Problem statement and data used:

The dataset I have used is bank marketing dataset. Here is a summary to present the basic data.

Observations	Variables
Orginal:11162	Original:17
After cleaning:9869	After cleaning:4

My research question is what conditions of people might borrow housing loan. In other words, I wanted to explore the relation between housing loan and different characters of people. Thus, the variables selected are housing[1], martial[2], and age[3]. It makes sense that how martial status and age could impact the decision of buying houses.

Regarding data cleaning, I first did a summary of the original dataset. Then I used subset function to make the variable marital only include "married" and "single" status, and I subset the whole dataset to create a new one only including martial, age and housing.

Planning:

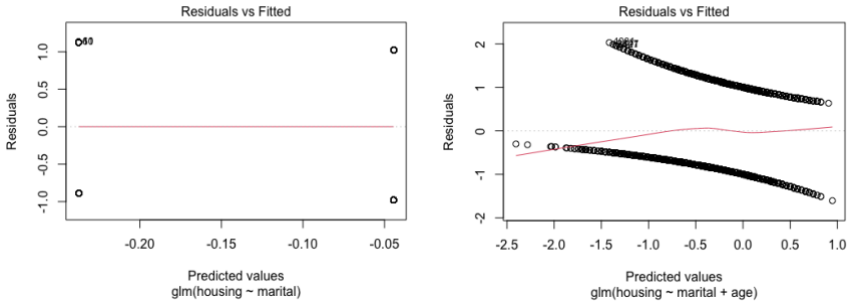
Housing is my dependent variable. My predictors are martial and age. I wanted to explore how marital, and age impact the possession of housing loan. In terms of doing logistic regression, I planned to compare two models and check the effect of each of them on housing. The first one was only involved with one predictor martial, and the second model included martial and age. I also planned to use jitter to plot the three variables, and checked assumptions of multi-collinearity, linearity, and independence of errors. For multi-collinearity, I used VIF function in R. For checking linearity, I calculated logit values and checked p-values of my continuous variable. I also plotted the graph of residuals versus fitted values to check independence of errors.

Analysis:

Firstly, I checked the assumptions of my logistic regression models. The VIFs of the two variables are around 1.3 which are fine. For linearity, the logit value of the variable age is -0.24351 with a p-value lower than 0.05. For independence of errors, when the predictor is

[1] Binary variable. This represents housing loan. Does this person have housing loan?
[2] Binary variable. Martial Status: Single or Married
[3] Continuous variable

marital only, it is a straight line, indicating it is fine; however, when adding age, there is slightly deviance from the straight line, but not much.



For the regression of predictor martial[4], when the martial status is single, the coefficient is -0.1932 with p-value lower than 0.05, meaning that single people have less motivation to borrow housing loan. When adding the second variable age, the coefficient of age is -0.040733 with p-value lower than 0.05, indicating that younger people also tended not to borrow housing loan. In addition, by checking the confidence intervals of the two variables, both the values within 2.5% and 97.5% do not cross 1, meaning that when predictors increase, the odds of having housing loan will decrease.[5] Comparing the deviance of the two models, the null deviance of the two models is 13650 which is the same; however, the residual deviance of the second model is 13232 which is less than the first one which is 13629. Besides, the value of AIC of the second model is also less than the first one.[6] Above all, we can conclude that model 2 with two predictors is more accurate and better than model 1 included marital only.

Conclusion

Based on my research question and analysis of the data, I can conclude that it seems like younger and single people tend to have less housing loan, meaning that younger and single people do not intend to buy houses. It makes sense that married people tend to buy a house to settle down and borrow housing loan. In addition, when age increases, people are more likely to get married and buy a house.

[4] See index for the model.
[5] See index for the data.
[6] 13238<13633

Index

glm(formula = housing ~ marital, family = binomial(), data = bank_data2)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.159	-1.159	-1.078	1.196	1.280

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.04441	0.02510	-1.769	0.0769	.
maritalsingle	-0.19320	0.04223	-4.575	4.76e-06	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 13650 on 9868 degrees of freedom

Residual deviance: 13629 on 9867 degrees of freedom

AIC: 13633

Call:

glm(formula = housing ~ marital + age, family = binomial(), data = bank_data2)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.597	-1.126	-0.775	1.173	1.807

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.761568	0.097130	18.14	<2e-16	***
maritalsingle	-0.662644	0.049363	-13.42	<2e-16	***
age	-0.040733	0.002127	-19.15	<2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 13650 on 9868 degrees of freedom

Residual deviance: 13232 on 9866 degrees of freedom

AIC: 13238

	2.5 %	97.5 %
(Intercept)	4.8159431	7.0477156
maritalsingle	0.4678559	0.5677438
age	0.9560720	0.9640761