Econ 104 Project 3

Group 12

2024-01-02

Contents

2. Binary Dependent Variables

(a) Briefly discuss your data and the question you are trying to answer with your model.

Citation

```
## To cite AER, please use:
##
##   Christian Kleiber and Achim Zeileis (2008). Applied Econometrics with
##   R. New York: Springer-Verlag. ISBN 978-0-387-77316-2. URL
##   https://CRAN.R-project.org/package=AER
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     title = {Applied Econometrics with {R}},
##     author = {Christian Kleiber and Achim Zeileis},
##     year = {2008},
##     publisher = {Springer-Verlag},
##     address = {New York},
##     note = {{ISBN} 978-0-387-77316-2},
##     url = {https://CRAN.R-project.org/package=AER},
##   }
```

The dataset chosen for this part of the project is `SwissLabor`, which contains 872 observations on 7 variables. `SwissLabor` is a binary dependent variable dataset that consists of one binary dependent variable and six independent variables, including:

Dependent Varaible

- `participation (binary variable)`: Factor. Did the individual participate in the labor force?

Independent Varaible

- `income`(continous variable): Logarithm of nonlabor income.
- `age`(discrete variable): Age in decades (years divided by 10).
- `education` (discrete variable): Years of formal education.
- `youngkids` (discrete variable): Number of young children (under 7 years of age).
- `oldkids` (discrete variable): Number of older children (over 7 years of age).
- `foreign` (indicator variable): Factor. Is the individual a foreigner (i.e., not Swiss)?

As shown above, there are 1 continuous, 4 discrete, and 1 indicator variable and the dependent variable is a binary variable.

The objective of this portion of the project is to identify how the different values of independent variables influence the probability of getting a certain outcome. We will be attempting to build a predictive model that predicts the outcome of the binary dependent variable (`participation`; yes or no) by investigating the relationship between the dependent and independent variables and estimating the marginal effect of each variable.

Furthermore, with the model built within the project, we will attempt to predict the outcome, and how likely a person would participate in the workforce in a hypothetical situation. (e.g. an x-year-old person with a log-non-labor income of y and z years of education, etc. . .)

The detailed descriptions of the variables in the dataset will be further discussed in part(b).

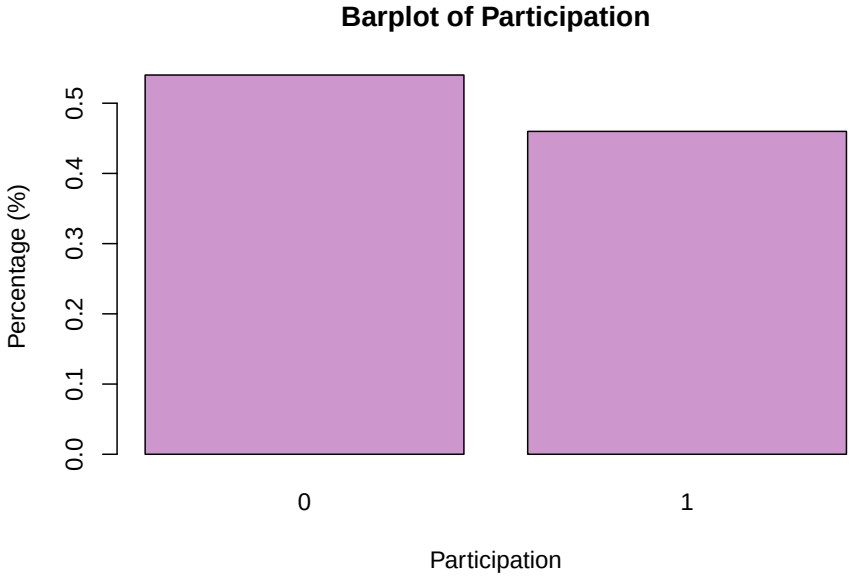
(b) Provide a descriptive analysis of your variables. This should include RELEVANT histograms and fitted distributions, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five- number summary). All figures must include comments. For binary variables, you can simply include the proportions of each factor.

Statistical Summaries: Before getting into the detailed descriptive analysis of each variable, the five number statistical summaries below provide a general idea of what variables we are dealing with.

participation	income	age	education	youngkids	oldkids	foreign
Min. :0.0000	Min. : 7.187	Min. :2.000	Min. : 1.000	Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:10.472	1st Qu.:3.200	1st Qu.: 8.000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median :0.0000	Median :10.643	Median :3.900	Median : 9.000	Median :0.0000	Median :1.0000	Median :0.0000
Mean :0.4599	Mean :10.686	Mean :3.996	Mean : 9.307	Mean :0.3119	Mean :0.9828	Mean :0.2477
3rd Qu.:1.0000	3rd Qu.:10.887	3rd Qu.:4.800	3rd Qu.:12.000	3rd Qu.:0.0000	3rd Qu.:2.0000	3rd Qu.:0.0000
Max. :1.0000	Max. :12.376	Max. :6.200	Max. :21.000	Max. :3.0000	Max. :6.0000	Max. :1.0000

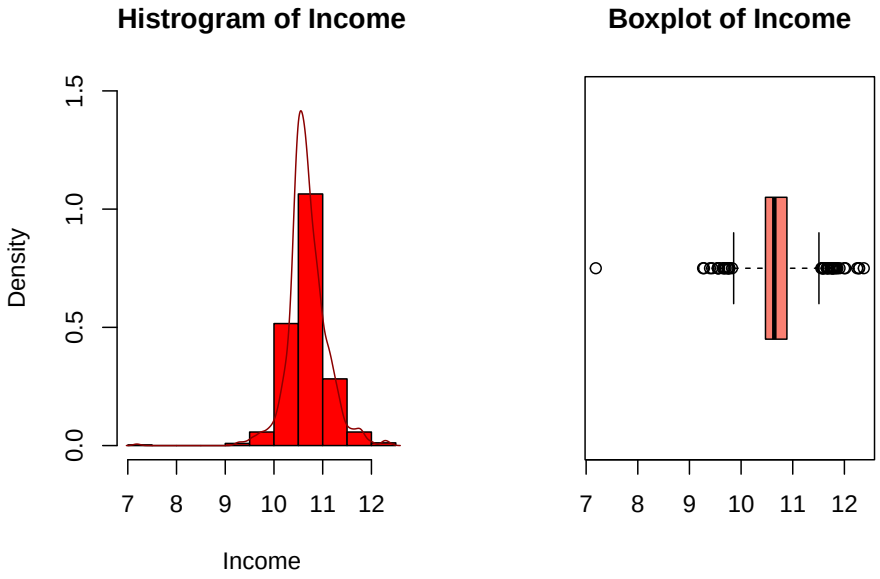
Binary Dependent Variable

`participation:` People who were participated in the labor force is 45.99% (401 observation) and the rest of the people, approximately 54% of the people (471 observation), did not participate in the workforce.

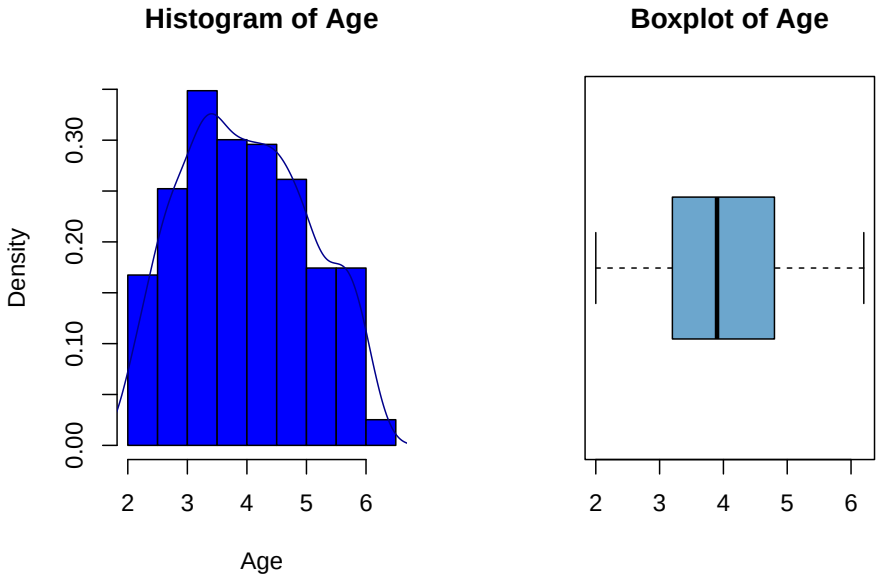


Independent Variables

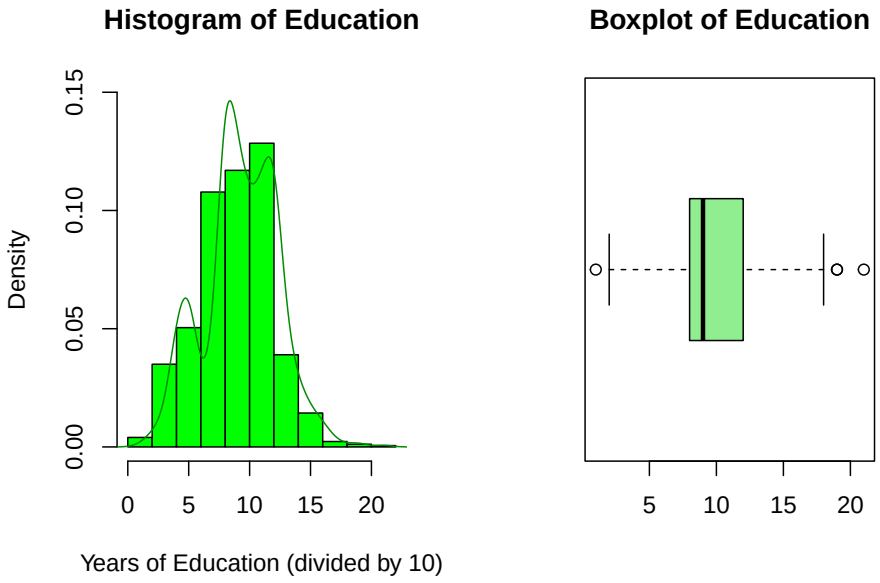
`income :` Logarithm of nonlabor income. The `income` variable is slightly left-skewed, and as seen in the 5-number statistical summaries and the box plot, more than 75% of the observations have log-nonlabor income of 10-11, which is about $22,000 - $60,000.



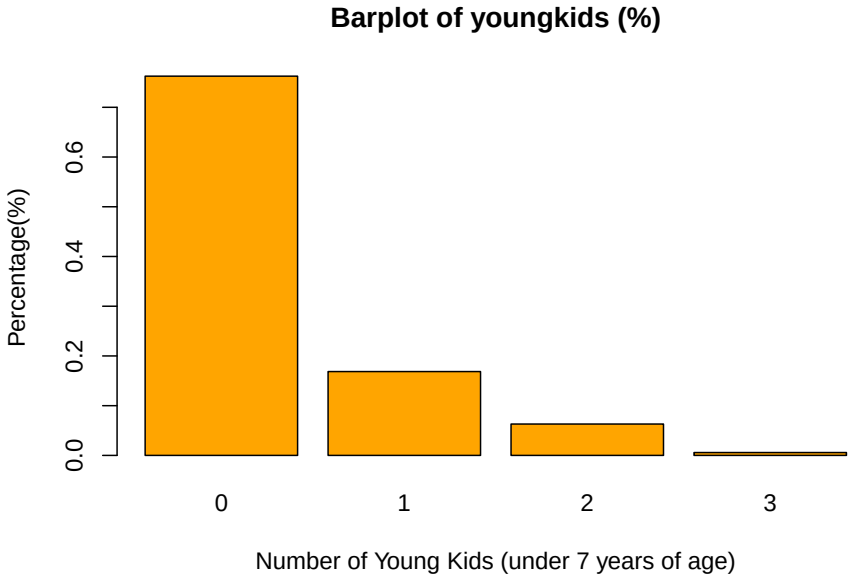
`age :` Age in decades (years divided by 10). The `age` variable is slightly right-skewed and is spread out.



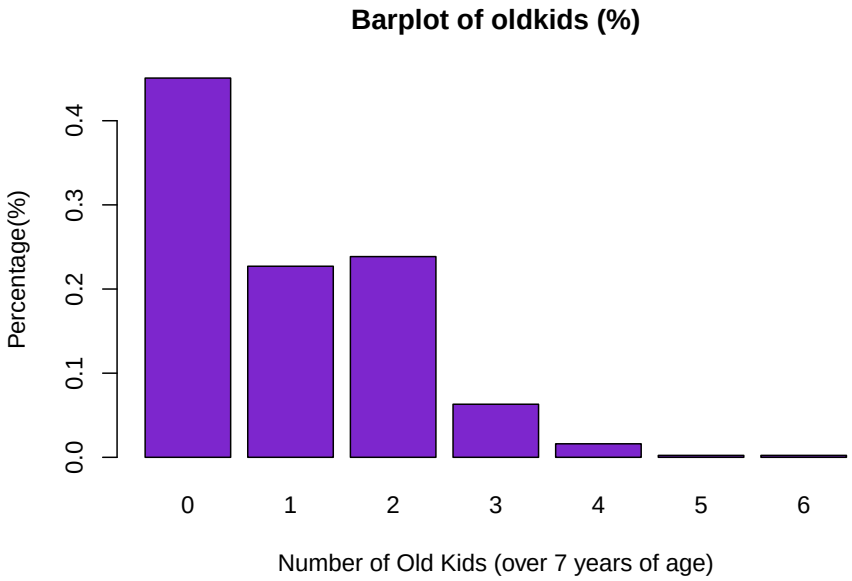
`education :` Years of formal education. The `education` variable seems to be slightly right-skewed.



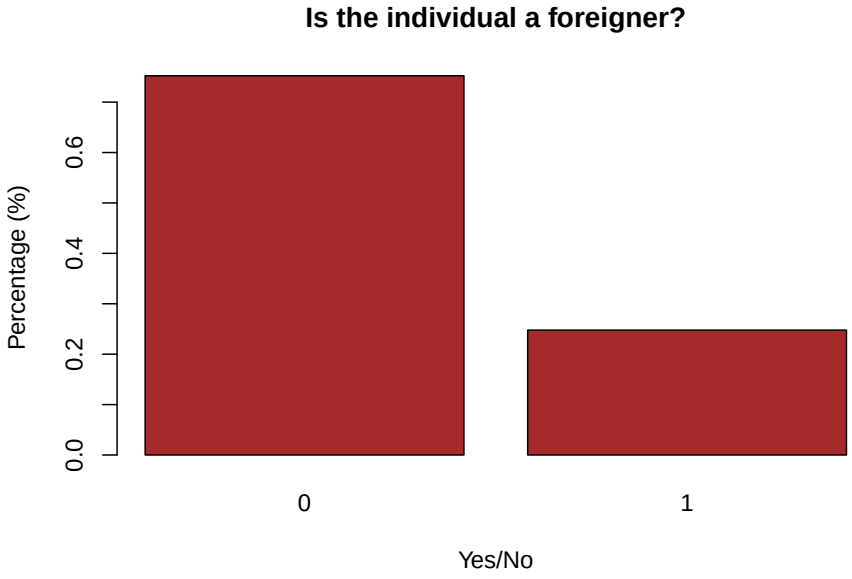
youngkids : Number of young children (under 7 years of age). Approximately, 76.3% of the people have no young kids, 16.9% have 1 young kid, about 6% have 2 young kids, and less than 1% of the people have 3 young kids.



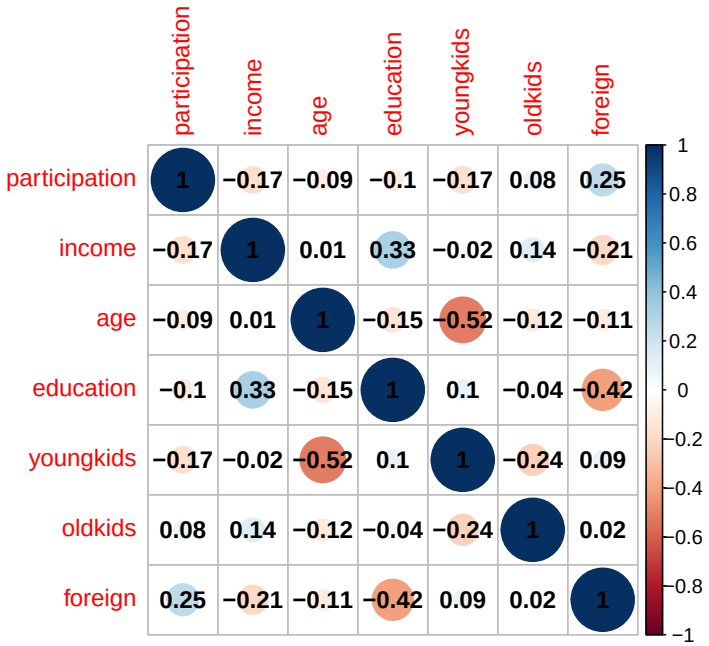
oldkids : Number of older children (over 7 years of age). Starting from the left bar, about 45% have no kids, 22.7% have 1 kid, 23.8% have 2 kids, 6% have 3 kids, 1.6% have 4 kids, 0.2% have 5 kids, and 0.2% have 6 kids that are over 7 years old.

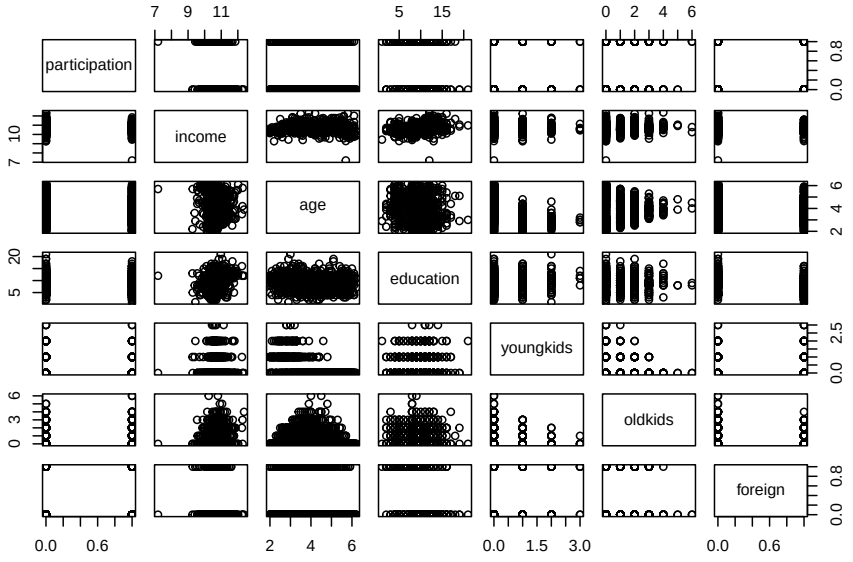


foreign : Factor. Is the individual a foreigner (i.e., not Swiss)? Approximately 75% are citizens of Switzerland and 25% of the people are foreigners (not Swiss).



Correlation Matrix and relationships between variables





The correlations between the dependent variable and the independent variables are not that strong, but we can still see that `income`, `age`, `education`, and `youngkids` are negatively correlated to our binary dependent variable, which means an increase in them leads to a decrease in the probability of participating in the workforce. On the contrary, `oldkids` and `foreign` are positively correlated, meaning that an increase in them results in an increase in the dependent variable `participation`.

Intuitively, these make some sense. For example, parents of young children often need more time to take care of them, resulting in them not working a job. Thus, resulting in the negative correlation for the `youngkids` variable and positive correlation for `oldkids`.

(c) Fit the three models below, and identify which model is your preferred one and why. Make sure to include statistical diagnostics to support your conclusion, and to comment on your findings.

- **Linear Probability Model**

term	estimate	std.error	statistic	p.value
(Intercept)	2.6245038	0.4262291	6.1574950	0.0000000
income	-0.1679172	0.0407161	-4.1240994	0.0000408
age	-0.1062254	0.0183799	-5.7794405	0.0000000
education	0.0072316	0.0060413	1.1970293	0.2316231
youngkids	-0.2595680	0.0317993	-8.1626906	0.0000000
oldkids	-0.0028598	0.0156508	-0.1827265	0.8550554
foreign	0.2847923	0.0405881	7.0166375	0.0000000

For the LPM model, interpreting the marginal effects of the model is relatively straight forward since the marginal effects are constant, which means the coefficient estimates are the marginal effects. But, the effects being constant also implies the following issues:

1. the probability of getting a certain outcome could exceed 1 or could be below 0 (negative)
2. The variance is heteroskedastic, which means robust standard errors or FGLS should be used

Correcting the standard errors using robust standard errors

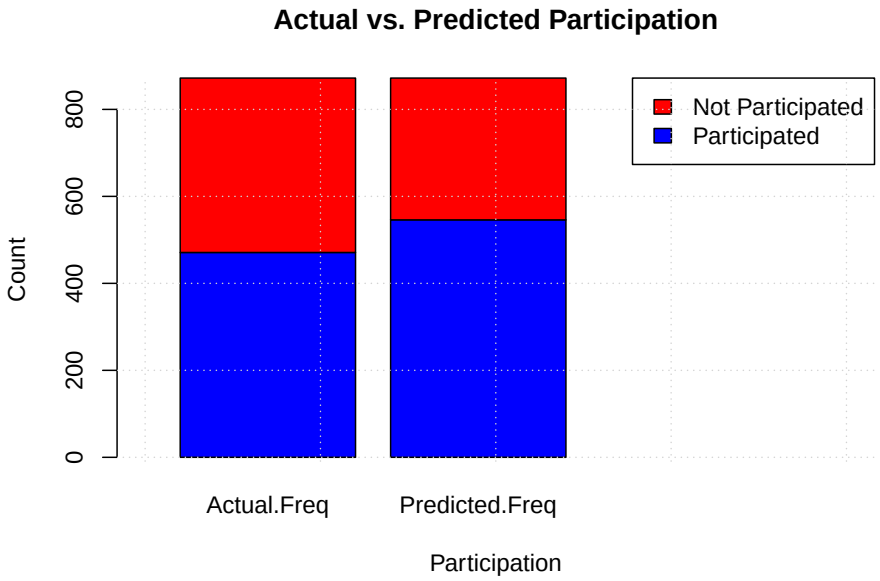
term	estimate	std.error	statistic	p.value
(Intercept)	2.6245038	0.3804837	6.8978088	0.0000000
income	-0.1679172	0.0364368	-4.6084439	0.0000047
age	-0.1062254	0.0175194	-6.0632927	0.0000000
education	0.0072316	0.0059747	1.2103794	0.2264640
youngkids	-0.2595680	0.0306195	-8.4772169	0.0000000
oldkids	-0.0028598	0.0156096	-0.1832086	0.8546773
foreign	0.2847923	0.0405970	7.0151087	0.0000000

interpreting the Marginal Effects:
For example, the average marginal effect (AME) for the `income` variable is **-0.1679172**. This means that every one-unit increase in `income` (Log-nonlabor income) leads to approximately 16.7% decrease in the probability of one participating in the labor force, and the same logic applies to the other variables as well.

Note: ***`education`*** *and* ***`oldkids`*** *variables are not statistically significant, but were still included in the model just for consistency with the other models*

Prediction with LPM
Our LPM predicted that 538 people would not participate and 334 people would participate in the workforce while the real outcome is 471 and 401 respectively.

	Actual.Freq	Predicted.Freq
No	471	546
Yes	401	326



Assessing the performance of LMP

```
## [1] 0.6662844
```

```
## [1] 1126.542
```

As shown above, the probability of LPM correctly predicting the true outcome is approximately 66.6%, and the AIC of the model is 1126.542. These results will be compared with the other models and will be further discussed.

Note: 0.5 threshold was used for this prediction.

- **Probit Model**

term	estimate	std.error	statistic	p.value
(Intercept)	6.3684681	1.2893945	4.9391153	0.0000008
income	-0.5025826	0.1225696	-4.1003862	0.0000412
age	-0.3108509	0.0542387	-5.7311687	0.0000000
education	0.0204050	0.0175281	1.1641344	0.2443695
youngkids	-0.7845398	0.1035216	-7.5785104	0.0000000
oldkids	-0.0134804	0.0448862	-0.3003241	0.7639300
foreign	0.8043432	0.1192583	6.7445484	0.0000000

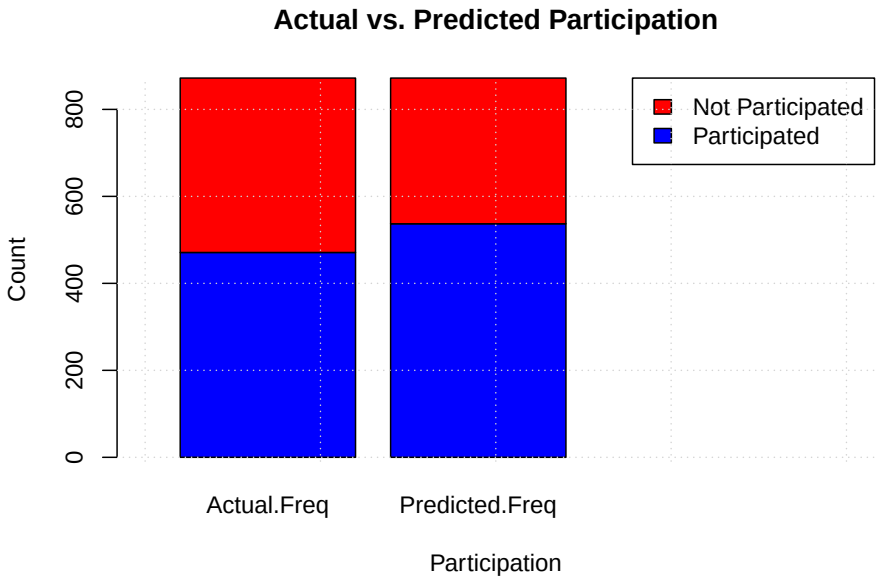
interpreting the Marginal Effects

term	estimate	std.error	statistic	p.value
age	-0.1069480	0.0176141	-6.0717307	0.0000000
education	0.0070203	0.0060165	1.1668408	0.2432747
foreign	0.2767335	0.0377575	7.3292264	0.0000000
income	-0.1729130	0.0409902	-4.2184026	0.0000246
oldkids	-0.0046379	0.0154409	-0.3003651	0.7638986
youngkids	-0.2699201	0.0321591	-8.3932756	0.0000000

Based on the result above (probit marginal effects), we can say that one unit increase in `income` (Log-nonlabor income) leads to approximately 17.3% decrease in the probability of one participating in the labor force, and the same logic applies to the other variables as well. Overall, the probit model estimates the marginal effects slightly higher than LPM did.

Prediction with the Probit Model:
Our probit model predicted that 537 people would not participate and 335 people would participate in the work force while the real outcome is 471 and 401 respectively.

	Actual.Freq	Predicted.Freq
No	471	537
Yes	401	335



Assessing the performance of Probit Model

```
## [1] 0.6651376
```

```
## [1] 1066.983
```

As shown above, the probability of the probit model correctly predicting the true outcome is approximately 66.5%, and the AIC of the model is 1066.983. while the percentage that the probit model and the LPM correctly predicted the outcomes were almost the same, the probit model indicates better performance in AIC.

Note: 0.5 threshold was used for this prediction.

- **Logit Model**

term	estimate	std.error	statistic	p.value
(Intercept)	2.6245038	0.4262291	6.1574950	0.0000000
income	-0.1679172	0.0407161	-4.1240994	0.0000408
age	-0.1062254	0.0183799	-5.7794405	0.0000000
education	0.0072316	0.0060413	1.1970293	0.2316231
youngkids	-0.2595680	0.0317993	-8.1626906	0.0000000
oldkids	-0.0028598	0.0156508	-0.1827265	0.8550554
foreign	0.2847923	0.0405881	7.0166375	0.0000000

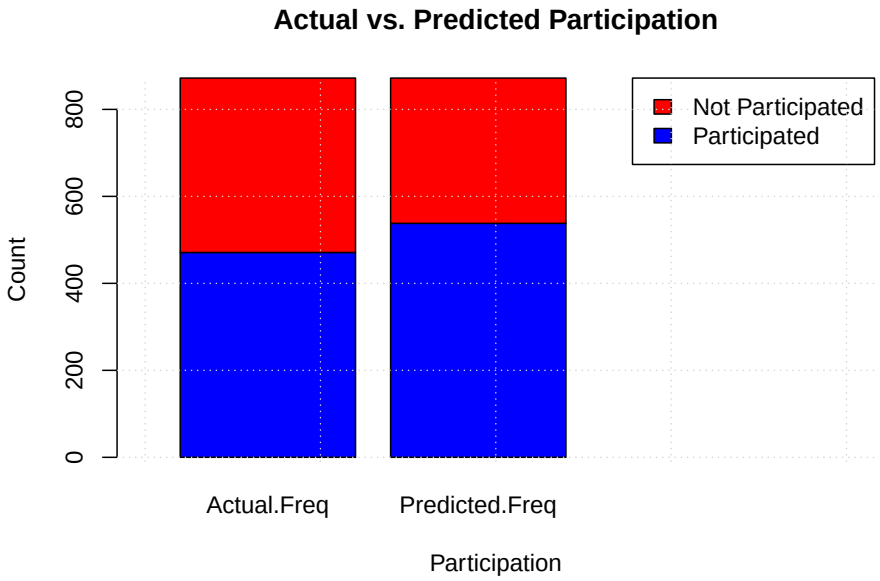
interpreting the Marginal Effects of the logit model

term	estimate	std.error	statistic	p.value
age	-0.1064081	0.0175899	-6.0493762	0.0000000
education	0.0066156	0.0060386	1.0955378	0.2732811
foreign	0.2732306	0.0377130	7.2449981	0.0000000
income	-0.1699428	0.0414555	-4.0994002	0.0000414
oldkids	-0.0045842	0.0153783	-0.2980955	0.7656303
youngkids	-0.2774672	0.0332563	-8.3433052	0.0000000

Based on the result above (logit marginal effects), we can say that one unit increase in `income` (Log-nonlabor income) leads to approximately 16.99% decrease in the probability of one participating in the labor force, and the same logic applies to the other variables as well. Overall, the logit model estimates the marginal effects slightly lower, compared to the probit model.

Prediction with the Logit Model:
Our logit model predicted that 538 people would not participate and 334 people would participate in the work force while the real outcome is 471 and 401 respectively.

	Actual.Freq	Predicted.Freq
No	471	538
Yes	401	334



Assessing the performance of Logit Model

```
## [1] 0.6662844
```

```
## [1] 1066.798
```

As shown above, the probability of the logit model correctly predicting the true outcome is approximately 66.6%, and the AIC of the model is 1066.793.

(d) Comparing the three models:

Although the AIC values are very close between the probit model and the logit model, the logit model shows a slightly better performance AIC as well as the probability of predicting the outcomes correctly.

```
##            LPM       Probit        Logit
## 1    0.6662844    0.6651376    0.6662844
## 2 1126.5415837 1066.9827120 1066.7975023
```

Therefore, the logit model best describes the data based on the predictive performance.

(e) Creating a function that could predict any inputs using the logit model:

Using the best model we have selected, I have created a function, using the logit model estimates, that takes any values of independent variables. So now, we can predict the probability of one participating in the labor force in a hypothetical situation.

```
## (Intercept)      income         age   education   youngkids     oldkids
## 10.37434616 -0.81504064 -0.51032975  0.03172803 -1.33072362 -0.02198573
##     foreign
##  1.31040497
```

Q. What is the probability that a 40-year-old foreigner with a nonlabor income of $20,000, 10 years of education, 2 young kids, and 1 old kid participates in the labor force in Swiss?

```
## [1] 0.3112107
```

The logit model outputs that the person in question would participate in the labor force in Switzerland with the probability of approximately 31%.