probability of voting SocDem. We can see a 3.39% increase in probability when we raise the level of interest in politics from "2" to "1". When we lower the value to "3", we instead get a -3.64% decrease in probability. Finally "4" gives us a decrease of -7.51%. Thus, when predicting probability, it would seem that a higher political interest has a higher correlation with voting SocDem if we apply the results from looking at 1 individual on the entire dataset.

Individual #18 is also a current member of a trade union. We increase the value of the var TradeUn to "2", which means he's a former member. According to our prediction, this makes him 9.67% more likely to vote SocDem. If we increase it further to "3", which means he's not a member, #18 is 16.6% more likely to vote SocDem. This is because of the relationship between being a trade union member and having a high value in the variable TradeUn is inverted, the higher value instead being given to individuals that are not members.

individual #18 has a familiy size of 3, a spouse and a child. We decrease the size of this family to just containing 1 member, in order to see how this affects the likelihood to vote for SocDem. This will allow us to see the change in probability when looking at a individual who is single compared to a member of a family. From our prediction, the change in probability is -4.67%. Similarly, individuals from families of a size of 2, i.e. a single parent or 2 spouses, gives a change in probability of -2.29%. If we instead increase from 3 to 4 members, we get an increase of 2.18%. This keeps increasing the more members above 3 we add - an individual from a family of 6 members has a probability increase of 6.24% for example. This might be because working class households (which are probably more likely to vote SocDem) trend towards having larger families, and these households in turn form a large part of the SocDem voters.

Table 5: Changes in individual #18

Nr	PoInt	TradeUn	FamSize
1	0.03	0.00	-0.05
2	0.00	0.10	-0.02
3	-0.04	0.17	0.00
4	-0.08	NA	0.02
5	NA	NA	0.04

4.4 Likelihood Ratio test

4.4.1 PoInt

We split the Political Intrest into dummy variables, now we have two models; the first one has Political interest, Membership in the Trade Union, Family size and are, the second one has those same variables but Political Interest is split into dummy variables. We then apply the Likelihood ratio test.

- Hypothesis: H_0: The Likert variable: Policial Intrest has a constant scale H_1: The Likert variable: Policial Intrest does not have a constant scale
- Significance level: $\alpha = 0.05$
- Estimators: $\hat{\beta}_1, \hat{\beta}_1^2, \hat{\beta}_1^3, \hat{\beta}_1^4$
- Assumptions: Large T
- Test statistic: $LR = -2 \ln \left[\frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)} \right] \sim \mathcal{X}^2$
- Decision rule and figure: We reject the null hypothesis if $LR > \mathcal{X}^2_{\alpha,2}$ or if $p - value < 0.05$

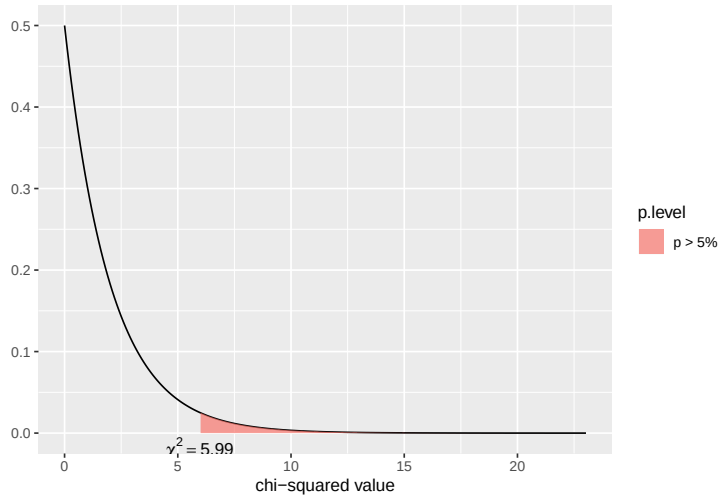


Figure 3: Chisquare Distribution

- Calculations and decision: p-value= 0.2563197>0.05. So we do not reject H_0.
- Conclusion: At the 5% significance level, we fail to reject the null that the Likert variable: Policial Intrest has a constant scale.

4.4.2 TradeUn

We split the membership in the trade union into dummy variables, now we have two models; the first one has Political interest, Membership in the Trade Union, Family size and are, the second one has those same variables but Membership in the Trade Union is split into dummy variables. We then apply the Likelihood ratio test.

- Hypothesis: H_0: The likert variable: membership in the trade union has a constant scale H_1: The likert variable: membership in the trade union does not have a constant scale
- Significance level: $\alpha = 0.05$
- Estimators: $\hat{\beta}_2, \hat{\beta}_2^2, \hat{\beta}_2^3$
- Assumptions: Large T
- Test statistic: $LR = -2 \ln \left[\frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)} \right] \sim \mathcal{X}^2$
- Decision rule and figure: We reject the null hypothesis if $LR > \mathcal{X}^2_{\alpha,1}$ or if $p - value < 0.05$

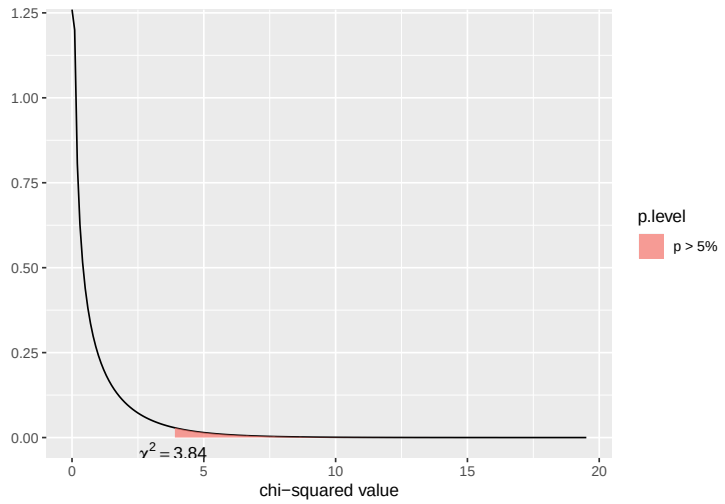


Figure 4: Chisquare Distribution

- Calculations and decision: p-value= 0.6550473>0.05. So we do not reject H_0.
- Conclusion: At the 5% significance level, we fail to reject the null that the Likert variable: Membership to the trade union has a constant scale.

4.4.3 FamSize

We split the Family size into dummy variables, now we have two models; the first one has Political interest, Membership in the Trade Union, Family size and are, the second one has those same variables but Family size is split into dummy variables. We then apply the Likelihood ratio test.

- Hypothesis: H_0: The likert variable: Family Size has a constant scale H_1: The likert variable: Family Size does not have a constant scale
- Significance level: $\alpha = 0.05$
- Estimators: $\hat{\beta}_3, \hat{\beta}_3^2, \hat{\beta}_3^3, \hat{\beta}_3^4, \hat{\beta}_3^5, \hat{\beta}_3^6, \hat{\beta}_3^7, \hat{\beta}_3^9, \hat{\beta}_3^{12}$
- Assumptions: Large T
- Test statistic: $LR = -2 \ln \left[\frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)} \right] \sim \mathcal{X}^2$
- Decision rule and figure: We reject the null hypothesis if $LR > \mathcal{X}^2_{\alpha,7}$ or if $p - value < 0.05$

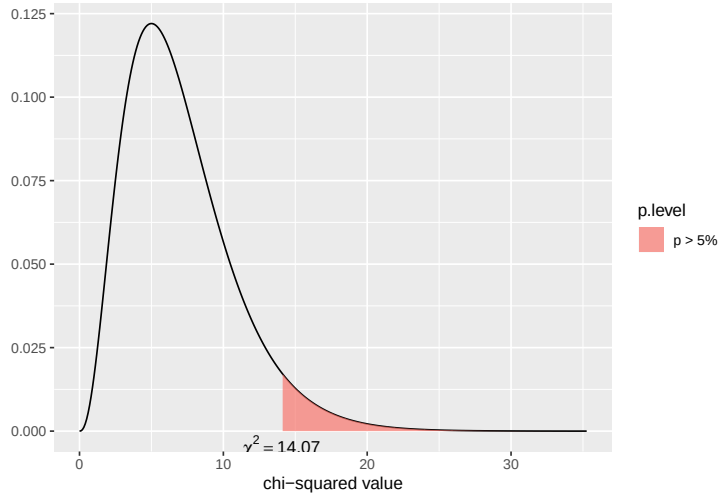


Figure 5: Chisquare Distribution

- Calculations and decision: p-value= 0.3512204>0.05. So we do not reject H_0.
- Conclusion: At the 5% significance level, we fail to reject the null that the Likert variable: Membership to the Family Size has a constant scale.

4.4.4 Area

We split the Area dummy variables, now we have two models; the first one has Political interest, Membership in the Trade Union, Family size and are, the second one has those same variables but Area is split into dummy variables. We then apply the Likelihood ratio test.

- Hypothesis: H_0: The likert variable: Area has a constant scale H_1: The likert variable: Are does not have a constant scale
- Significance level: $\alpha = 0.05$
- Estimators: $\hat{\beta}_4, \hat{\beta}_4^2, \hat{\beta}_4^3, \hat{\beta}_4^4, \hat{\beta}_4^5$
- Assumptions: Large T
- Test statistic: $LR = -2 \ln \left[\frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)} \right] \sim \mathcal{X}^2$
- Decision rule and figure: We reject the null hypothesis if $LR > \mathcal{X}^2_{\alpha,1}$ or if $p - value < 0.05$

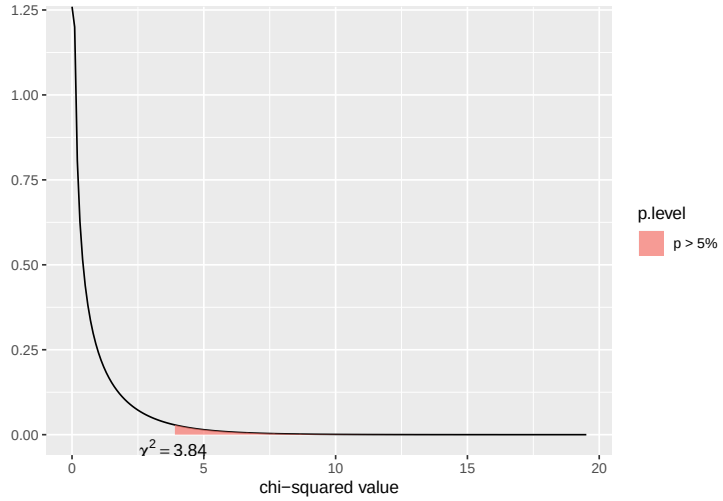


Figure 6: Chisquare Distribution

- Calculations and decision: p-value= 7.3115681×10^{-4}<0.05. So we reject H_0.
- Conclusion: At the 5% significance level, we reject the null that the Likert variable: Area has a constant scale.

4.4.5 Summary for constant scale test

The recommended practice in this situation is to treat Political Intrest, trade union membership, and Family size as Lickert variables, but to attribute different parameters to each dummy of the Area variable.

4.5 Comparison between observed values and expected values

Table 6: Predicted and observed values

	Obs. 0	Obs. 1
Pred 0	1294	472
Pred 1	0	0

At the 0.5 cut-off, the model does not predict anyone to vote for the social democrates. The model always guesses correctly that a person would not vote for the social democrates if in reality they did not vote for them. The model always guesses incorrectly that a person would vote for the social democrates if in reality they did vote for them. So, the model decided that in every case the individual did not vote for the social democrates. We suspect that we got these results due to fitting the model to an unbalanced dataset; the dataset contains 1294 individuals who did not vote for the social democrates and only 472 who did vote for them. Thus, the choice of having a cut-off point at 0.5 may not be adequate. We try to come up with the best cut-off point by plotting the performance of the model for various cut-off points versus what true positive values it should give us. The cut-off point should be the intersection between those two.

The goodness-of-fit measure that we choose is the accuracy,. It is the ration of the correctly predicted observations to the total observations. Is calculated by: Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

where: TP = True positive; FP = False positive; TN = True negative; FN = False negative.

So the McFadden R2 for our model is approximately 0.0399946 and the accuracy for our model is 0.7327293

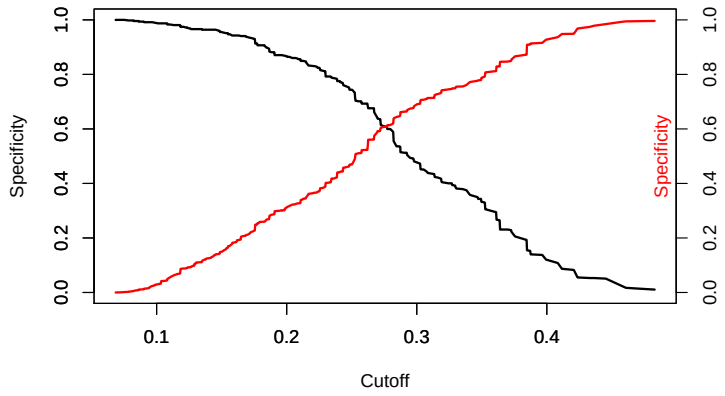


Figure 7: Plot for Cut-off Point

The intersection point is 0.28. Hence, it is the best cut-off point. So we try applying it.

Table 7: Predicted and observed values

	Obs. 0	1
Pred 0	802	194
Pred 1	492	278

With this new cut-off point, 802 individuals who did not actually vote for the social democrats were actually predicted correctly, however the model got 492 individuals who did not actually vote for the social democrats wrong. The model now is able to predict correctly 278 of the individuals who voted for the social democrats, however it failed with 194 individuals from this category. Overall, the new model with cut-off point of 0.28 gives better results than the model with cut-off point =0.5.

So the McFadden R2 for our model is approximately 0.0399946 and the accuracy for our model is 0.6115515

4.6 Probit model

Table 8: Probit Model

Coefficients	Estimate	Significant
Intercept	-0.34	NA
PoInt	0.10	Yes
TradeUn	-0.31	Yes
FamSize	-0.06	Yes
Area	0.04	No

After estimating our probit model, we obtain $\beta^0, ..., \beta^4$ estimates of the parameters of the model. $\beta^0, ..., \beta^3$ are significant and β^4 is not significant.

- For the person who is one level less interested in politics, the log odds of voting for the Social democrates increases by 0.0920531, holding other variables constant.
- For the person who is a member of the Trade Union compared to the one who was a member or the person who was a member of the Trade Union compared to the person who was not a member, the log odds of voting for the Social democrates decreases by -0.3144133, holding other variables constant.
- For every one additional member increase in family size, the log odds of voting for the Social democrates decreases by -0.0576068, holding other variables constant.
- The parameter for the variable Area is not significant so it does not contribute any significant explanatory power.

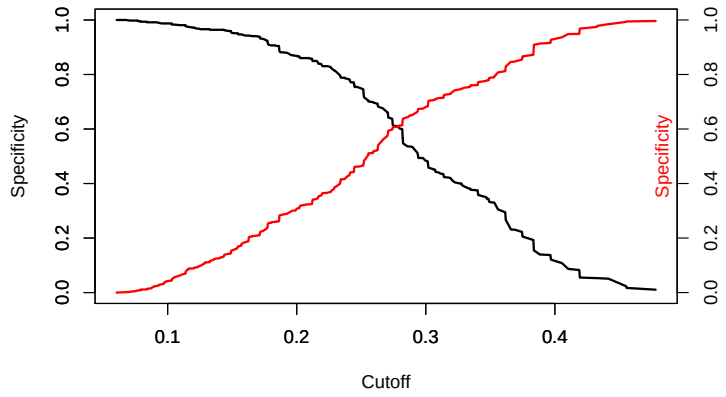


Figure 8: Plot for Cutoff Point

We notice that the best cut-off point for the probit model is 0.3. With that we calculate the predictions of this model and draw the confusion matrix.

Table 9: Predicted and observed values for Logit model

	Obs. 0	Obs. 1
Pred 0	802	194
Pred 1	492	278

Table 10: Predicted and observed values for Probit model

	Obs. 0	Obs. 1
Pred 0	884	245
Pred 1	410	227

With the probit model 818 individuals who did not actually vote for the social democrats were actually predicted correctly, however the model got 476 individuals who did not actually vote for the social democrats wrong. The model now is able to predict correctly 265 of the individuals who voted for the social democrats, however it failed with 107 individuals from this category.

So the McFadden R2 for our model is 0.040293 and the accuracy for our model is 0.6291053

The probit model's predictions are better with individuals who voted for the social democrats and for those who did not compared to the logit model. The probit model has a slightely better accuracy than the logit model so we conclude that the probit model is better than the logit model.

5 Conclusion

In this assignment, we have performed a study regarding voting habits, based on a dataset from the European Social Survey, from which we used a small number of variables. We focused on the Social Democratic Party of Sweden and this became our dependant variable - weather an individual in our sample had voted for this party or not.

For modeling, we used Logit and Probit models. We chose a couple of variables we thought would have an impact on voting for the Social Democrats, such as membership in a trade union, or measured political interest. Prediction is used to determine how much the chance of voting for Social Democrats changed when we change the values of a single individual, and found that in some cases this can impact quite a lot, such as one example: If the individual was living alone, then s/he would have a 4.67% lower chance to vote for the Social Democrats compared to if s/he would have a spouse.

A likelihood ratio test was also applied to our variables and only one of our variables has a non-constant scale. We evaluated our logit model by measuring its accuracy, which was found to be 61.1%. Finally, we compared our logit model with a probit model using the same dataset. We found the accuracy of our probit model to be 61.3% - a difference of 0.02% percent, so rather small difference, but a difference nonetheless.

6 Appendix

6.1 Predicted outcome in the Probit model

Table 11: Predicted outcome in the Probit model (Head 20)

observed	predicted
0	0
0	0
0	0
1	1
0	0
0	1
0	0
0	0
0	0
1	1
0	0
0	0
0	0
1	0
1	0
0	0
0	0
0	0
0	0
1	1

6.2 Code for the whole study

```
## ----setup, include=FALSE, message=FALSE-----------------------------------------------------------
knitr::opts_chunk$set(fig.pos = 'H', echo = FALSE, warning = FALSE)
library(knitr)
library(ggplot2)
library(tidyverse)
library(dummies)
library(DT)
library(lmtest)
library(sjPlot)
library(ROCR)
library(PDCRE)
library(caret)
library(e1071)


## ----fig.cap="Linear Regression Model", out.width="50%", fig.align="center"------------------------
table = data.frame(x = 1:10,
                   y = c(rep(0,5), rep(1,5)))
table %>%
```

```r
   ggplot(aes(x = x, y = y)) +
   geom_point(color = "red") +
   theme_bw() +
   geom_abline(aes(intercept=-0.3, slope=0.15), color = "blue")


## ----fig.cap="Logit Model", out.width="50%", fig.align="center"----------------------------------
table %>%
   ggplot(aes(x = x, y = y)) +
   geom_point(color = "red") +
   theme_bw() +
   geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE)


## ------------------------------------------------------------------------------------------------
table = data.frame(Obs_Pred = c("Pred 0","Pred 1"),
                   Obs_0 = c("True Negative(TN)","False Positive(FP)"),
                   Obs_1 = c("False Negative(FN)","True Negative(TN)"))

# Then for PDF:
kable(table, caption = "Confusion Matrix", digits = 2)


## ------------------------------------------------------------------------------------------------
dat_vote <- read.csv("vote_parties.csv", na.strings = "No answer")
dat <- read.csv("vote_parties.csv", na.strings = c("No answer", "Don't know"))
dat <- na.omit(dat)
names(dat)[names(dat) == "a5"] <- "PaperTime"
names(dat)[names(dat) == "b1"] <- "PoInt"
names(dat)[names(dat) == "b11"] <- "VoteLast"
names(dat)[names(dat) == "b12"] <- "VoteParty"
names(dat)[names(dat) == "c1"] <- "HapScale"
names(dat)[names(dat) == "c5"] <- "Crime5year"
names(dat)[names(dat) == "c15"] <- "HealthLvl"
names(dat)[names(dat) == "c28"] <- "SweBorn"
names(dat)[names(dat) == "c33"] <- "FSweBorn"
names(dat)[names(dat) == "c35"] <- "MSweBorn"
names(dat)[names(dat) == "d4"] <- "Spouse3year"
names(dat)[names(dat) == "d6"] <- "Married"
names(dat)[names(dat) == "d9"] <- "TotalChildren"
names(dat)[names(dat) == "e4"] <- "OptLvl"
names(dat)[names(dat) == "e5"] <- "PosLvl"
names(dat)[names(dat) == "e7"] <- "SatLvl"
names(dat)[names(dat) == "f1"] <- "FamSize"
names(dat)[names(dat) == "f2"] <- "Gender"
names(dat)[names(dat) == "f3"] <- "Bdate"
names(dat)[names(dat) == "f5"] <- "Area"
names(dat)[names(dat) == "f27"] <- "FormUnemp"
names(dat)[names(dat) == "f30"] <- "TradeUn"
names(dat)[names(dat) == "f32"] <- "IncLvl"
```

```
## ------------------------------------------------------------------------------------------------

dat_t = dat

dat_t$PaperTime = as.character(dat_t$PaperTime)
dat_t$PaperTime[dat_t$PaperTime == "No time at all"] = 0
dat_t$PaperTime[dat_t$PaperTime == "Less than 0,5 hour"] = 1
dat_t$PaperTime[dat_t$PaperTime == "0,5 hour to 1 hour"] = 2
dat_t$PaperTime[dat_t$PaperTime == "More than 1 hour, up to "] = 3
dat_t$PaperTime[dat_t$PaperTime == "More than 1,5 hours, up "] = 4
dat_t$PaperTime[dat_t$PaperTime == "More than 2 hours, up to"] = 5
dat_t$PaperTime[dat_t$PaperTime == "More than 2,5 hours, up to"] = 6
dat_t$PaperTime[dat_t$PaperTime == "More than 2,5 hours, up "] = 6
dat_t$PaperTime[dat_t$PaperTime == "More than 3 hours"] = 7
dat_t$PaperTime[dat_t$PaperTime == "Don't know"] = 88

dat_t$PoInt = as.character(dat_t$PoInt)
dat_t$PoInt[dat_t$PoInt == "Very interested"] = 1
dat_t$PoInt[dat_t$PoInt == "Quite interested"] = 2
dat_t$PoInt[dat_t$PoInt == "Hardly interested"] = 3
dat_t$PoInt[dat_t$PoInt == "Not at all interested"] = 4
dat_t$PoInt[dat_t$PoInt == "Don't know"] = 88

dat_t$VoteLast = as.character(dat_t$VoteLast)
dat_t$VoteLast[dat_t$VoteLast == "Yes"] = 1
dat_t$VoteLast[dat_t$VoteLast == "No"] = 2
dat_t$VoteLast[dat_t$VoteLast == "Not eligible to vote"] = 3
dat_t$VoteLast[dat_t$VoteLast == "Don't know"] = 88

dat_t$HapScale = as.character(dat_t$HapScale)
dat_t$HapScale[dat_t$HapScale == "Extremely unhappy"] = 0
dat_t$HapScale[dat_t$HapScale == "Extremely happy"] = 10
dat_t$HapScale[dat_t$HapScale == "Refusal"] = 99
dat_t$HapScale[dat_t$HapScale == "Don't know"] = 88

dat_t$HealthLvl = as.character(dat_t$HealthLvl)
dat_t$HealthLvl[dat_t$HealthLvl == "Very good"] = 1
dat_t$HealthLvl[dat_t$HealthLvl == "Good"] = 2
dat_t$HealthLvl[dat_t$HealthLvl == "Fair"] = 3
dat_t$HealthLvl[dat_t$HealthLvl == "Bad"] = 4
dat_t$HealthLvl[dat_t$HealthLvl == "Very bad"] = 5
dat_t$HealthLvl[dat_t$HealthLvl == "Don't know"] = 88

dat_t$SweBorn = as.character(dat_t$SweBorn)
dat_t$SweBorn[dat_t$SweBorn == "Yes"] = 1
dat_t$SweBorn[dat_t$SweBorn == "No"] = 2
dat_t$SweBorn[dat_t$SweBorn == "Don't know"] = 88

dat_t$FSweBorn = as.character(dat_t$FSweBorn)
dat_t$FSweBorn[dat_t$FSweBorn == "Yes"] = 1
dat_t$FSweBorn[dat_t$FSweBorn == "No"] = 2
```

```r
dat_t$FSweBorn[dat_t$FSweBorn == "Don't know"] = 88

dat_t$MSweBorn = as.character(dat_t$MSweBorn)
dat_t$MSweBorn[dat_t$MSweBorn == "Yes"] = 1
dat_t$MSweBorn[dat_t$MSweBorn == "No"] = 2
dat_t$MSweBorn[dat_t$MSweBorn == "Don't know"] = 88

dat_t$Crime5year = as.character(dat_t$Crime5year)
dat_t$Crime5year[dat_t$Crime5year == "Yes"] = 1
dat_t$Crime5year[dat_t$Crime5year == "No"] = 2
dat_t$Crime5year[dat_t$Crime5year == "Don't know"] = 88

dat_t$Spouse3year = as.character(dat_t$Spouse3year)
dat_t$Spouse3year[dat_t$Spouse3year == "Yes"] = 1
dat_t$Spouse3year[dat_t$Spouse3year == "No"] = 2
dat_t$Spouse3year[dat_t$Spouse3year == "Don't know"] = 88

dat_t$Married = as.character(dat_t$Married)
dat_t$Married[dat_t$Married == "Yes"] = 1
dat_t$Married[dat_t$Married == "No"] = 2

dat_t$TotalChildren = as.character(dat_t$TotalChildren)
dat_t$TotalChildren[dat_t$TotalChildren == "Not applicable"] = 0
dat_t$TotalChildren[dat_t$TotalChildren == "1.000000"] = 1
dat_t$TotalChildren[dat_t$TotalChildren == "2.000000"] = 2
dat_t$TotalChildren[dat_t$TotalChildren == "3.000000"] = 3
dat_t$TotalChildren[dat_t$TotalChildren == "4.000000"] = 4
dat_t$TotalChildren[dat_t$TotalChildren == "5.000000"] = 5
dat_t$TotalChildren[dat_t$TotalChildren == "6.000000"] = 6
dat_t$TotalChildren[dat_t$TotalChildren == "7.000000"] = 7
dat_t$TotalChildren[dat_t$TotalChildren == "8.000000"] = 8
dat_t$TotalChildren[dat_t$TotalChildren == "9.000000"] = 9
dat_t$TotalChildren[dat_t$TotalChildren == "10.000000"] = 10
dat_t$TotalChildren[dat_t$TotalChildren == "11.000000"] = 11
dat_t$TotalChildren[dat_t$TotalChildren == "12.00000"] = 12

dat_t$OptLvl = as.character(dat_t$OptLvl)
dat_t$OptLvl[dat_t$OptLvl == "Agree strongly"] = 1
dat_t$OptLvl[dat_t$OptLvl == "Agree"] = 2
dat_t$OptLvl[dat_t$OptLvl == "Neither agree nor disagr"] = 3
dat_t$OptLvl[dat_t$OptLvl == "Disagree"] = 4
dat_t$OptLvl[dat_t$OptLvl == "Disagree strongly"] = 5
dat_t$OptLvl[dat_t$OptLvl == "Don't know"] = 88

dat_t$PosLvl = as.character(dat_t$PosLvl)
dat_t$PosLvl[dat_t$PosLvl == "Agree strongly"] = 1
dat_t$PosLvl[dat_t$PosLvl == "Agree"] = 2
dat_t$PosLvl[dat_t$PosLvl == "Neither agree nor disagr"] = 3
dat_t$PosLvl[dat_t$PosLvl == "Disagree"] = 4
dat_t$PosLvl[dat_t$PosLvl == "Disagree strongly"] = 5
dat_t$PosLvl[dat_t$PosLvl == "Don't know"] = 88
```

```r
dat_t$SatLvl = as.character(dat_t$SatLvl)
dat_t$SatLvl[dat_t$SatLvl == "Agree strongly"] = 1
dat_t$SatLvl[dat_t$SatLvl == "Agree"] = 2
dat_t$SatLvl[dat_t$SatLvl == "Neither agree nor disagr"] = 3
dat_t$SatLvl[dat_t$SatLvl == "Disagree"] = 4
dat_t$SatLvl[dat_t$SatLvl == "Disagree strongly"] = 5
dat_t$SatLvl[dat_t$SatLvl == "Don't know"] = 88

dat_t$Gender = as.character(dat_t$Gender)
dat_t$Gender[dat_t$Gender == "Male"] = 1
dat_t$Gender[dat_t$Gender == "Female"] = 2

dat_t$FormUnemp = as.character(dat_t$FormUnemp)
dat_t$FormUnemp[dat_t$FormUnemp == "Yes"] = 1
dat_t$FormUnemp[dat_t$FormUnemp == "No"] = 2
dat_t$FormUnemp[dat_t$FormUnemp == "Don't know"] = 88

dat_t$TradeUn = as.character(dat_t$TradeUn)
dat_t$TradeUn[dat_t$TradeUn == "Yes, currently"] = 1
dat_t$TradeUn[dat_t$TradeUn == "Yes, previously"] = 2
dat_t$TradeUn[dat_t$TradeUn == "No"] = 3
dat_t$TradeUn[dat_t$TradeUn == "Don't know"] = 88

dat_t$IncLvl = as.character(dat_t$IncLvl)
dat_t$IncLvl[dat_t$IncLvl == "J"] = 1
dat_t$IncLvl[dat_t$IncLvl == "R"] = 2
dat_t$IncLvl[dat_t$IncLvl == "C"] = 3
dat_t$IncLvl[dat_t$IncLvl == "M"] = 4
dat_t$IncLvl[dat_t$IncLvl == "F"] = 5
dat_t$IncLvl[dat_t$IncLvl == "S"] = 6
dat_t$IncLvl[dat_t$IncLvl == "K"] = 7
dat_t$IncLvl[dat_t$IncLvl == "P"] = 8
dat_t$IncLvl[dat_t$IncLvl == "D"] = 9
dat_t$IncLvl[dat_t$IncLvl == "H"] = 10
dat_t$IncLvl[dat_t$IncLvl == "U"] = 11
dat_t$IncLvl[dat_t$IncLvl == "N"] = 12
dat_t$IncLvl[dat_t$IncLvl == "Refusal"] = 77
dat_t$IncLvl[dat_t$IncLvl == "Don't know"] = 88

dat_t$Area = as.character(dat_t$Area)
dat_t$Area[dat_t$Area == "A big city"] = 1
dat_t$Area[dat_t$Area == "Suburbs or outskirts of "] = 2
dat_t$Area[dat_t$Area == "Town or small city"] = 3
dat_t$Area[dat_t$Area == "Country village"] = 4
dat_t$Area[dat_t$Area == "Farm or home in countrys"] = 5
dat_t$Area[dat_t$Area == "Don't know"] = 88


## ----out.width="80%"-----------------------------------------------------------------------------
dat %>%
  filter(VoteParty != "NOT APPLICABLE") %>%
```

```
  filter(VoteParty != "REFUSAL") %>%
  filter(VoteParty != "DONT KNOW") %>%
  filter(VoteParty != "NO ANSWER") %>%
  count(VoteParty) %>%
  mutate(VoteParty = fct_reorder(VoteParty, n, .desc = FALSE)) %>%
  ggplot() +
  geom_bar(aes(x = VoteParty, y = (n)/sum(n), fill = VoteParty), stat = 'identity') +
  scale_y_continuous(labels=scales::percent) +
  ylab("Percentage") +
  guides(fill = "none") +
  coord_flip() +
  theme_bw()


## ------------------------------------------------------------------------------------------------

table = data.frame(VarName = c("TradeUn", "PoInt", "FamSize", "Area"),
                   Measure = c("Trade Union membership", "Political interest", "Size of family", "Urb
                   ExpEffect = c("Neg.", "Neg.", "Pos.", "Pos."))

kable(table, caption = "Chosen variables", digits = 2)



## ----echo=FALSE----------------------------------------------------------------------------------

dat_t$PoInt <- as.numeric(dat_t$PoInt)
dat_t$TradeUn <- as.numeric(dat_t$TradeUn)
dat_t$FamSize <- as.numeric(dat_t$FamSize)
dat_t$Area <- as.numeric(dat_t$Area)

m <- glm(as.factor(socialdemocrats) ~ PoInt+TradeUn+FamSize+Area, data = dat_t,
         family = binomial(link='logit'))

# Getting log likelihood value (for test of likert restriction)


## ------------------------------------------------------------------------------------------------
table2 = data.frame(Coefficients = c("Intercept", "PoInt", "TradeUn", "FamSize", "Area"),
                   Estimate = c(-0.50969, 0.17109, -0.53338 , -0.10878, 0.06648),
                   Significant = c("-", "Yes", "Yes", "Yes", "No"))

kable(table2, caption = "Logit Model", digits = 2)


## ------------------------------------------------------------------------------------------------
preds <- predict.glm(m, dat_t, type="response")


## ----eval=FALSE----------------------------------------------------------------------------------
## # Predict individ 18
```

```
## id18 = dat_t[18, ]
## # response returns predictions as probability (not log odds)
## pred1 <- predict.glm(m, id18, type="response")
##
## # change a value of id15 to
## id18$PoInt = 2
## pred2 = predict.glm(m, id18, type="response")
##
## # Difference in probability
## pred1-pred2


## ----eval=FALSE----------------------------------------------------------------------------------
## # Predict individ 18
## id18 = dat_t[18, ]
## # response returns predictions as probability (not log odds)
## pred3 <- predict.glm(m, id18, type="response")
##
## # change a value of id15 to
## id18$TradeUn = 2
## pred4 = predict.glm(m, id18, type="response")
##
## # Difference in probability
## pred3-pred4


## ----eval=FALSE----------------------------------------------------------------------------------
## # Predict individ 18
## id18 = dat_t[18, ]
## # response returns predictions as probability (not log odds)
## pred5 <- predict.glm(m, id18, type="response")
##
##
## # change a value of id18 to
## id18$FamSize = 5
## pred6 = predict.glm(m, id18, type="response")
##
## # Difference in probability
## pred5-pred6


## ------------------------------------------------------------------------------------------------
table3 = data.frame(Nr = c(1, 2, 3, 4, 5),PoInt = c(0.03394278, 0, -0.03646351, -0.0751499 , NA),
                    TradeUn = c(0, 0.09676355, 0.1668725, NA, NA),
                    FamSize = c(-0.04676795, -0.0229091 , 0, 0.02188785, 0.04270068 ))

kable(table3, caption = "Changes in individual #18", digits = 2)


## ------------------------------------------------------------------------------------------------
```

```r
m1 <- glm(as.factor(socialdemocrats) ~ as.factor(PoInt)+TradeUn+FamSize+Area, data = dat_t,
          family = binomial(link='logit'))


lr = lrtest(m, m1)


## ----fig.cap="Chisquare Distribution", out.width="60%", fig.align="center"-----------------------
dist_chisq(p = 0.05, deg.f = 2)


## ----------------------------------------------------------------------------------------------


m1 <- glm(as.factor(socialdemocrats) ~ PoInt+as.factor(TradeUn)+FamSize+Area, data = dat_t,
          family = binomial(link='logit'))



lr = lrtest(m, m1)


## ----fig.cap="Chisquare Distribution", out.width="60%", fig.align="center"-----------------------
dist_chisq(p = 0.05, deg.f = 1)


## ----------------------------------------------------------------------------------------------


m1 <- glm(as.factor(socialdemocrats) ~ PoInt+TradeUn+as.factor(FamSize)+Area, data = dat_t,
          family = binomial(link='logit'))



lr = lrtest(m, m1)


## ----fig.cap="Chisquare Distribution", out.width="60%", fig.align="center"-----------------------
dist_chisq(p = 0.05, deg.f = 7)


## ----------------------------------------------------------------------------------------------
m1 <- glm(as.factor(socialdemocrats) ~ PoInt+TradeUn+FamSize+as.factor(Area), data = dat_t, family =
m1_null <- glm(as.factor(socialdemocrats) ~ 1, data = dat_t, family = binomial(link='logit'))
R2_logit = 1 - logLik(m1)/logLik(m1_null)
lr = lrtest(m, m1)


## ----fig.cap="Chisquare Distribution", out.width="60%", fig.align="center"-----------------------
dist_chisq(p = 0.05, deg.f = 1)
```

```
## ------------------------------------------------------------------------------------------------------
preds <- predict.glm(m1, dat_t, type="response")
for(i in 1:length(preds))
{
  if (preds[i]<0.5)
  {
    preds[i]=0
  }
  else {preds[i]=1}
}

preds <- as.data.frame(preds)
pred_obs <- cbind(dat_t$socialdemocrats, preds)
pred_obs <- as.data.frame(pred_obs)
cm =confusionMatrix(as.factor(pred_obs$preds), as.factor(pred_obs$`dat_t$socialdemocrats`), positive


## ------------------------------------------------------------------------------------------------------
rownames(cm$table) = c("Pred 0", "Pred 1")
colnames(cm$table) = c("Obs. 0", "Obs. 1")
kable(cm$table, caption = "Predicted and observed values", digits = 2)


## ----fig.cap="Plot for Cut-off Point", out.width="60%", fig.align="center"-------------------------
preds <- predict.glm(m1, dat_t, type="response")
pred = prediction(preds, dat_t$socialdemocrats)
plot(unlist(performance(pred, "sens")@x.values), unlist(performance(pred, "sens")@y.values),
     type="l", lwd=2, ylab="Specificity", xlab="Cutoff")
par(new=TRUE)
plot(unlist(performance(pred, "spec")@x.values), unlist(performance(pred, "spec")@y.values),
     type="l", lwd=2, col='red', ylab="", xlab="")
axis(4, at=seq(0,1,0.2))
mtext("Specificity",side=4, padj=-2, col='red')


## ------------------------------------------------------------------------------------------------------
preds <- predict.glm(m1, dat_t, type="response")
for(i in 1:length(preds))
{
  if (preds[i]<0.28)
  {
    preds[i]=0
  }
  else {preds[i]=1}
}

preds <- as.data.frame(preds)
pred_obs <- cbind(dat_t$socialdemocrats, preds)
pred_obs <- as.data.frame(pred_obs)
cm =confusionMatrix(as.factor(pred_obs$preds), as.factor(pred_obs$`dat_t$socialdemocrats`), positive
```

```r
## ------------------------------------------------------------------------------------------------------
rownames(cm$table) = c("Pred 0", "Pred 1")
colnames(cm$table) = c("Obs. 0", " 1")
kable(cm$table, caption = "Predicted and observed values", digits = 2)


## ------------------------------------------------------------------------------------------------------
m <- glm(as.factor(socialdemocrats) ~ PoInt+TradeUn+FamSize+as.factor(Area), data = dat_t,
         family = binomial(link='probit'))
m_null = glm(as.factor(socialdemocrats) ~ 1, data = dat_t, family = binomial(link='probit'))
R2_probit = 1 - logLik(m)/logLik(m_null)


## ------------------------------------------------------------------------------------------------------
table2 = data.frame(Coefficients = c("Intercept", "PoInt", "TradeUn", "FamSize", "Area"),
                    Estimate = c(-0.33873, 0.10292, -0.30996 , -0.06322, 0.04038),
                    Significant = c(NA, "Yes", "Yes", "Yes", "No"))

kable(table2, caption = "Probit Model", digits = 2)


## ----fig.cap="Plot for Cutoff Point", out.width="60%", fig.align="center"-------------------------
preds <- predict.glm(m, dat_t, type="response")
pred = prediction(preds, dat_t$socialdemocrats)
plot(unlist(performance(pred, "sens")@x.values), unlist(performance(pred, "sens")@y.values),
     type="l", lwd=2, ylab="Specificity", xlab="Cutoff")
par(new=TRUE)
plot(unlist(performance(pred, "spec")@x.values), unlist(performance(pred, "spec")@y.values),
     type="l", lwd=2, col='red', ylab="", xlab="")
axis(4, at=seq(0,1,0.2))
mtext("Specificity",side=4, padj=-2, col='red')


## ------------------------------------------------------------------------------------------------------
preds <- predict.glm(m, dat_t, type="response")
for(i in 1:length(preds))
{
  if (preds[i]<0.3)
  {
    preds[i]=0
  }
  else {preds[i]=1}
}

preds <- as.data.frame(preds)
pred_obs <- cbind(dat_t$socialdemocrats, preds)
pred_obs <- as.data.frame(pred_obs)


cm1 =confusionMatrix(as.factor(pred_obs$preds), as.factor(pred_obs$`dat_t$socialdemocrats`), positive
```

```r
## ------------------------------------------------------------------------------------------------------------
rownames(cm$table) = c("Pred 0", "Pred 1")
colnames(cm$table) = c("Obs. 0", "Obs. 1")
rownames(cm1$table) = c("Pred 0", "Pred 1")
colnames(cm1$table) = c("Obs. 0", "Obs. 1")
kable(cm$table, caption = "Predicted and observed values for Logit model", digits = 2)
kable(cm1$table, caption = "Predicted and observed values for Probit model", digits = 2)


## ------------------------------------------------------------------------------------------------------------
table = pred_obs
names(table) = c("observed", "predicted")
kable(table[1:20,], caption = "Predicted outcome in the Probit model (Head 20)")


## ------------------------------------------------------------------------------------------------------------
library(InformationValue)
optCutOff <- optimalCutoff(dat_t$socialdemocrats, preds)[1]


## ------------------------------------------------------------------------------------------------------------
#purl("HW1.Rmd", output = "code1.R")
```