Lab06: Logistic Regression

**Handed out:** Thursday, November 12, 2020

**Return date:** Tuesday, November 24, 2020

**Grading:** This lab counts 6 % towards your final grade

**Format of answer:** Your answers (statistical figures and verbal description) should be submitted as ***hardcopy***. Add a running title with the following information: Lab06, your name and page numbers. You may use this document as template. Copy the requested statistical figures into your document. Trial and error answers will lead to a deduction of points. Label each answer properly with the bold task headings. You are expected to hand in professionally formatted answers: use a fixed pitch font, like **Courier New**, for any  code and output. Copy and paste figures into your document. Make sure that each figure has a proper ***caption*** describing its content.

# Part 1: Logistic Regression Model for a Binary Outcome [4 points]

## Data

You will be working with the data set **Mroz** which is in the **car** library.   
The data can be read with   
**> library(car)  
> data(Mroz)  
> attach(Mroz)**

The dependent variable in the data set is the wife's labor-force participation.

|  |  |
| --- | --- |
| Variable | Description |
| **lfp** | wife labor-force participation; a factor with levels: 'no'; 'yes' |
| **k5** | number of children 5 years old or younger |
| **k618** | number of children 6 to 18 years old |
| **age** | wife’s age in years |
| **wc** | wife's college attendance; a factor with levels: 'no'; 'yes' |
| **hc** | husband's college attendance; a factor with levels: 'no'; 'yes' |
| **lwg** | log expected wage rate; for women in the labor force, the actual wage rate; for women not in the labor force, an imputed value based on the regression of 'lwg' on the other variables. |
| **inc** | family income exclusive of wife's income |

*More information on this data set can be found in the online help of the car library.*

**Task 1:** Specify with common sense arguments into which directions ***all*** independent variables may influence the wife’s propensity to participate in the labor force. Use a ***table*** to with the headings [a] variable name, [b] argument and [c] null and alternative hypotheses for your answer. [0.5 point]

**Task 2:** Model discussion [1.5 points]

[a] Build a logistic regression model for the probability of **lfp** with these independent variables and give the 95% confidence intervals around the estimated logistic regression parameters

[b] ***Discuss*** your model output in the light of your stated hypotheses from task 1.

[c] Interpret the calibrated logistic regression model in terms of ***probabilities*** by using an ***all effects plot*** (i.e., the “other” variables are at their average level).

**Task 3:** Perform one likelihood ratio tests [0.5 point]

Refine the model from task 2 by dropping all variables which you deem to be not relevant. Test whether these variables jointly have explanatory power or not. Properly state the null and the alternative hypotheses.

**Task 4:** Conditional effects plots [1.5 points]

Generate conditional effects plots based on the refined model for the probability of labor force participation for the income variable **inc**. Interpret the plots.

Assume two scenarios with the following values levels of the additional independent variables in the logistic regression model:

|  |  |  |
| --- | --- | --- |
| Variable | Low Probability | High Probability |
| k5 | 2 | 0 |
| age | 49 | 36 |
| wc | 'no' | 'yes' |
| lwg | 0.81 | 1.40 |

Discuss your plots for the two scenarios.

# Part 2: Logistic Regression for Rates [2 points]

Continue using the **TexMix** package. The script **LogisiticPctPopPov.r** is setting your data up for a logistic regression analysis of the percentage of the population below the poverty threshold.

**Task 5:** Before you can run a logistic regression analysis on the variable **PCTPOPPOV** you need to consider two issues. [0.5 points]

1. The measurement units of **PCTPOPPOV** do not match the expected range of probabilities . Perform a proper transformation of the variable **PCTPOPPOV**. Justify your selected transformation.
2. The variable **PCTPOPPOV** has an underlying denominator which needs to be used in **glm( )** function to account for the size of each census tract. Justify your choice of the underlying denominator?

**Task 6:** Run a logistic regression model of the transformed **PCTPOPPOV** with a proper **weights** setting using the set of 4 independent variables **PCTUNEMP+PCTPUB2WRK+PCTASIAN+PCTUNIVDEG**. [0.5 points]

**Task 7:** Interpret the model by using default conditional effects plots for the 4 independent variables. [1.0 point]