Lab05: Logistic & Poisson Regression

**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 type-setting when equations are required. 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 [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]

|  |  |  |
| --- | --- | --- |
| Variable Name | Common sense assumption | Hypothesis |
| k5 | More children with a 5-year-old or younger, a higher probability of moms staying at home due to the responsibility. |  |
| k618 | More children within 6 to 18 would lead to a lower probability of labor-force participation of wives since those children still need to take care of. |  |
| age | Older age with a less probability of labor-force participation since they may have enough money at that time. |  |
| wc | higher education with a higher probability of labor-force participation since they have more chance and motivation. |  |
| hc | Higher husband education rates would lead to a lower probability of labor-force participation of wives considering they may have higher incomes. |  |
| lwg | Of course, a higher expected wage rate corresponding to a higher probability of labor-force participation since it increases their motivation. |  |
| inc | Higher family income would lead to a lower probability of labor-force participation of wives since they have lower motivation. |  |

**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

|  |
| --- |
| GLM.01 <- glm(lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family=binomial(logit), trace=TRUE, data=Mroz)  summary(GLM.01) #slope is for logit, not for probability  glm(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family = binomial(logit),  data = Mroz, trace = TRUE)  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) 3.182140 0.644375 4.938 7.88e-07 \*\*\*  k5 -1.462913 0.197001 -7.426 1.12e-13 \*\*\*  **k618 -0.064571 0.068001 -0.950 0.342337**  age -0.062871 0.012783 -4.918 8.73e-07 \*\*\*  wcyes 0.807274 0.229980 3.510 0.000448 \*\*\*  **hcyes 0.111734 0.206040 0.542 0.587618**  lwg 0.604693 0.150818 4.009 6.09e-05 \*\*\*  inc -0.034446 0.008208 -4.196 2.71e-05 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  **Null deviance: 1029.75** on 752 degrees of freedom  **Residual deviance: 905.27** on 745 degrees of freedom  AIC: 921.27  confint(GLM.01, level=0.95, type="Wald",trace = FALSE)  2.5 % 97.5 %  (Intercept) 1.93697359 4.46630794  k5 -1.86089654 -1.08747196  **k618 -0.19839650 0.06867096**  age -0.08830325 -0.03813509  wcyes 0.36099360 1.26377557  **hcyes -0.29200419 0.51679061**  lwg 0.31402218 0.90697688  inc -0.05099767 -0.01877093 |

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

|  |
| --- |
| Almost all variables fit the previous assumption with a significant impact on the dependent variable except 'k618' and 'hc'.  For 'k618', I guess since children within 6 to 18 would stay at schools most of the time, it is not a big problem for moms to take care of them after work.  For "husband education", it surprised me it does not have a significant influence on the dependent variable. However, family income still has a significant impact but a low slope, which means women are more independent than I assumed. |

[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).

|  |
| --- |
| plot(allEffects(GLM.01), type="response", ylim=c(0,1), ask=FALSE) |
| K5: A higher value leads to a lower probability. For most families have 0 or 1 child with the age from 0 to 5, they have 60% and 30% probability to take work, respectively. If this value larger than 1, wives are very unlikely to take works. |
| K618: The regression line is almost flat, which means not many influences from it. The average around 58% means there are 58% of our observed records have a work. |
| Age: Higher age lower probability to work. The average age of 45 years old corresponds to a 50% probability to work. |
| Wife Education: A higher education level corresponds to a higher probability to work. |
| Husband education: flat, not many impacts. |
| Family income: A Higher family income would lower the probability to work. |
| Expected wage rate：A Higher expected wage rate would increase the probability to work, and this slope is sharpest since the wage is the first considering when decided to work or not. |

**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.

|  |
| --- |
| ( LR <- -2\*(GLM.02$deviance-GLM.02$null.deviance) )  ( pchisq(LR[1], df=2, lower.tail=F) )  **2.853685e-54** |
| **Null hypothesis:** The slope of all independent variables is equal to 0. (intercept model)  **Alternative hypothesis:** At least one of those slopes are not 0. |
| **Result:** The deviance of the intercept model is 1029.75, and for the full model (unrestricted model) is 905.27. From the likelihood ratio test, we could find the full model reduce the deviance substantially, so our model is acceptable. The independent variables have enough power to explain the variance among dependent records. |

**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? (conditional effects plot)

|  |
| --- |
| # Low prob respondent  eff.GLM.low <- effect("inc",GLM.02, given.values = c(k5 = 2,age = 49,"wcyes" = 0,lwg = 0.81))  plot(eff.GLM.low, type="response", ylim=c(0,1), ylab=expression(Pr(Y[i]=="Close")),  main="Low Probability Respondents")  # High prob respondent  eff.GLM.hi <- effect("inc",GLM.02, given.values=c(k5 = 0,age = 49,"wcyes" = 1,lwg = 1.40))  plot(eff.GLM.hi, type="response", ylim=c(0,1), ylab=expression(Pr(Y[i]=="Close")),  main="High Probability Respondents")  **Low Probability Case:** If there is a 49 years old woman with 2 children under 5, and she does not have a college degree, it is very unlikely for her to work. As shown in graphs, unless the family income is pretty low, which means she may have to work for a living, the probability of working is near to 0.  **High Probability Case:** There is a 36 years old woman without children under 5. Additionally, she has a college degree, and the wage for her is over her expected. Unless her family is rich enough, she has a strong motivation to work. But when the probability goes up, the uncertainty also raises. Since not many people have that high family income, the sample size is small. |

# 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**.

|  |
| --- |
| glm(formula = L\_WM\_P2\_CN ~ URBRUR + RAD\_MD + I(RAD\_MD^2) + TOBACCO,  **family = poisson(link = "log")**, data = cancer, **weights = (POP1980/2)**, **offset = log(POPATRISK1982))**  Deviance Residuals:  Min 1Q Median 3Q Max  -65045 -3254 -484 1586 95432  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -4.757e+00 8.916e-06 -533494 <2e-16 \*\*\*  URBRURurban 6.286e-01 5.144e-06 122220 <2e-16 \*\*\*  RAD\_MD 6.577e-01 1.090e-05 60315 <2e-16 \*\*\*  I(RAD\_MD^2) -2.503e-01 4.288e-06 -58360 <2e-16 \*\*\*  TOBACCO 1.790e-02 4.536e-07 39474 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for poisson family taken to be 1)  Null deviance: 7.2989e+10 on 507 degrees of freedom  Residual deviance: 4.1508e+10 on 503 degrees of freedom  AIC: 4.2671e+10  Number of Fisher Scoring iterations: 5 |

**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 (half of the population) proper weight variable to account for the population at risk, i.e., half of the population in 1980. [1 points]

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

|  |
| --- |
| cancer$L\_WM\_P2\_RT\_rate <- scales::rescale(L\_WM\_P2\_RT, to = c(0, 1))  glm(formula = L\_WM\_P2\_RT\_rate ~ URBRUR + RAD\_MD + I(RAD\_MD^2) +  TOBACCO, **family = binomial(logit), data = cancer, weights = (POP1980/2)**, trace = TRUE)  Deviance Residuals:  Min 1Q Median 3Q Max  -602.56 -57.80 7.34 78.64 465.77  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -9.381e-01 1.166e-03 -804.5 **<2e-16 \*\*\***  URBRURurban 1.312e-02 3.917e-04 33.5 **<2e-16 \*\*\***  **RAD\_MD -8.782e-01 8.651e-04 -1015.1 <2e-16 \*\*\***  **I(RAD\_MD^2) 1.706e-01 2.628e-04 649.2 <2e-16 \*\*\***  TOBACCO 9.113e-02 5.246e-05 1737.3 **<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: 11640706 on 507 degrees of freedom  Residual deviance: 7607219 on 503 degrees of freedom  AIC: 7613369  Number of Fisher Scoring iterations: 4 |

**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]

|  |
| --- |
| eff.GLM.average <- effect("RAD\_MD",GLM.01)  plot(eff.GLM.average, ylim=c(0,1), type="response", ylab=expression(Pr(Y[i]=="Close")), main="Average Probability Respondents")    It shows the effects of radon gas have a turning point around 2, which means when the density of it is very low, they would not have a significant harmful impact. But when the density keeps increasing, the effects raise sustainably since it is a quadric term. |

**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]

|  |
| --- |
| glm(formula = L\_WM\_P2\_RT\_rate ~ URBRUR + RAD\_MD + I(RAD\_MD^2) +  TOBACCO, family = quasibinomial, data = cancer, weights = (POP1980/2), trace = TRUE)  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -0.938137 0.139689 -6.716 5.06e-11 \*\*\*  **URBRURurban 0.013123 0.046927 0.280 0.78**  RAD\_MD -0.878201 0.103634 -8.474 2.64e-16 \*\*\*  I(RAD\_MD^2) 0.170623 0.031485 5.419 9.30e-08 \*\*\*  TOBACCO 0.091135 0.006284 14.503 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for quasibinomial family taken to be 14350.21)  Null deviance: 11640706 on 507 degrees of freedom  Residual deviance: 7607219 on 503 degrees of freedom  AIC: NA  Number of Fisher Scoring iterations: 4  All regression coefficients remain the same value (unbiased) but scales shrink 100 times. This difference would also spill out to the standard error, and then shown in the t-test result. That is why the Rural-Urban variable becomes no longer significant. |

# 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]

|  |
| --- |
| glm(formula = MIJ ~ lnPI + lnPJ + lnDIJ, **family = poisson(log),**  data = upfing, weights = CWT)  Deviance Residuals:  Min 1Q Median 3Q Max  -253.43 -60.75 -32.42 5.52 408.13  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -7.0409657 0.0241716 -291.3 <2e-16 \*\*\*  lnPI 0.7142001 0.0009334 765.2 <2e-16 \*\*\*  lnPJ 0.5838710 0.0009029 646.7 <2e-16 \*\*\*  lnDIJ -0.3154750 0.0011822 -266.9 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for poisson family taken to be 1)  Null deviance: 2097541 on 89 degrees of freedom  Residual deviance: 891051 on 86 degrees of freedom  AIC: 891967  Number of Fisher Scoring iterations: 5 |
| glm(formula = normal\_MIJ ~ lnPI + lnPJ + lnDIJ, **family = quasibinomial**,data = upfing, trace = TRUE)  Deviance Residuals:  Min 1Q Median 3Q Max  -0.88449 -0.14258 -0.07105 0.00757 1.63916  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -23.3224 3.4282 -6.803 1.29e-09 \*\*\*  lnPI 0.9508 0.1414 6.726 1.83e-09 \*\*\*  lnPJ 0.8032 0.1378 5.830 9.48e-08 \*\*\*  lnDIJ -0.5797 0.1862 -3.114 0.00251 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for quasibinomial family taken to be **0.1222711**)  Null deviance: 21.7069 on 89 degrees of freedom  Residual deviance: 9.9829 on 86 degrees of freedom  AIC: NA  Number of Fisher Scoring iterations: 6 |

**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]

|  |
| --- |
| Since the dispersion parameter is << 1, which means our observations are under-dispersion, so we applied quasi model here.  For the regression coefficients, the number of populations on both destinations and origins has a positive impact on migration. It makes sense since more population influx and outflux in metropolitans than small towns. When the population number goes larger, the curvature of the regression line also goes sharper.  Interprovincial distances harm migrations. But when the distance goes large enough, the influence becomes minor, which means beyond this distance, people concerns about other things more. |