Probability of a Term Deposit.

ST404 Assignment 4 Logistic Regression

Executive Summary

This summary will explain how I selected the final model's variables from a choice of 12 to effectively predict whether a customer will accept a term deposit from the latest campaign. We only care if they will subscribe and not the actual amount if they do. Naturally, an equation which can give us a probability of success (where the customer says 'yes') will be adequate. We will in fact express this probability as an odds ratio OR which is equal to the probability of success divided by failure. Our final model is:

$$\left(\frac{P(Yes)}{P(No)}\right) = \mathbf{0.0565}$$
$$* \mathbf{0.699}(bluecollar) * \mathbf{2.143}(retired) * \mathbf{2.176}(student) * \mathbf{0.812}(unemployed)$$
$$* \mathbf{1.477}(tertiaryeducation)$$
$$* \mathbf{0.528}(housingloan)$$
$$* \mathbf{0.488}(personalloan)$$
$$* \mathbf{1.004}\hat{}(durationoflastcontactinseconds)$$
$$* \mathbf{1.592}(poutcomeother) * \mathbf{10.329}(poutcomesuccess) * \mathbf{0.568}(poutcomeunknown)$$

where our base individual is a white collar, non tertiary educated person with no loans who did not subscribe to a product in the previous campaign and who was not contacted at all in this current campaign. He/she would have an odds ratio of 0.0565, the intercept i.e. there is a 5.35% chance they will accept.

Here is a more basic example. Consider two individuals where one was successfully targeted in the previous campaign and the other was a failure. If we hold the other variables constant, we can see that the one who was succesfully targeted before is now between (6.29,16.96) times more likely to say yes to the term deposit in the latest campaign 95% of the time.

A quick look at the data tells us that 521/4521 people said yes so we would expect a random person to have have roughly 1/9 chance of saying yes. This is good variability and sample size so we can be confident in continuing.

Exploratory analysis of the explanatory variables generally indicated what was to be expected, such as older people are more likely to be married and younger people are likely to be single. Many typical stereotypes were confirmed especially in the job variable where the 12 different jobs showed high variation when compared to the other explanatory variables. Intuitively, knowing someone's job and marital status can tell you a lot of other information about them such as age. This was an early sign that there could be a lot of overlap in information used to predict the OR and to remediate this I considered collapsing some variables for simplicity. Certainly, the job variable was very hard to interpret given its current state.

By performing univariable analysis on each variable vs the response I could quickly determine which variables offered little benifit in predicting. The intuition is that if the variable alone cannot be useful in predicting then it would be unlikely to be useful in a multivariate model. Both default and balance variables stood out as being of little use so they were marked for potential removal. Of course, we have to check for confounding before removing them as they may have an efffect on other variables.

Running a simple multivaraite model with the original 12 variables unchanged vs the response and then comparing to the univariate models can help check for confounding. If the confidence intervals of the coefficients, odds ratios and standard errors noticeably change then there could be a confounding variable. By comparing the odds ratio confidence intervals side by side, it was obivous that many of the job variables fluctuated in the multivarariate model as did education.

There is evidence to warrant recollapsing of the job variable. There was an elegant and natural way of regrouping many of the specific jobs into a white collar group which complements the existing blue collar group. The white collar group also contained those with similar odds ratios which made even more sense for this grouping. Education was also expressed as those with tertiary education and those without instead of primary, seconday, tertiary and unknown. This was because the tertiary group has a stark difference compared to the other groups. I also confirmed to remove default and balance without even trying to group them first and retesting.

With 10 variables, 2 of which have been regrouped it was necessary to check for collinearity to try and remove any more unneccessary variables. Correlations between the continuos variables showed almost no correlation and chi square tests of independence showed that the job, marital and education variables had higher dependence. Previous outcome was the most independent variable compared to others, and given its strong predictive power I was quite sure that this variable would be very important in the final model.

To get a sense of direction I used automatic variable selecting algorithms based on AIC and BIC, which are measures of the tradeoff between the goodness of fit of potential models and their complexity. This gave us a rough boundary where it was suggested the model at least include job, loan, housing, poutcome (previous outcome) and duration with maybe education, marital and campaign aswell in order of most useful to least. Likelihood ratio tests which were done manually also suggested that these sets of variables were relevant enough to be included, and the others (age and previous) should be removed.

My approach was to create a base model with the 5 minimum variables suggested with education aswell, so a total of 6 variables. I was able to execute a hosmer lemehow test which showed that the model had a sensitivity (correctly predicting yes) of 30.13 % and a specificity (correctly predicting no) of 97.8%. This was assuming a cutoff point of 50%. So, if the fitted values are greater than 0.5 we would assume the model is prediciting that they will say yes. Likewise, a fitted value of less than 0.5 would mean a no prediction for that individual. How does the sensitivity and specificity change as the cutoff changes?

A receiver operating characteristic ROC curve tells us exactly this. It is used in signal detection thoery and can be applied to this case as we are detecting whether we have correctly predicted yes or no. We can calculate the area under the curve AUC which is equal to the probability that the model will rank a randomly chosen yes individual higher than a randomly chosen one who says no. If we find the number of possible pairs of yes and no (521*4000) and divide the Wilcoxon statistic by this combination we shall obtain the AUC = 87.24% which is a good result. We are very likely to predict the yes result with this model given any pair of contrasting individuals.

Possible interactions were not ignored. Even though a likelihood bases test showed that an interaction between duration and poutcome could be useful, I felt that on the whole the model was approaching it's upper limit of desired complexity. It would also add 3 more variables in our final model which would be cumbersome to interpret.

Satisfied with the model fit and without the need for interactions, the base model was left unchanged and was the final model.

Customer service is crucial in any industry and especially banking where the most loyal customers will be the ones who are most likely to not only stay but also subscribe to new products. It comes as no surprise that the most important factors are if a customer has previously accepted a product and has had a longer duration of their last contact with the bank employees and advertisers. Of course, there is additional cost if longer time is spent on the line with a potential client. Therefore, is it important to start targeting the most likely groups first. Focusing on those without any existing loans will result in higher success as these groups are likely to have more liquid funds so they can subscribe to a term deposit. Retired people are especially likely to subscribe. Perhaps they are most easy to convince and be sold to. If cost is a major constraint, spending less time pursuing leads on blue collar workers and could prove to be cost saving. In short, spending quality time targeting a select few by considering mainly campaign related factors is much more effective than considering personal client variables such as age and marital status. By focusing more on campaign variables, we can see directly over time and over different campaigns which strategies are most effective, so there is scope for continued improvement.

Probability of a Term Deposit.

ST404 Assignment 4 Logistic Regression

Technical Report

Contents

1 Exploratory Analysis

1.1 Introduction

Direct marketing is an advertising strategy used in this case by a bank hoping to attract customers to make a term deposit. By targeting specific groups of individuals with certain characteristics, the bank can improve it's advertising success by giving more attention to those groups which are statistically more likely to accept. This marketing campaign also has another advantage. The bank now has data from previous campaigns, about the client and the previous campaign data for each individual, so it can alter its strategy in future. The main question is: which groups are the most likely to accept and how should the bank contact the client?

The explanatory variables we have at hand are client related variables: "age", "job", "marital", "education", "default", "balance", "housing", "loan". We also have variables related to the campaign: "duration", "campaign", "previous", "poutcome". We will use this data to help us predict whether the client says "yes" or "no" to a term deposit, and not necessarily how much they deposit if they do.

1.2 Response Variable "Outcome"

In the data this is labelled "y", I will use these terms interchangebly. This piece of data answers one question: "has the client subscribed a term deposit from the current campaign?" with answers "yes" and "no". With 4000 people saying "no" and 521 saying "yes" this is a large sample with good variabilty. We would expect roughly 1/9 chance for someone to so say "yes" if he/she was picked at random.

Figure 1 shows pairwise relationship between the response variable and the explanatory variables. If we take a look at the job graph we can also see that retired people are most likely to say yes compared to other jobs. Students are also likely to take up a term deposit, which is surprising given the stereotype of cash-strapped students. Figure 2 shows us clearly that as age increases the percentage of yes decreases very slowly but as soon as age 60 is reached the chances of yes dramatically increases. This cannot be a coincidence as this bank is based in Portugal where the retirement age of residents is 65. There is some variation amongst the y vs marital and y vs education graphs whereas default shows almost no variability. Duration has a steep trend. As the duration increases the chances of yes increases. "Previous" also exhibits this stepwise pattern though to a lesser extent. The y vs poutcome graph is very interesting. It shows that if the previous outcome was a "success" then the chances of a customer accepting this term deposit is roughly 60%. This could be a very important factor to observe.

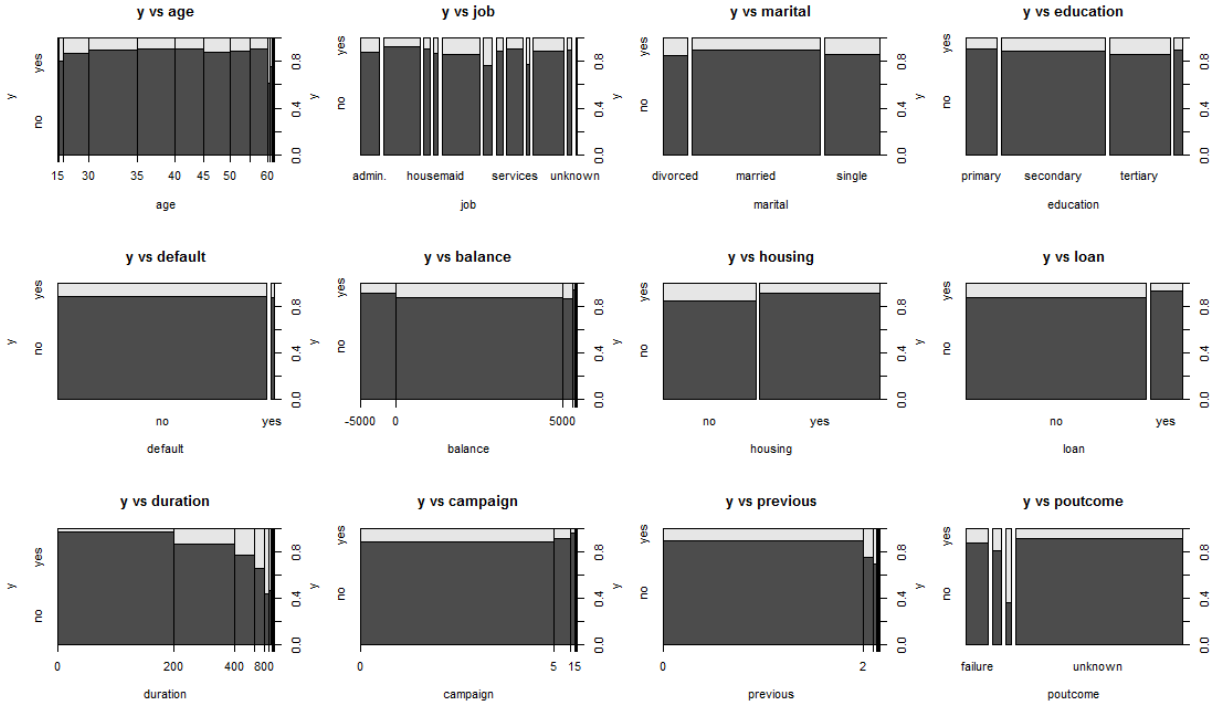

Figure 1: response y vs explanatory variables.

On the whole, campaign related variables may prove to be most useful though client variables certainly will improve our model. At this stage we can see already some potential collinearity and interactions. Below is a blown up image of y vs age.

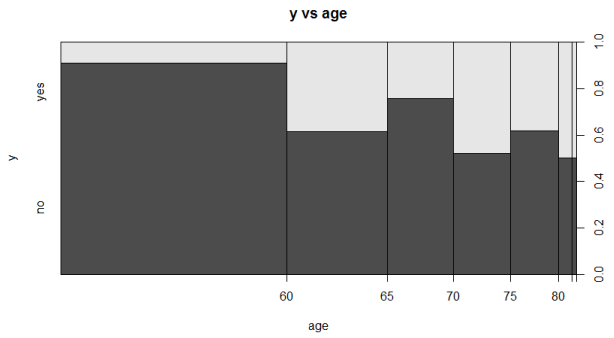

Figure 2: This shows y vs the 95th age percentile upwards.

1.3 Explanatory Variables

1.3.1 Age

Figure 3 shows pairwise relationship between the age variable and the explanatory variables. This figure shows us that those who are retired are around 60 years old and that single people tend to be be younger. This is as expected, and the other plots in this figure show us little variability. The campaign and previous plots show us that as middle aged people tend to have the largest figures for these variables. But that is most likely because the majority of individuals in the data are middle aged as age has a mean of 41.17 with interquartile ranges of 33, 39 and 49.

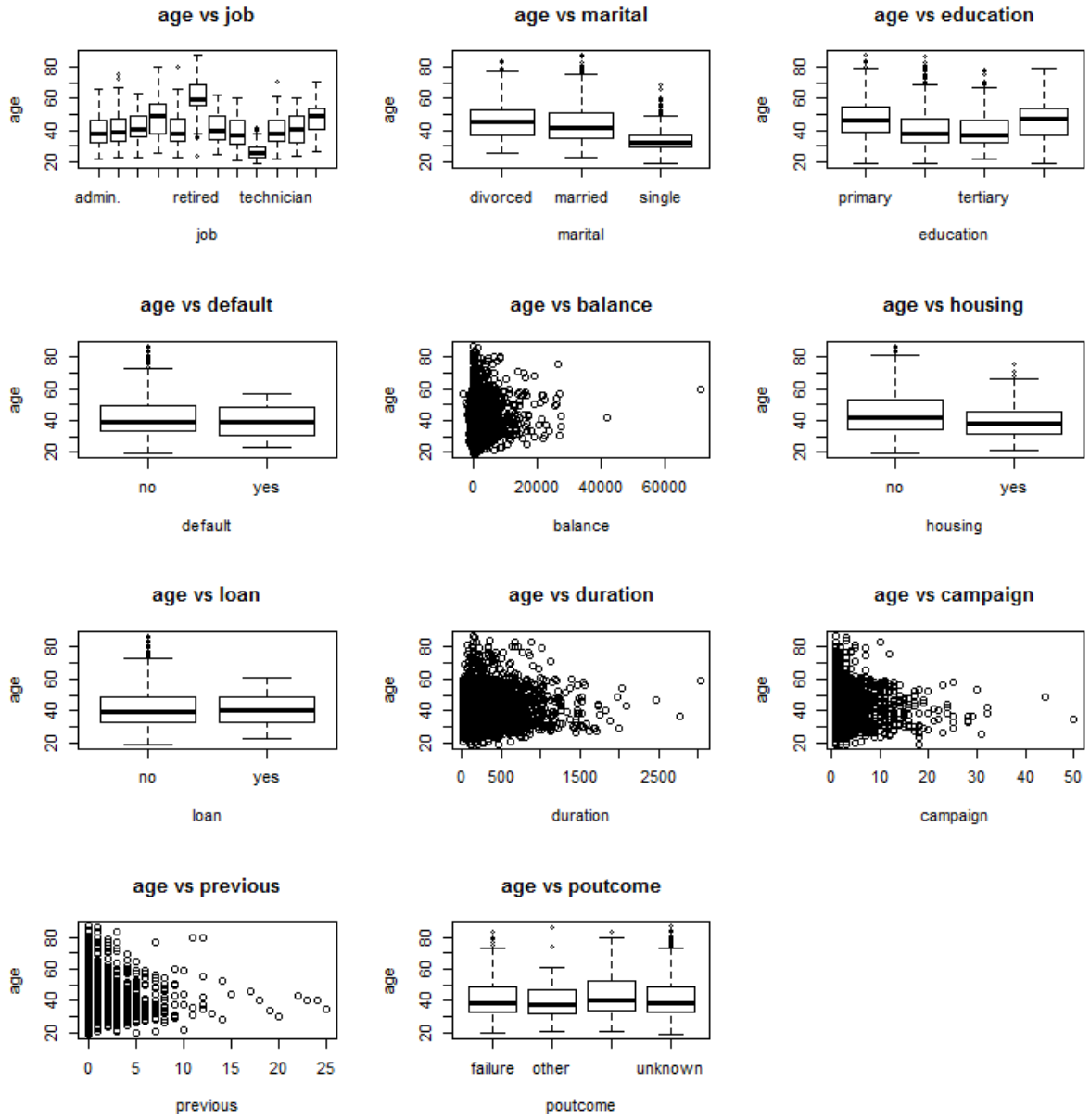

Figure3: age vs other explanatory variables

1.3.2 Job

These plots show us that there is high variation between job vs marital, education and housing. As one might expect, students tend to be single, without a house and have the lowest amount of people with just a primary education. Contrastingly, blue collar workers have the highest number of individuals with primary education but they also seem to be more stable in terms of having the highest amount of housing loans and more likely to be married. Housemaids are an interesting case, as the vast majority will be women so there could be possible confounding. It may be that their gender is mostly affecting their results. In fact, this could be said of other jobs such as blue collar. I would argue that housemaid should be in blue collar. The wikipedia definition of blue collar is "A blue-collar worker is a member of the working class who performs manual labor. Blue-collar work may involve skilled or unskilled, manufacturing, mining, construction, mechanical, maintenance, technical installation and many other types of physical work.

Often something is physically being built or maintained." Housemaids fall into criteria by this definition. Clearly there are far too many different job variables which could be confusing for advertising data inputters who may assign people incorrectly. I believe there is potential and motivation for grouping jobs together. For instance, there are obvious overlaps between self employment and entrepeneur. Confusion may arise from individuals reporting their job and also those who collect the information.

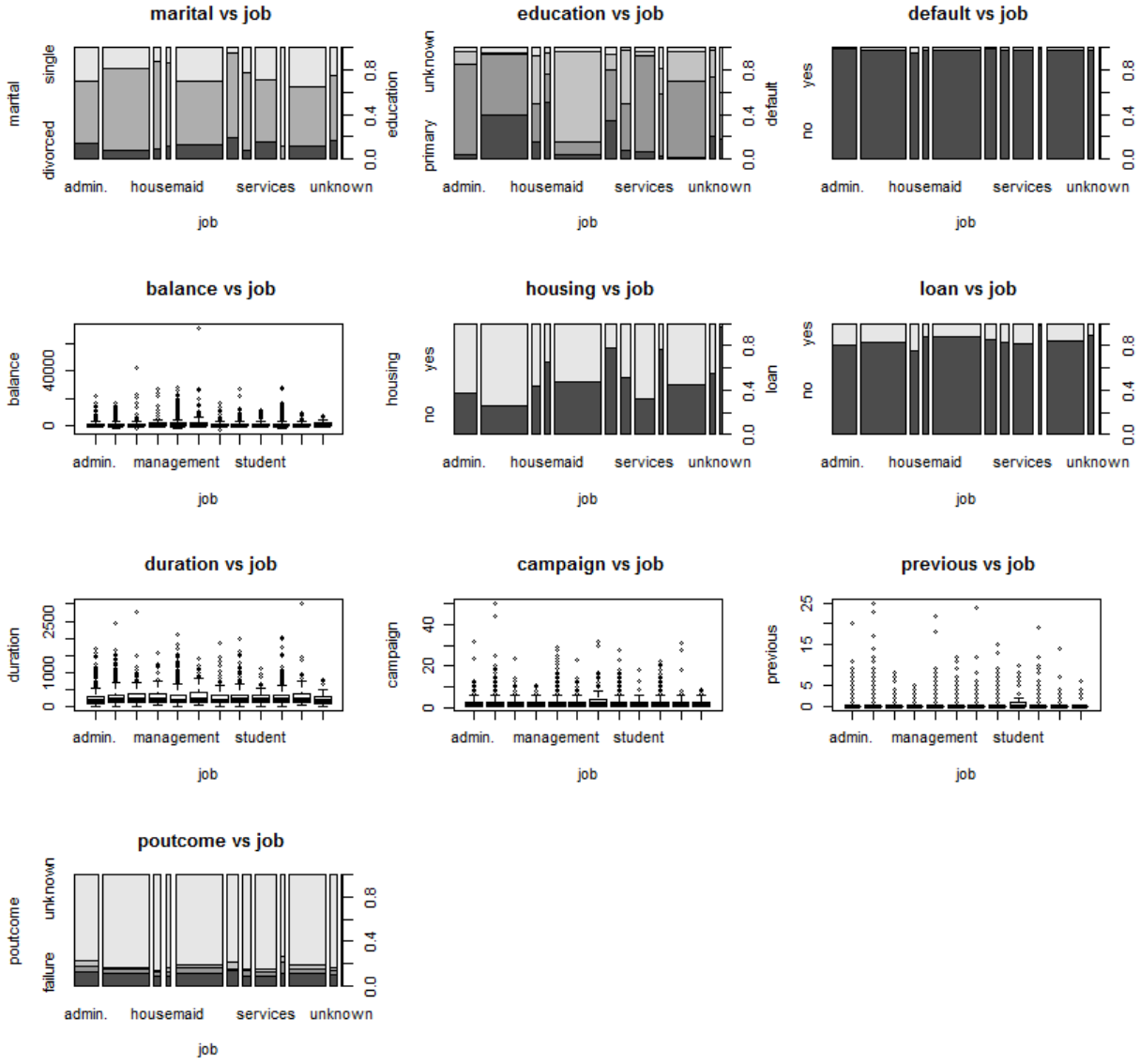

Figure 4: This shows other explanatory variables vs job. Job has too many categories so is on the x axes.

1.3.3 Marital

These plots show that there is little variation amongst marriage vs other explanatory variables. However, marital status and education show a pattern. Divorced rates are the same but as the level of education increases the percentage of married people are lower and consequently the percentage of singles increases. If an individual has a high balance then they are more likely to single individuals.

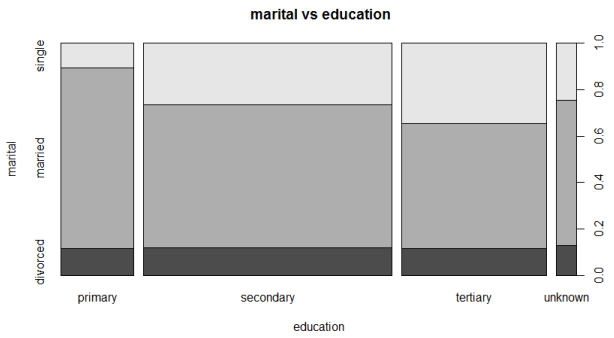

Figure 5: marital vs education. Notice the drop in marriage percentage as known individuals have higher levels of education.

1.3.4 Education

There is little variation amongst those with primary education. Whereas those with defaults, loans and housing loans are less likely to be tertiary educated individuals. Those with higher balances tend to be those with tertiary education.

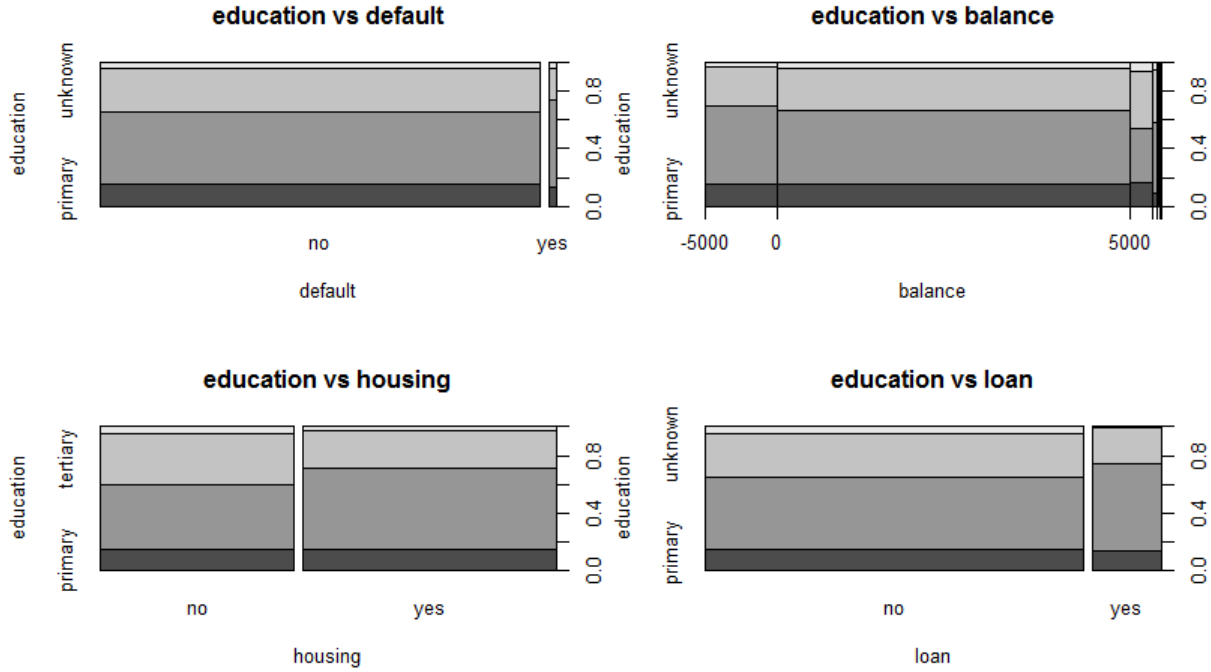

Figure 6: education vs relevant explanatory variables.

1.3.5 Default

This piece of data shows that there were 4445 people with no defaults and 76 with. 76 is a relatively low number, and this data may be inconclusive unless an unusally high or low amount of positive coutcomes are received from those with defaults. A table shows us that of the 76 who have defaults, 9 said yes to a term deposit which is roughly what we would expect from a random sample anyway. Plots can also illustrate how default gives us little information apart from it is clear that those with low balances have credit in default. This should be obvious by definition. I suspect this variable will not be of much use.

1.3.6 Balance

Balances range from -3313 to 71190 (euros) which is a wide range. A box plot shows us that there are some outliers with a high balance but the interquartile ranges are narrow being 69 and 1480. The mean is 1423 close to the 3rd quartile. Simple plots are of not much use as balance figures are too widely spread and measured in thousands compared to the other continous variables.

1.3.7 Housing

Those with housing loans are perhaps less likely to have been successfuly targeted in the previous campaign. This is unsurprising as most people who are paying off a housing loan may not have the funds to subscribe to a term deposit. However, those who were contacted more previously (>4) may be more likely to have had a housing loan.

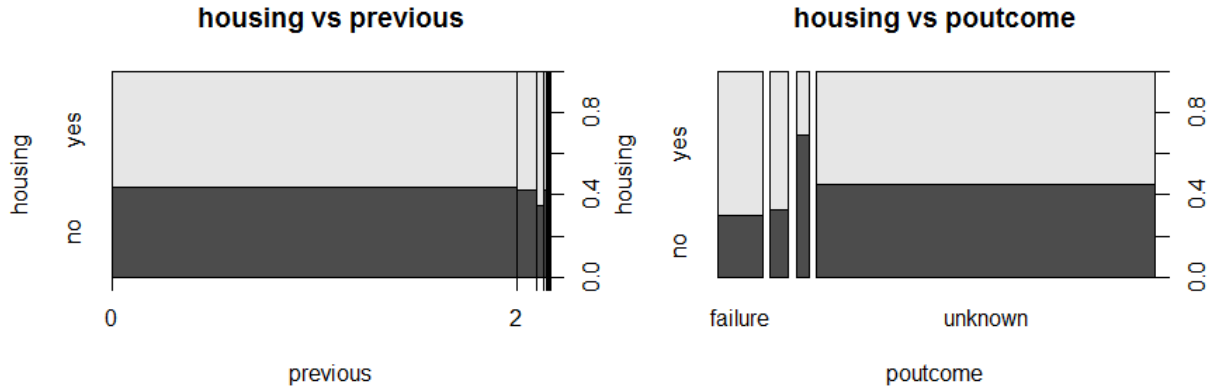

Figure 7: housing vs relevant explanatory variables.

1.3.8 Loan

Those with loans are also less likely to have been successfuly targeted in the previous campaign just like housing. Increasing the number of contacts in the campaign had a small effect on increasing the chances of them saying yes. One might ask if there's any significance in dividing loans between personal and housing?

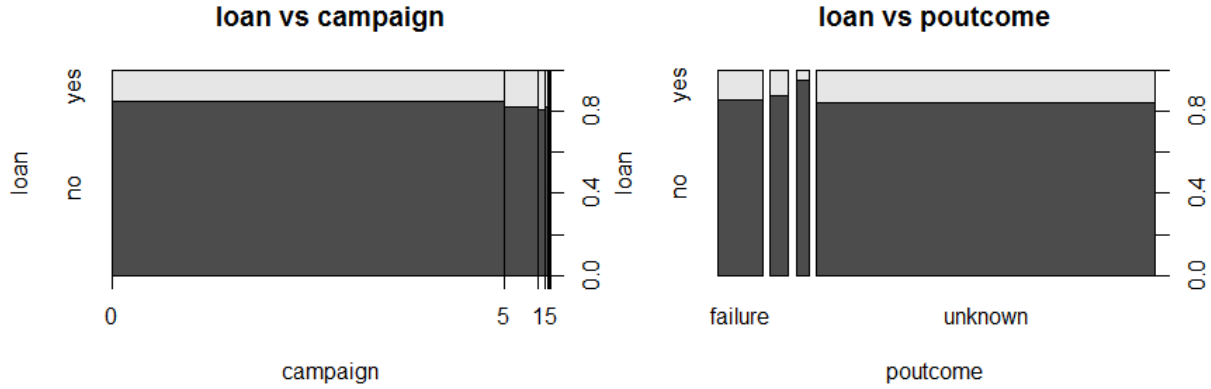

Figure 8: loan vs relevant explanatory variables.

1.3.9 Campaign Related Variables

Duration is the time in seconds of the last contact. We can see that those with succesful previous outcomes had a higher number of time spent on their last contact. However, duration by itself may not tell us much more. The information may give us polarised results, a high duration could mean that a successful contact was made or it could mean a lot of time was spent speaking to an individual who had trouble understanding resulting in failure. However, we expect in general, a low duration should imply that there is little chance of a customer saying yes.

Campaign tells us the number of contacts that the client had during this campaign including the last contact. The campaign plots don't tell us much as there is little variation. However, there may be scope for interaction between duration and campaign which we can investigate later.

The plot between previous and campaign shows us that the bank decided to target more of those in this campaign who were were not targeted at all previously. Those who were previously contacted for numerous times (over 10) were more likely to have failed in the previous campaign.

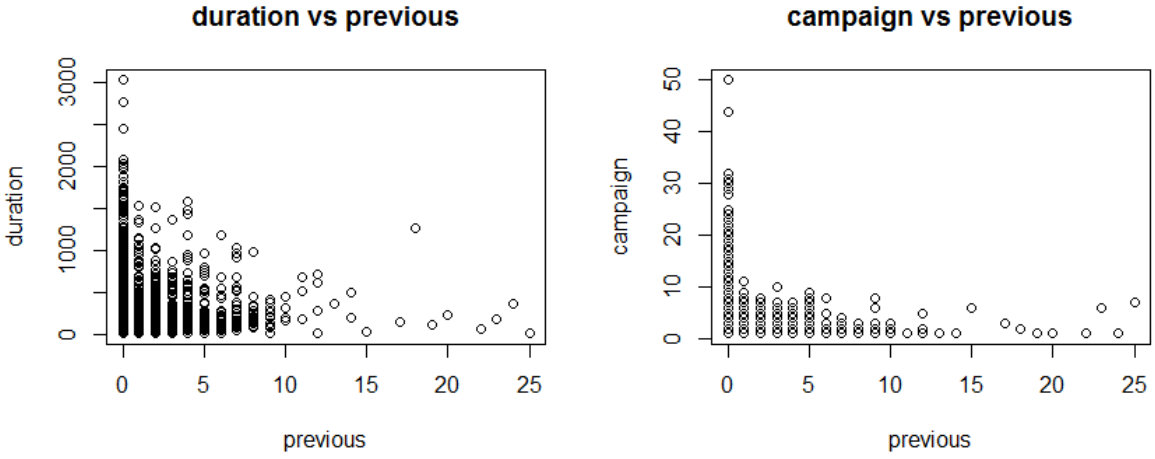

Figure 9: This shows duration and current campaign variables vs previous.

2 Univariate Analysis

Testing the association of each individual variable with the response variable is a quick and easy way to spot explanatory variables which do not show a significant pattern with y. Creating a function "logistic.regression.or.ci" can also tell us the 95% confidence intervals for all the coefficients and also for the odds ratios.

Figure 10 shows a table which summarises the confidence intervals for the odds ratio coefficients.

Age has positive (>1) odds ratio(OR) of 1.013 which means with every decade increase in age there is an increase of 12.99% of a customer saying yes. At the lower confidence boundary there may be just a 4.56% increase for every decade passed up to 21.49% at the upper boundary. So there is obviously a positive effect, which we saw in the exploratory analysis aswell.

The univariable anlaysis of job tells a lot. Clearly some jobs have a very high OR relative to admin such as student and retired which we saw earlier. Blue collar workers have a much lower OR (<1). Having checked the confidence intervals for all the ORs of the jobs, we can see that many of them contain 1. Unkown has a fairly high OR of 1.635 but it too contains 1 in the confidence interval which is very wide, because there are only 38 cases of unknown jobs. There is a strong case for combining some of these jobs into white collar, blue collar, students, retired, and unemployed together with unknown.

Married people are roughly half as likely to accept than divorced people and single is similar to divorced in terms of odds ratios.

For education, there is strong effect for those with university education compared to those without. Those with tertiary education are about 60% more likely to accept than those without.

The confidence intervals of OR for those who have defaulted is very wide and the point estimate is very close to 1. The anova results also shows below that the chi square test shows that this data may not be significant.

```
Response: y
Terms added sequentially (first to last)

        Df  Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     4520       3231
default  1 0.0076122      4519       3231   0.9305
```

We can remove the default variable from any further analysis from now on. Similarly, for balance the confidence intervals of OR lie around 1 and the anova results show that there is insignificant association between the response variable and balance so we should also remove this variable:

```
Response: y
Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    4520     3231.0
balance  1   1.3285      4519     3229.7   0.2491
```

For housing loans and personal loans, both show that those already with a loan are about half as likely to accept than those without.

The duration variable has an odds ratio tightly spread between (1.00322,1.00390) which tells us that for every additional 120 increase (2 minutes) then the chances of success increases by 47.06% which is a very large factor. However, case 569 with the highest duration figure of 3025 shows us what we suspected before. This individual is 59 years old with no job and 0 in the balance yet spent almost an hour in contact with the bank to which the end result was no.

As the number of contacts made during the campaign increases, the success rate falls slightly which tells us there is a negative association. However, could this be because those who were contacted most those were failures from a previous campaign, which would mean they would have less chance anyway?

The previous variable shows the opposite trend. There is a slight increase in OR for those who were contacted the most number of times in the previous campaign. Perhaps they were people who ended up subscribing to the old product so may be interested in the new one.

This brings us to the last univariable test for previous outcome. We can see that this is a very significant variable with a near 0 p value showing that it is almost definitely not insignificant. Indeed the OR of those who were a success in the past tells us they are at least 7.82 times more likely to accept in this campaign.

	univariate				simple multivariate		
	2.50%	OR CI	97.50%		2.50%	OR CI	97.50%
AGE	[1]	1.004559	1.021487		[1,]	0.9875529	1.0141483
JOB blue	[1,]	0.3943173	0.8231818		[2,]	0.3753869	0.9392575
ent	[2,]	0.3907365	1.2899066		[3,]	0.3125089	1.3132362
hous	[3,]	0.5544545	1.9301035		[4,]	0.3048103	1.4596479
man	[4,]	0.8133496	1.5755054		[5,]	0.5774845	1.4449736
ret	[5,]	1.4740527	3.3488197		[6,]	1.0170111	3.2543881
self	[6,]	0.5180583	1.5238730		[7,]	0.3696020	1.3950748
serv	[7,]	0.4713678	1.1183338		[8,]	0.4553284	1.2745232
stud	[8,]	1.1848776	3.7813736		[9,]	0.8582582	3.4927588
tech	[9,]	0.6141511	1.2535522		[10,]	0.4984659	1.1906968
unemp	[10,]	0.4334406	1.5459807		[11,]	0.2292904	1.1357702
unknown	[11,]	0.6885417	3.8831577		[12,]	0.5619587	4.6093100
MARITAL mar	[1,]	0.4906670	0.8447793		[13,]	0.4761460	0.9163078
sing	[2,]	0.7100079	1.2726548		[14,]	0.5811997	1.2487154
EDUCATION sec	[1,]	0.8537969	1.523345		[15,]	0.7526750	1.6115927
tert	[2,]	1.1865826	2.158374		[16,]	0.9703507	2.3341153
unknown	[3,]	0.6324757	1.861322		[17,]	0.3402746	1.2739251
DEFAULT yes	[1]	0.5114458	2.0818229		[18,]	0.7471940	3.7647518
BALANCE yes	[1]	0.9999896	1.0000433		[19,]	0.9999705	1.0000389
HOUSING yes	[1]	0.4314253	0.6244326		[20,]	0.4178609	0.6703436
LOAN yes	[1]	0.3368325	0.6428729		[21,]	0.3311734	0.7021085
DURATION	[1]	1.003219	1.003893		[22,]	1.0035938	1.0043492
CAMPAIGN	[1]	0.8679521	0.9506307		[23,]	0.8912762	0.9873519
PREVIOUS	[1]	1.107636	1.202212		[24,]	0.9223632	1.0712910
POUTCOME oth	[1]	1.041322	2.519780		[25,]	0.9628170	2.6509813
success	[2,]	7.821671	19.121221		[26,]	6.4897142	17.6900052
unknown	[3,]	0.508856	0.903848		[27,]	0.3977995	0.8795831

Figure 10: This shows the OR confidence intervals for the coefficients in univariate models and multivariate. The colour coding in the job represents potential groupings. The orange represents the variables which I am going to remove. Ignore the right table for now.

2.1 Collapsing Categorical Variables

2.1.1 Job

There are 12 levels of jobs and there is a natural way mapping them into blue collar, white collar, students, retired, other. Blue collar will contain (blue collar, housemaid). We will conflict with how the bank campaign team have allocated these individuals but since there are only 112 housemaids and they have similar plots to blue collar. We can test for confounding later. White collar will contain (admin., entrepeneur, management, self-employed, services, technician). Unemployed and unknown can be grouped together into a big unemployed group as there are only a few unknown 38/4521 cases. We can safely assume those who are unknown are quite likely to be unemployed as well, not much damage can come from combining the two together. This leaves us with retired and student which will be in their own categories. It would be unwise to group them, as there are major differences between the groups such as education and marital, not to mention age and the generation gap that comes with it.

So now we have 5 groups of jobs (white-collar, blue-collar, retired, student, unemployed) with more than half (2983) the people with white collar jobs which is a bit skewed. We can correct this later if needed.

2.1.2 Education

I also decided to turn the education category into tertiary and non tertiary as the tertiary coefficient in the univariable analysis stands above the other types of education which are similar in OR. This should give us more degrees of freedom and also make the model simpler to execute and interpret. Despite primary having a stronger effect than secondary, it still makes sense to group the two together intuitively as often the lifestyle choices (such as job and other variables) of those who left either are similar.

2.2 Confounding

Before we agree to recategorise our variables and delete both default and balance it may make sense to check for confounding. I built a simple multivariate logistic regression model from the existing data and compared the OR confidence intervals with the univariate analysis. What we are looking for are discrepencies between the figures so that we can see if an explanatory variable is affecting the relationship amongst the other variables. As you can see, there are a lot different OR's for the job variables and education in particular. So, it may make sense to recategorise these. Please refer back to Figure 10 to check the discrepencies in the the univariate model vs the multivariate model.

3 Collinearity

We now have 10 variables to work with, 2 of which have been collapsed. There may be correlation between variables which would mean that one would suffice to explain the variability in the y outcome.

3.1 Quantitative

For the quantitative variables the PPMC values between age,duration,campaign and previous are all very close to 0 showing no linear trend.

```
>cor(age,duration)  > cor(age,campaign)  > cor(age,previous)
[1] -0.002366889    [1] -0.005147905      [1] -0.003510917

> cor(duration,campaign) > cor(duration,previous) > cor(campaign,previous)
[1] -0.068382            [1] 0.01808032           [1] -0.06783263
```

3.2 Categorical

Since we have the job variable, it makes it much easier to perform relevant tests such as chi square and fishers. It may not be possible to use the fisher's exact test in R, so the chi square test may be used instead. Our null hypothesis is that the variables are independent. If we obtain a p value <0.05 we should fail to accept the null hypothesis that they are independent. After doing tests for each categorical variable vs each other here is a table of the results of the p values

	jobcat	marital	edutert	housing	loan	poutcome
jobcat						
marital	< 2.2e-16					
edutert	< 2.2e-16	2.34E-15				
housing	< 2.2e-16	0.01605	4.15E-11			
loan	0.0003859	0.004339	0.003326	0.2266		
poutcome	0.009406	0.3265	0.2726	< 2.2e-16	0.00067	

Figure 11: A table of the p values from chi square and fisher test statistics.

We can see that the poutcome variable has a few high p values, with marital and education. We should assume it is independent of these variables. The job category has produced many low p values. However depsite reducing the number of categories within job, there may still be too many. Certainly, one cannot perform a fisher's exact test with this many variables, and can assume a chi square test may not be as convincing. However, we would expect job to have many affects with the other explanatory variables. The question is which ones to remove? Job, education and marital all have very low p values so there may be inter-multi-collinearity here. Housing and loan exhibit high independence given that it was a 2x2 table we could calculate the fishers statistic accurately which produced an OR of 1.11 and we accept the null hypothesis that the OR is 1. However, they both have low p values compared to other variables so it may be too hasty to remove one at the moment.

4 Multivariate Analysis

4.1 Initial Model

Our first model is a logisitic regression with all the variables which we have kept and recategorised. For referencing purposes the recategorised variables for job and education are called "jobcat" and "edutert".

```
Call:
glm(formula = y ~ age + jobcat + marital + edutert + housing +
    loan + duration + campaign + previous + poutcome, family = "binomial",
    data = bank)
Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
-4.0411  -0.4029  -0.2881  -0.2071   2.8593

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.4707361  0.4116527  -6.002 1.95e-09 ***
age                0.0003627  0.0066549   0.054 0.956537
jobcatblue-collar -0.3327426  0.1595663  -2.085 0.037043 *
jobcatretired      0.7621806  0.2474870   3.080 0.002072 **
jobcatstudent      0.6749920  0.3185180   2.119 0.034077 *
jobcatunemployed  -0.2130691  0.3000014  -0.710 0.477563
maritalmarried    -0.4244058  0.1657491  -2.561 0.010451 *
maritalsingle     -0.1633187  0.1940289  -0.842 0.399943
eduterttertiary    0.3869159  0.1242711   3.113 0.001849 **
housingyes        -0.6375889  0.1188200  -5.366 8.05e-08 ***
loanyes           -0.7023123  0.1896260  -3.704 0.000213 ***
duration           0.0039350  0.0001901  20.704  < 2e-16 ***
campaign          -0.0636926  0.0258379  -2.465 0.013698 *
previous          -0.0045393  0.0375949  -0.121 0.903896
poutcomeother      0.4815459  0.2573530   1.871 0.061324 .
poutcomesuccess    2.3670730  0.2539172   9.322  < 2e-16 ***
poutcomeunknown   -0.5243047  0.2009093  -2.610 0.009063 **
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3231.0  on 4520  degrees of freedom
Residual deviance: 2331.9  on 4504  degrees of freedom
AIC: 2365.9
```

Both age and previous have high p values which show that they may not be contributing much significance to the outcome. We can remove these two and run another model to see if the coefficients change. If they do then there may be confounding. Throughout this model building stage we will be wary of stepwise variable selections. Maritalsingle is a variable with a high p value 0.400, but we can't consider removing this by itself, we would have to remove the whole marriage category which we won't consider at this stage.

4.1.1 Second Initial Model

We removed age and previous and ran another simple logistic regression model. We see that the coeffiecients of the remaining variables are unaffected. An anova test can show whether it was worth including age and previous.

```
> anova(log2,log1, test="Chisq")
Analysis of Deviance Table
Model 1: y ~ jobcat + marital + edutert + housing + loan + duration +
    campaign + poutcome
Model 2: y ~ age + jobcat + marital + edutert + housing + loan + duration +
    campaign + previous + poutcome
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      4506     2331.9
2      4504     2331.9  2 0.017872   0.9911
```

We can see that the residual deviance lowered by a very small amount with the full model. A p value close to 1 confirms that we should not consider age and previous any further.

4.2 Automatic Variable Selection

To give us an idea what direction to take in the variable selection procedure it can be useful to employ some algorithms which will help select models based on AIC an BIC. There is a tradeoff between number of variables and information gained, and there is no perfect amount of variables to have in a model. However, I would like to see around 5 variables at the end which should sufficiently explain the chances of individuals accepting a term deposit.

4.2.1 step(aic)

Running a "step" function on our second initial model shows no further subsets would give a better result. Perhaps we have gone too far with our second initial model? To check backwards by running the function on our first initial model, we find that we end up with our current model. So, this is as complex as our model ought to be.

4.2.2 bic.glm

Using the package "BMA" we can use the bic.glm function to test for potential models. Usually this will give us a much stricter set of variables, we can think of this as our minimum model. This function has given us 4 potential models ranked in order of highest posterior probabilities. We have no priori or reason to believe any particular variable should definitely be included so we allow the function to run itself.

```
Call: bic.glm.formula(f = y ~ jobcat + marital + edutert + housing +    loan +
duration + campaign + poutcome, data = bank, glm.family = "binomial",     maxCol = 12)
 Best  4  models (cumulative posterior probability =  1 ):

                   p!=0    EV        SD         model 1     model 2     model 3     model 4
Intercept          100    -2.767434  0.2228246  -2.873e+00  -2.758e+00  -2.658e+00  -2.541e+00
jobcat             100.0
      .blue-collar         -0.408107  0.1716981  -3.588e-01  -3.641e-01  -5.133e-01  -5.219e-01
      .retired              0.714022  0.2150960   7.624e-01   7.434e-01   6.261e-01   6.051e-01
      .student              0.746789  0.3052572   7.774e-01   7.665e-01   6.896e-01   6.783e-01
      .unemployed          -0.232858  0.3005016  -2.083e-01  -2.229e-01  -2.711e-01  -2.893e-01
marital              0.0
       .married             0.000000  0.0000000       .           .           .           .
       .single              0.000000  0.0000000       .           .           .           .
edutert             69.5
       .tertiary            0.273131  0.2083028   3.897e-01   3.992e-01       .           .
housing            100.0
       .yes                -0.648822  0.1171623  -6.394e-01  -6.405e-01  -6.686e-01  -6.710e-01
loan               100.0
    .yes                   -0.719100  0.1898993  -7.173e-01  -7.021e-01  -7.404e-01  -7.254e-01
duration           100.0    0.003937  0.0001891   3.938e-03   3.952e-03   3.919e-03   3.930e-03
campaign            34.6   -0.022421  0.0343972       .      -6.540e-02       .      -6.323e-02
poutcome           100.0
        .other              0.470721  0.2556219   4.647e-01   4.851e-01   4.616e-01   4.812e-01
        .success            2.326251  0.2530435   2.335e+00   2.332e+00   2.310e+00   2.306e+00
        .unknown           -0.551947  0.1696064  -5.653e-01  -5.120e-01  -5.814e-01  -5.301e-01
nVar                                              6           7           5           6
BIC                                            -3.560e+04  -3.560e+04  -3.560e+04  -3.560e+04
post prob                                       0.443       0.251       0.211       0.094
```

A check of the mle's and se's also show that they are very similar so there is no confounding evident. We can see that we should definitely include 5 variables (jobcat, housing, loan, duration, poutcome) at the minimum unless we regroup the data further. There is scope for including edutert and perhaps campaign too.

4.3 Selection Criteria

The nested models have provided us a useful method of testing the importance of variables. By checking the wald chi square test values we have removed age and previous succesfully. Whilst the wald chi square test is not as powerful as the likeliood ratio tests, the results of the age and previous coefficient p values were still extreme which warranted their removal. The lower the value of AIC or BIC the better the model.The algorithms which we used have given us a good set of variables producing low criterion figures.

5 Model Building

We have a base model obtained from the bic.glm algorithm containing variables (jobcat, edutert, housing, loan, duration, poutcome). This has 6 variables, close to what we would like in our final model. However we can still think about transforming these variables and start to think about interactions. Let us test the fit of the model first before we refine it.

5.1 Model Fit

For these tests I have turned the y response variable into a numeric binary variable of 1 and 0 instead of yes and no.

5.1.1 Hosmer Lemeshow Test

We can sort the data by their fitted risk value (whether they say yes) and then test the goodness of fit for each decile or subgroup. Because we will divide the 4521 into 10 groups of 452, we will miss the last value. However, on the whole it should not affect the results too much. A more elegant method would be to sort into 11 groups of 411.

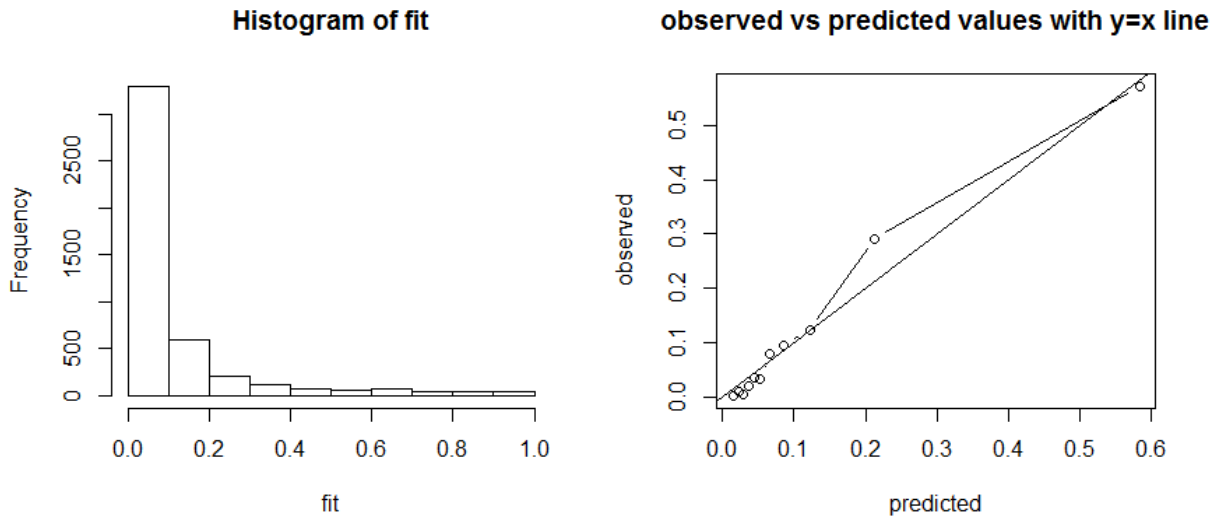

Figure 12: The frequency of predicted values and compared to observed.

We shall test with the 11 groups of 411. The plot shows that with the 45% line fitted, we have a good fit, with only the second to last group being a bit off the trend line. Most of the groups contain predicted values less than 10% so it could be quite prone to fluctuations.

If we choose a cutoff of 50% so that we can see those which are predicted correctly and observed correctly we see whether our cutoff of 50% is an appropriate value.

Looking at our data we can see that 521 values were yes of which 157 were correctly observed (fit >0.5). 4000 were no, of which 3915 were correctly predicted (fit <0.5). A sensitivity of 157/521= 30.13% and specificity of 3915/4000=97.88% show us that the choice of cutoff was very useful for predicting those who won't accept a term deposit. This makes intuitive sense, it is much easier to see those who are hard to sell than those who are not.

	observed yes	observed no	total
predicted yes	157	85	242
predicted no	364	3915	4279
total	521	4000	4521

Figure 13: Table of observed and predicted binary outcomes for the repsonse variable.

5.1.2 ROC

We can now try to calculate a receiver operating characteristic ROC curve. This can tell us about the performance of our model as the cutoff point is varied. It is used in signal detection thoery and can be applied to this case as we are detecting whether we have correctly predicted yes or no. We can calculate the area under the curve AUC which is equal to the probability that the model will rank a randomly chosen yes individual higher than a randomly chosen one who says no.

We can do a plot of the cutoff values ranging from 0 to 100% in 10% increments. If we plot the sensitivity and 1-specificity we can see the concavity. The more concave, the better our model is at predicting as it will have 100% sensitivity and 100% specificity.

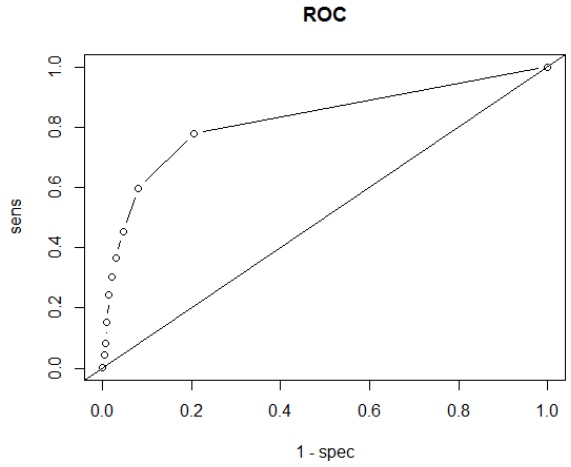

Figure 14: Sensitivity against 1-Specificity with a y=x line

Calculating the AUC vs the line through the origin at y=x will give us the Wilcoxon Mann Whitney test test statistic, which is equivalent to probability of successfully selecting the yes from a pair of yes or no.

```
> wilcox.test(x=fit.pos, y=fit.neg)

Wilcoxon rank sum test with continuity correction

data:  fit.pos and fit.neg
W = 1817999, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

If we find the number of possible pairs of yes and no (521*4000) and divide the W statistic by this combination we shall obtain the AUC = 87.24% which is a good result. We are very likely to predict the yes result with this model given any pair of contrasting individuals.

5.2 Interactions

A possible interaction we could test for is between housing and loan. Having a housing loan and a personal loan may be rare and tell us something more about the individual's finance appetite. If they have both loans or neither we may expect there to be a multiplicative effect on the odds ratio. A logistic regression model including this interaction gives the Wald statistic of 0.312 which shows that it perhaps is insignificant. A likelihood ratio test gives the same result. An interaction term for these two variables should not be included. The AIC actually increases marginally in the more complex model.

```
Model 1: y ~ jobcat + edutert + housing + loan + duration + poutcome
Model 2: y ~ jobcat + edutert + housing + loan + housing:loan + duration +
    poutcome
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      4509     2347.5
2      4508     2346.5  1   1.0322   0.3096
```

Another possible interaction is between poutcome and duration. We may expect that those who did take up a product previously and those whose length of last contact was low to be unusually extremely unlikely to accept the term deposit. A likelihood ratio test shows us that there is a likely to be case for this inclusion.

```
Model 1: y ~ jobcat + edutert + housing + loan + duration + poutcome
Model 2: y ~ jobcat + edutert + housing + loan + duration + poutcome +
    duration:poutcome
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      4509     2347.5
2      4506     2337.8  3   9.7666  0.02066 *
```

However, I think an interaction term may be too complex for this model. We want the model to be easily interprested and this may cause greater difficulty. The p value is low, but the wald test statistics with this interaction show that the wald values are not very low. It would add 3 more variables in our final model. For parsimony sake I will leave the interactions out.

6 Final Model

For logistic regression models, the model equation is of the form:

$$log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

where $\theta(x)$ is the probability of a yes outcome in this campaign. The β's represent the coefficients of the the intercept and the variables and the x's represent the values for each variable.

A summary of the final model coefficients:

```
Call:
glm(formula = y ~ jobcat + edutert + housing + loan + duration +
    poutcome, family = "binomial", data = bank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0662  -0.3997  -0.2912  -0.2130   2.8550

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.8730163  0.1954987 -14.696  < 2e-16 ***
jobcatblue-collar -0.3588007  0.1586950  -2.261 0.023763 *
jobcatretired      0.7623791  0.2068638   3.685 0.000228 ***
jobcatstudent      0.7773580  0.3038400   2.558 0.010514 *
jobcatunemployed  -0.2082711  0.2990741  -0.696 0.486187
eduterttertiary    0.3896971  0.1234569   3.157 0.001597 **
housingyes        -0.6393886  0.1164952  -5.489 4.05e-08 ***
loanyes           -0.7172690  0.1899533  -3.776 0.000159 ***
duration           0.0039383  0.0001887  20.873  < 2e-16 ***
poutcomeother      0.4647079  0.2553299   1.820 0.068755 .
poutcomesuccess    2.3349551  0.2529518   9.231  < 2e-16 ***
poutcomeunknown   -0.5653074  0.1671912  -3.381 0.000722 ***

    Null deviance: 3231.0  on 4520  degrees of freedom
Residual deviance: 2347.5  on 4509  degrees of freedom
AIC: 2371.5


$OR
jobcatblue-collar     jobcatretired     jobcatstudent
        0.6985135         2.1433694         2.1757165
jobcatunemployed   eduterttertiary        housingyes
        0.8119869         1.4765334         0.5276149
          loanyes          duration     poutcomeother
        0.4880834         1.0039460         1.5915492
poutcomesuccess   poutcomeunknown
      10.3289960         0.5681854
```

```
$OR.ci
            [,1]       [,2]
 [1,] 0.5117919  0.9533585
 [2,] 1.4289407  3.2149915
 [3,] 1.1994253  3.9466753
 [4,] 0.4518317  1.4592220
 [5,] 1.1591953  1.8807451
 [6,] 0.4199099  0.6629457
 [7,] 0.3363609  0.7082434
 [8,] 1.0035748  1.0043174
 [9,] 0.9649013  2.6251688
[10,] 6.2913685 16.9578620
[11,] 0.4094272  0.7885033
```

So by taking the inverse of the equation of the general model we can input the point estimates of the OR's to produce our final model:

$$\begin{aligned}\left(\frac{P(Yes)}{P(No)}\right) = \mathbf{0.0565} \\ &* \mathbf{0.699}(bluecollar) * \mathbf{2.143}(retired) * \mathbf{2.176}(student) * \mathbf{0.812}(unemployed) \\ &* \mathbf{1.477}(tertiaryeducation) \\ &* \mathbf{0.528}(housingloan) \\ &* \mathbf{0.488}(personalloan) \\ &* \mathbf{1.004}\hat{}(durationoflastcontactinseconds) \\ &* \mathbf{1.592}(poutcomeother) * \mathbf{10.329}(poutcomesuccess) * \mathbf{0.568}(poutcomeunknown)\end{aligned}$$

where our base individual is a white collar, non tertiary educated person with no loans and who did not subscribe to a product in the previous campaign and who was not contacted at all in this current campaign. He/she would have an odds ratio of 0.0565, the intercept i.e. there is a 5.35% chance they will accept.

Contrastingly, imagine a retired person with tertiary education and no loans, who has been contacted for 2 minutes in their last contact and have also signed up to a product in the previous campaign. Then, he/she would have an odds ratio of 2.963 and a probability of saying yes in this campaign of 74.77% which is almost 3/4 likely!

7 Conclusion

With an AUC of at least 87% we can safely say that our model is quite accurate in its predictions.

Throughout this model building process we have tried to look both forwards and backwards in each step so that we do not miss out or include incorrectly a variable in our final model. Confounding was checked thoroughly as possible and as the final variables are independent to each other (apart from poutcome but this can't be excluded as it is far too important) it shouldn't be a problem. However, we could have been more rigorous to get a more precise model. Yet, despite the direct approach in this project we have

produced a final model which is fairly strong.

The inclusion of edutert is questionable. It of course helps to have this extra variable but makes our model a bit more complex. However, since I recategorised it into tertiary and non tertiary this has given it more reason to fit into our model. Despite still having 6 different variables, we have managed to halve the amount of variables from the original set which is a vast improvement. We also know that the minimum set of variables should be around 5 from the bic.glm and step functions and we have included these 5. Since this model is used for direct marketing, which is tailored for the customer it would make sense to include a good amount of variables. Certainly, too few variables will be far more dangerous for the bank advertisers than having a slight more cumbersome model. We have made this model easier to interpret with the job categories recollapsed and have not included many continuous variables. Only duration remains and we could even think about turning this into a group in future.

One of the biggest criticisms for the model may be the data from the very beginning. This set of 4521 is only a tenth of the full set of data from the UCI repository. A lot of variables were not even there for us to consider. However, since our data was randomly selected from the full dataset this rules out any bias selection which is encouraging.

Our model is also based on the assumption that continous variables (duration) are linear. Without this assumption our model would break down so in order to test for this we could use the Box Tidwell method in future.

8 References

archive.ics.uci.edu/ml/datasets/Bank+Marketing (Insight into full original data)

www.medicine.mcgill.ca/epidemiology/joseph - Lawrence Joseph (Adapted output function with OR and CI for glm, ROC curve method).

sydney.edu.au/vetscience/biostat/macros/logistictutmodel (Project Structure)

9 Code

```
setwd("~/Warwick/2012-2013 morse3/st404/ass4")
bank <- read.table("A4bank.txt", header=T, quote="\"")
bank
attach(bank)
summary(bank)

#exploratory plots

par(mfrow=c(3,4))
plot(y~age, main="y vs age")
plot(y~job, main="y vs job")
plot(y~marital, main="y vs marital")
plot(y~education, main="y vs education")
plot(y~default, main="y vs default")
plot(y~balance, main="y vs balance")
plot(y~housing, main="y vs housing")
plot(y~loan, main="y vs loan")
plot(y~duration, main="y vs duration")
plot(y~campaign, main="y vs campaign")
plot(y~previous, main="y vs previous")
plot(y~poutcome, main="y vs poutcome")

par(mfrow=c(1,1))
plot(y~age, main="y vs age", xlim=c(.95,1))

par(mfrow=c(4,3))
plot(age~job, main="age vs job")
plot(age~marital, main="age vs marital")
plot(age~education, main="age vs education")
plot(age~default, main="age vs default")
plot(age~balance, main="age vs balance")
plot(age~housing, main="age vs housing")
plot(age~loan, main="age vs loan")
plot(age~duration, main="age vs duration")
plot(age~campaign, main="age vs campaign")
plot(age~previous, main="age vs previous")
plot(age~poutcome, main="age vs poutcome")

par(mfrow=c(4,3))
plot(marital~job, main=" marital vs job")
plot(education~job, main=" education vs job")
plot(default~job, main=" default vs job")
plot(balance~job, main=" balance vs job")
plot(housing~job, main=" housing vs job")
plot(loan~job, main=" loan vs job")
plot(duration~job, main=" duration vs job")
plot(campaign~job, main=" campaign vs job")
plot(previous~job, main=" previous vs job")
plot(poutcome~job, main=" poutcome vs job")

par(mfrow=c(3,3))
plot(marital~education, main=" marital vs education")
plot(marital~default)
plot(marital~balance)
plot(marital~housing)##--
```

```
plot(marital~loan)
plot(marital~duration)
plot(marital~campaign)
plot(marital~previous)
plot(marital~poutcome)

par(mfrow=c(1,1))
plot(marital~education, main=" marital vs education")

par(mfrow=c(4,2))
plot(education~default)
plot(education~balance)
plot(education~housing)
plot(education~loan)
plot(education~duration)
plot(campaign~education)
plot(previous~education)
plot(education~poutcome)

par(mfrow=c(2,2))
plot(education~default, main="education vs default")
plot(education~balance, main="education vs balance")
plot(education~housing, main="education vs housing")
plot(education~loan, main="education vs loan")

par(mfrow=c(4,2))
plot(default~balance)
plot(default~housing)
plot(loan~default)
plot(default~duration)
plot(campaign~default)
plot(previous~default)
plot(poutcome~default)

summary(default)
table(y,default)


par(mfrow=c(3,2))
plot(balance~housing)
plot(balance~loan)
plot(balance~duration)
plot(balance~campaign)
plot(balance~previous)
plot(balance~poutcome)
summary(balance)
par(mfrow=c(1,1))
boxplot(balance)

par(mfrow=c(3,2))
plot(housing~loan)
plot(housing~duration)
plot(campaign~housing)
plot(housing~previous)
plot(housing~poutcome)
par(mfrow=c(1,2))
plot(housing~previous,main="housing vs previous")
plot(housing~poutcome, main="housing vs poutcome")
```

```
par(mfrow=c(2,2))
plot(loan~duration)
plot(loan~campaign)
plot(loan~previous)
plot(loan~poutcome)
par(mfrow=c(1,2))
plot(loan~campaign, main="loan vs campaign")
plot(loan~poutcome, main="loan vs poutcome")

par(mfrow=c(2,2))
plot(duration~campaign)

plot(duration~poutcome)


par(mfrow=c(2,1))
plot(campaign~previous)
plot(campaign~poutcome)

plot(previous~poutcome)


par(mfrow=c(1,2))
plot(duration~previous, main="duration vs previous")
plot(campaign~previous, main="campaign vs previous")




logistic.regression.or.ci <- function(regress.out, level=0.95)
{
  ##############################################
  # #
  # This function takes the output from a glm #
  # (logistic model) command in R and provides not #
  # only the usual output from the summary command, but #
  # adds confidence intervals for all coefficients and OR's. #
  # #
  # This version accommodates multiple regression parameters #
  # #
  ##############################################
  usual.output <- summary(regress.out)
  z.quantile <- qnorm(1-(1-level)/2)
  number.vars <- length(regress.out$coefficients)
  OR <- exp(regress.out$coefficients[-1])
  temp.store.result <- matrix(rep(NA, number.vars*2), nrow=number.vars)
  for(i in 1:number.vars)
  {
    temp.store.result[i,] <- summary(regress.out)$coefficients[i] +
      c(-1, 1) * z.quantile * summary(regress.out)$coefficients[i+number.vars]
  }
  intercept.ci <- temp.store.result[1,]
  slopes.ci <- temp.store.result[-1,]
  13
  OR.ci <- exp(slopes.ci)
  output <- list(regression.table = usual.output, intercept.ci = intercept.ci,
```

```
                slopes.ci = slopes.ci, OR=OR, OR.ci = OR.ci)
  return(output)
}

#univariate analysis

aglogit <- glm(y ~ (age), data= bank, family = "binomial")
logistic.regression.or.ci(aglogit)
plot(aglogit)
anova(aglogit, test="Chisq")


jologit <- glm(y ~ job, data= bank, family = "binomial")
logistic.regression.or.ci(jologit)
anova(jologit, test="Chisq")
summary(job)
plot(jologit)

malogit <- glm(y ~ marital, data= bank, family = "binomial")
logistic.regression.or.ci(malogit)
anova(malogit, test="Chisq")


edlogit<-glm(y ~ education, data= bank, family = "binomial")
logistic.regression.or.ci(edlogit)
anova(edlogit, test="Chisq")

delogit<- glm(y ~ default, data= bank, family = "binomial")
logistic.regression.or.ci(delogit)
anova(delogit, test="Chisq")


balogit<- glm(y ~ balance, data= bank, family = "binomial")
logistic.regression.or.ci(balogit)
anova(balogit, test="Chisq")


hologit<- glm(y ~ housing, data= bank, family = "binomial")
logistic.regression.or.ci(hologit)
anova(hologit, test="Chisq")


lologit<- glm(y ~ loan, data= bank, family = "binomial")
logistic.regression.or.ci(lologit)
anova(lologit, test="Chisq")



dulogit<- glm(y ~ duration, data= bank, family = "binomial")
logistic.regression.or.ci(dulogit)
anova(dulogit, test="Chisq")
summary(duration)
plot(dulogit)

calogit<- glm(y ~ campaign, data= bank, family = "binomial")
logistic.regression.or.ci(calogit)
anova(calogit, test="Chisq")
```

```
prlogit<- glm(y ~ previous, data= bank, family = "binomial")
logistic.regression.or.ci(prlogit)
anova(prlogit, test="Chisq")

pologit<- glm(y ~ poutcome, data= bank, family = "binomial")
logistic.regression.or.ci(pologit)
anova(pologit, test="Chisq")
plot(pologit)


#collapsing categories

education
levels(education) <- c("other","other","tertiary","other")
edutert<-education
bank$edutert<- edutert
plot(education)

summary(job)
job
levels(job) <- c("white-collar","blue-collar","white-collar","blue-collar",
"white-collar","retired","white-collar","white-collar",
"student","white-collar", "unemployed","unemployed")
jobcat<-job
bank$jobcat<- jobcat
summary(job)
summary(bank$job)
summary(bank$jobcat)

#simple multi
logit<- glm(y ~ age + job + marital + education + default +balance
+ housing+ loan + duration + campaign + previous + poutcome,
data= bank, family = "binomial")
logistic.regression.or.ci(logit)



#correlation

cor(age,duration)
cor(age,campaign)
cor(age,previous)
cor(duration,campaign)
cor(duration,previous)
cor(campaign,previous)

#collinearity tests

table1<- table(job, marital)
table1
chisq.test(table1)
fisher.test(table1)

table2<- table(job, education)
chisq.test(table2)
fisher.test(table2)
kappa(table2)
```

```
table3<- table(job, housing)
chisq.test(table3)
fisher.test(table3)

table4<- table(job, loan)
chisq.test(table4)
fisher.test(table4)

table5<- table(job, poutcome)
chisq.test(table5)
fisher.test(table5)


table2<- table(marital, education)
table2
chisq.test(table2)
fisher.test(table2)
kappa(table2)

table3<- table(marital, housing)
chisq.test(table3)
fisher.test(table3)

table4<- table(marital, loan)
chisq.test(table4)
fisher.test(table4)

table5<- table(marital, poutcome)
table5
chisq.test(table5)
fisher.test(table5)


table3<- table(education, housing)
table3
chisq.test(table3)
fisher.test(table3)

table4<- table(education, loan)
chisq.test(table4)
fisher.test(table4)

table5<- table(education, poutcome)
table5
chisq.test(table5)
fisher.test(table5)

table4<- table(housing, loan)
chisq.test(table4)
fisher.test(table4)
kappa(table4)

table5<- table(housing, poutcome)
table5
chisq.test(table5)
fisher.test(table5)

table5<- table(loan, poutcome)
```

```
table5
chisq.test(table5)
fisher.test(table5)

#initial models

log1<- glm(formula = y ~ age + jobcat + marital + edutert + housing
 + loan + duration + campaign + previous+ poutcome,
 family = "binomial", data = bank)
summary(log1)
plot(log1)

log2<- glm(formula = y ~ jobcat + marital + edutert + housing + loan
 + duration + campaign + poutcome, family = "binomial", data = bank)
summary(log2)
plot(log2)

anova(log2,log1, test="Chisq")

#autoselection
steplog2 <- step(log2)
summary(steplog2)

#package required (BMA)
biclog <- bic.glm(y ~ jobcat + marital + edutert + housing + loan
+ duration + campaign + poutcome, glm.family = "binomial",
                  data=bank, maxCol = 12)
summary(biclog)
biclog$label
biclog$probne0 #prob of being included
biclog$mle #check confound
biclog$se#^


#base model and #final model
log3<- glm(formula = y ~ jobcat + edutert + housing + loan
+ duration + poutcome, family = "binomial", data = bank)
logistic.regression.or.ci(log3)
summary(log3)
plot(log3)

#interactions
log4<- glm(formula = y ~ jobcat + edutert + housing + loan
+ housing:loan + duration + poutcome, family = "binomial", data = bank)
summary(log4)
plot(log3)
anova(log3,log4, test="Chisq")


log5<- glm(formula = y ~ jobcat + edutert + housing + loan
+ duration + poutcome + duration:poutcome, family = "binomial", data = bank)
summary(log5)
plot(log5)
anova(log3,log5, test="Chisq")


#model fit
ybin<- ifelse(bank$y == "yes", 1, 0)
```

```
ybin

#chisq
fit <- log3$fitted
hist(fit)

r <- (ybin - fit)/(sqrt(fit*(1-fit)))
r
sum(r^2)

1- pchisq(7660.882, df=4509)

#hosmer lem
index <- sort.list(fit)
hosmer <- matrix(c(ybin[index], fit[index]), byrow=F, nrow=4521)
hosmer
hosmer[1:1000,1:2]
hosmer[1001:2000,1:2]
hosmer[2001:3000,1:2]
hosmer[3001:4000,1:2]
hosmer[4001:4521,1:2]


observed <- rep(NA, 11)
for (i in 1:11) {observed[i] <- sum(hosmer[(411*(i-1)+1):(411*i),1])/411}

predicted <- rep(NA, 11)
for (i in 1:11) {predicted[i] <- sum(hosmer[(411*(i-1)+1):(411*i),2])/411}

plot(predicted, observed, type="b", main="observed vs predicted values with y=x line")
abline(a=0, b=1)

up<- hosmer[-(1:4279),]
up
colSums(up)

down<-hosmer[(1:4279),]
down
colSums(down)

#roc
sens <- rep(NA, 11)
spec <- rep(NA, 11)

sens[1] <- 1
spec[1] <- 0

# Cutoff of 10% for positivity (occurs at index 3298)
sens[2] = sum(hosmer[3298:4521,1])/521
spec[2] = sum(1-hosmer[1:3297,1])/4000

# Cutoff of 20% for positivity (occurs at index 3889)
sens[3] = sum(hosmer[3889:4521,1])/521
spec[3] = sum(1-hosmer[1:3888,1])/4000

# Cutoff of 30% for positivity (occurs at index 4101)
sens[4] = sum(hosmer[4101:4521,1])/521
spec[4] = sum(1-hosmer[1:4100,1])/4000
```

```
# Cutoff of 40% for positivity (occurs at index 4210)
sens[5] = sum(hosmer[4210:4521,1])/521
spec[5] = sum(1-hosmer[1:4209,1])/4000

# Cutoff of 50% for positivity (occurs at index 4279)
sens[6] = sum(hosmer[4279:4521,1])/521
spec[6] = sum(1-hosmer[1:4278,1])/4000

# Cutoff of 60% for positivity (occurs at index 4334)
sens[7] = sum(hosmer[4334:4521,1])/521
spec[7] = sum(1-hosmer[1:4333,1])/4000

# Cutoff of 70% for positivity (occurs at index 4404)
sens[8] = sum(hosmer[4404:4521,1])/521
spec[8] = sum(1-hosmer[1:4403,1])/4000

# Cutoff of 80% for positivity (occurs at index 4448)
sens[9] = sum(hosmer[4448:4521,1])/521
spec[9] = sum(1-hosmer[1:4447,1])/4000

# Cutoff of 90% for positivity (occurs at index 4483)
sens[10] = sum(hosmer[4483:4521,1])/521
spec[10] = sum(1-hosmer[1:4482,1])/4000

# Cutoff of 100% for positivity (occurs at index 4520)
sens[11] = sum(hosmer[4520:4521,1])/521
spec[11] = sum(1-hosmer[1:4520,1])/4000

sens
spec

par(mfrow=c(1,1))
plot(1-spec, sens, type="b", main="ROC")
abline(a=0, b=1)

#auc
fit.pos <- fit[ybin==1]
fit.pos
fit.neg <- fit[ybin==0]
wilcox.test(x=fit.pos, y=fit.neg)
```