Lab05: Logistic & Poisson Regression

**Handed out:** Monday, April 19, 2021

**Return date:** Monday morning, May 3, 2021 (this is a ***firm date***: the sample answer will be posted afterwards)

**Grading:** This lab counts 12 % 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: Lab05, 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. Use mathematical typesetting when equations are required. Copy and paste figures into your document. Make sure that each figure has a proper ***caption*** describing its content.

**Notes:**

* Special office hours will be held on Sunday May 9, 2021, from 2:00 to 4:00 pm via MS Teams for last minute questions.
* The final exam will be held via eLearning on Monday May 10, 2021, for three hours in the 24 hours window midnight to midnight central time.

# Part 1: Logistic Regression Model for a Binary Outcome [6 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. [1 point]

**Task 2:** Model discussion [2 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 a likelihood ratio test [1 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 in statistical terminology the null and the alternative hypotheses.

**Task 4:** Conditional effects plots [2 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: How does the labor force participation probability vary for women in both groups?

# Part 2: Poisson and Logistic Regression [4 points]

Use the data-frame **cancer** in the library **CancerSEA**. You can install the library with the  command **install.packages("*Drive:*\\*Path*\\CancerSEA\_0.9.6.tar.gz", repos=NULL)**. Show your results and briefly discuss them.

**Task 5:** Run a Poisson regression model on the annual ***raw counts of white male lung cancer deaths*** for the period 1970 to 1994. Make sure to use a ***proper offset*** in the link-function specification to account for the ***expected number of death*** based on the population size and age distribution in each State Economic Area. [1 points]

Select as independent variables **~URBRUR+RAD\_MD+I(RAD\_MD^2)+TOBACCO**.

**Task 6:** Run a logistic regression model for the ***directly age-standardized*** ***white male lung cancer death rates*** per 100.000 persons at risk. Caution: you need to re-scale the rates, so they become probabilities. Since we are dealing with a binomial distribution rather than a binary distribution you need to specify a proper weight variable to account for the population at risk, i.e., half of the population in 1980. [1 point]

Select as independent variables **~URBRUR+RAD\_MD+I(RAD\_MD^2)+TOBACCO**.

**Task 7**: For the logistic regression model from task 6 generate a conditional effects plot with respect to **RAD\_MD** (all other variable at their average levels). Make sure to have probabilities on y-axis and not logits. [1 point]

**Task 8:** Rerun the model from task 6 allowing explicitly modeling potential ***over-dispersion***. Compare both models and interpret the estimated over-dispersion parameter. [1 point]

# Part 3: Modeling Interregional Migration with Poisson Regression [2 points]

The dataset **UPFING.SAV** holds information about the 1976 to 1981 migration flows among the 10 Canadian provinces in the variable **MIJ**, where “I” stands for the origin and “J” for the destination. Additional variables are **PI** and **PJ** for the provincial population counts of the origin and destination as well as **DIJ** for the interprovincial distances between the main provincial cities in kilometers. Note: Internal provincial migration flows and internal provincial distances are not available. Therefore, observations for which the origin and destination are identically (i.e., **I==J**) need to be ***excluded*** from the analysis.

**Task 9:** Estimate the basic gravity model with Poisson regression and transforming the right-hand-side of the equation into a linear equation in the unknown regression coefficients and by applying the logarithm. [1 point]

**Task 10:** Interpret the estimate regression coefficients in terms of their estimated signs. How do the origin and destination populations as well as the interprovincial distances influence the migration flows? [1 point]