Sample Answer Lab05: Logistic Regression

Handed out: Thursday, November 14, 2019

Return date: Friday, November 22, 2019

Grading: This lab counts 8 % towards your final grade

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]

Independent Variable	Influence on the dependent variables	Hypothesis
K5	Negative direction. Because a mother usually spends a lot of time to take care of a 5-year-old or younger child. All children rely on their mothers until they can be independent. So the mothers does not have much time to work. The more toddlers there are in a household, the less possible it becomes for the wife to work.	$H_0: \beta_{K5} \geq 0$ $H_1: \beta_{K5} < 0$
K618	Negative direction. Although the children are older, they still require some attention. However, since they are for the most time at school, it gives their parents more flexibility. The probability for wives to	$H_0: \beta_{K618} \geq 0$ $H_1: \beta_{K618} < 0$

	work is higher than that of those wives with 5-year-old or younger children.	
age	Negative direction. When persons get older, they may not have enough energy and their skills may have become rusty and therefore they are not as attractive to employers anymore. So the probability decreases as the age increases.	$H_0: \beta_{age} \geq 0$ $H_1: \beta_{age} < 0$
wc	Positive direction. If a wife has a college attendance, she gains knowledge and skills to work. In addition, wives who have higher education degrees usually prefer the satisfaction and independence that a good job brings. The wife's labor-force participation probability increases as the wife's college attendance increases.	$H_0: \beta_{wc} \leq 0$ $H_1: \beta_{wc} > 0$
hc	Positive direction. If a husband has college attendance, it is becomes easier for him to support his wife's to find a job. The wife's labor-force participation probability increases as the husband's college attendance increases.	$H_0: \beta_{hc} \leq 0$ $H_1: \beta_{hc} > 0$
lwg	Positive direction. Larger expected wages encourage wives to work. The wife's labor-force participation probability increases as the log expected wage rate increases.	$H_0: \beta_{lwg} \leq 0$ $H_1: \beta_{lwg} > 0$
inc	Negative direction. In affluent families, wives usually do not need to work to support their families. The wife's labor-force participation probability decreases as family income increases.	$H_0: \beta_{inc} \geq 0$ $H_1: \beta_{inc} < 0$

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.

```
log1<- glm(lfp ~ k5 + k618 + age + wc + hc + lwg + inc,
           family=binomial(logit), trace=TRUE, data=Mroz)
summary(log1)

Call:
glm(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family = binomial(logit),
    data = Mroz, trace = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1062  -1.0900   0.5978   0.9709   2.1893

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
```

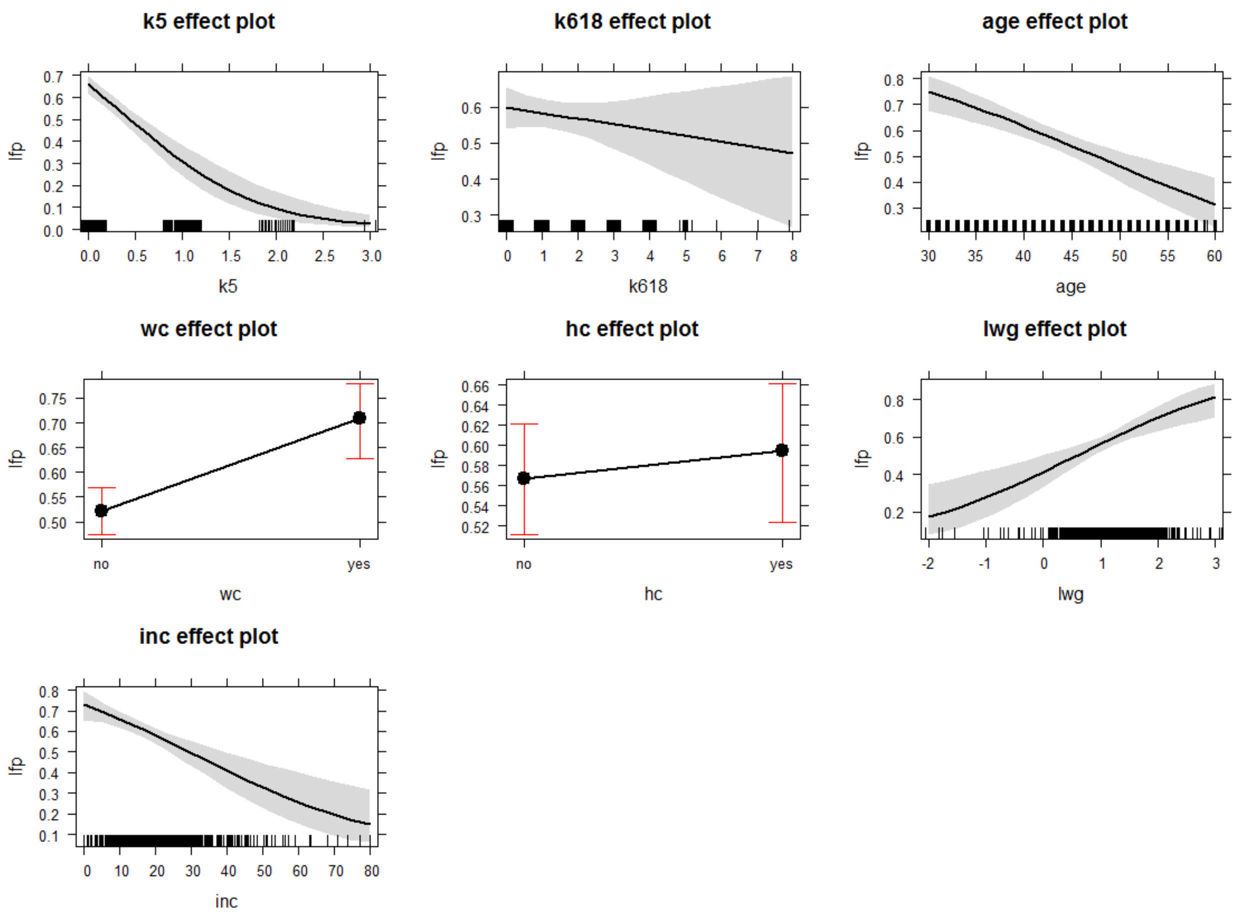
```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  905.27  on 745  degrees of freedom
AIC: 921.27

Number of Fisher Scoring iterations: 4
#5% confidence level
confint(log1, level=0.95, type="Wald")
                  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.

Comments: If the confidence interval of a coefficient includes 0, this coefficient is not significant. When we observe the confidence interval of other significant coefficients, only the confidence interval of k618 and hcyes includes 0. Therefore, except k618 and hcyes, all coefficients are significant.
The direction of all regression coefficients are the same as my hypotheses stated in the previous task.

- K5, k618, age and inc have the negative affect on the wife's labor-force participation.
- While wcyes, hcyes, and lwg have the positive effect on the probability of wife's labor-force participation.
- Within those significant regression coefficients, 5-year-old or younger kids have the largest negative affect on the probability of wife's labor-force participation.
- Wives who have the college attendance and log expected wage rate have largest positive effects.
- Other variables have little effect on the wife's labor-force participation, regarding to their small absolute regression coefficients.

[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).



Comments: The estimated probability for the observation can be expressed by $\hat{\pi}_i = \frac{1}{1+\exp(-x_i^t * b)}$. Thus, the interpretation in terms of probability is not the same as that of the linear regression. The slope of the logistic curve for a given probability $\hat{\pi}_i$ is $b_k \cdot \hat{\pi}_i \cdot (1 - \hat{\pi}_i)$. The effects plot can be used to interpret how one variable affect the probability when other variables are held at fixed levels; in this case the mean. If a family has no children under 5-year old, the probability of women would like to work is 65%. However, if a woman has three children under 5-year old, the probability drops to 5%. If a woman does not have children between 6 to 18 years old, the probability she works is 60%. When the number of 6 to 18 years old children increases to eight, this probability decreases to 48%. When a woman's age increases from 30 to 60, her willing to work decreases from 75% to 35%. Moreover, a woman with college education has 72% probability to work whereas only 52% if she does not have the college degree. If a woman's husband has college education, the probability for that woman to work is 59%, whereas 57% if her husband does not have a college degree. When the log wage rate raises from -2 to 3, the probability for a woman would to work increases from 18% to 78%. When other variables in the model are held at the mean level, the increasing family income causes the wife labor-force participation probability decreases from 70% to 15%.

Task 3: Perform one likelihood ratio tests [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 the null and the alternative hypotheses.

```
#refine the model
log2<- glm(lfp ~ k5 + age + wc + lwg + inc, family=binomial(logit), trace=TRUE,
data=Mroz)
summary(log2)
Call:
glm(formula = lfp ~ k5 + age + wc + lwg + inc, family = binomial(logit),
    data = Mroz, trace = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0428  -1.0853   0.6015   0.9697   2.1801

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.90193    0.54290   5.345 9.03e-08 ***
k5          -1.43180    0.19320  -7.411 1.25e-13 ***
age         -0.05853    0.01142  -5.127 2.94e-07 ***
wcyes        0.87237    0.20639   4.227 2.37e-05 ***
lwg          0.61568    0.15014   4.101 4.12e-05 ***
inc         -0.03367    0.00780  -4.317 1.58e-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:  906.46  on 747  degrees of freedom
AIC: 918.46

Number of Fisher Scoring iterations: 3

#log likelihood test
logLik(log2)
'log Lik.' -453.2277 (df=6)
logLik(log1)
'log Lik.' -452.633 (df=8)
LR<- -2 *(logLik(log2) - logLik(log1))
pchisq(LR[1], df = 2, lower.tail = F)
[1] 0.5516976

> anova(log2, log1, test = "LRT")
Analysis of Deviance Table

Model 1: lfp ~ k5 + age + wc + lwg + inc
Model 2: lfp ~ k5 + k618 + age + wc + hc + lwg + inc
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       747     906.46
2       745     905.27  2   1.1895   0.5517
```

Comments: In the first model, the variables k618 and hcyes are not significant. The null hypothesis in this likelihood ratio test is $\mathrm{H}_0: \beta_{k618} = \beta_{hcyes} = 0$. The alternative hypothesis is β_{k618} and β_{hcyes} are not equal to zero simultaneously. The p-value is 0.5517, so we fail to reject the null hypothesis. k618 and hc are not jointly significant so we do not need to include these variables into the model.

Task 4: Conditional effects plots [2 points]

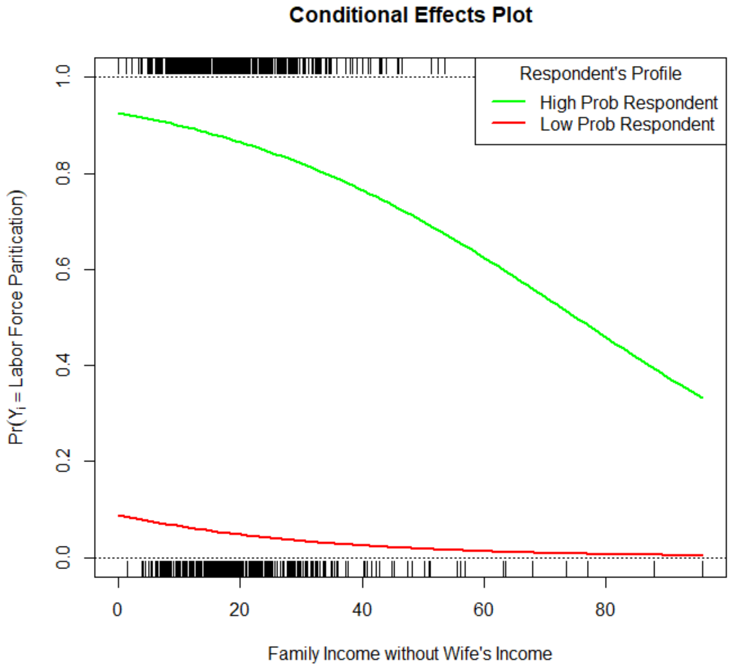
Generate conditional effects plots based on the refined model for the probability of labor force participation for the income variable **inc**. Interpret the plots.

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.

```
library(effects)
summary(Mroz)
invlogit <- function(x) 1/(1+exp(-x))

plot(Mroz$inc, Mroz$lfp, type="n", ylim=c(0,1), ylab=expression(Pr(Y[i]=="Labor Force
Paritication")),
       xlab="Family Income without Wife's Income", main="Conditional Effects Plot")
rug(Mroz$inc[as.numeric(Mroz$lfp)-1==0])
rug(Mroz$inc[as.numeric(Mroz$lfp)-1==1], side=3)
abline(h=c(0,1),lty=3)
curve(invlogit(cbind(1,x,2,49,0,0.81) %*% coef(log2)), col="red", lwd=2, add=TRUE)
curve(invlogit(cbind(1,x,0,36,1,1.40) %*% coef(log2)), col="green", lwd=2, add=TRUE)
legend("topright",legend=c("High Prob Respondent", "Low Prob Respondent"),
       title="Respondent's Profile", lwd=2,col=c("green","red"))
```



Comments: The conditional effects plot shows how the probabilities of labor force participation varies with regards to the family income for either a woman having a high likelihood of participating in the labor force or not. These likelihoods are based on specific characteristics of the woman and her family situation. Conditional effect plots are used to demonstrate the non-linear relationship between two variables as well as the dependence of this relationship on the remaining variables in the model.

The plot clearly shows that family income does not have a strong effect on the propensity of participating in the labor force for women who are tied up at home taking care of children, who are older and who do not have much of a prospect of adding substantially to the family income. It starts at a probability of 15% for a low family income and approaches zero rapidly for increasing family income. In contrast, women who are more flexible, younger and highly educated start at a labor force participation of 95% for a low family income. The propensity of labor force participation gradually declines with increasing income and reaches 30% for a family income of $100,000.

Part 2: Logistic Regression for Rates [2 points]

Continue using the **DallasTracts** 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]

a. The value range of **PCTPOPPOV** does not match the expected range of probabilities from ***zero to one***. Perform a proper transformation of the variable **PCTPOPPOV**. Justify your transformation.

```
> tractShp$ProbPctPopPov <- tractShp$PCTPOPPOV/100
```

b. 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?

   The nighttime population is the denominator of the rate of persons living in poverty in a census tract.

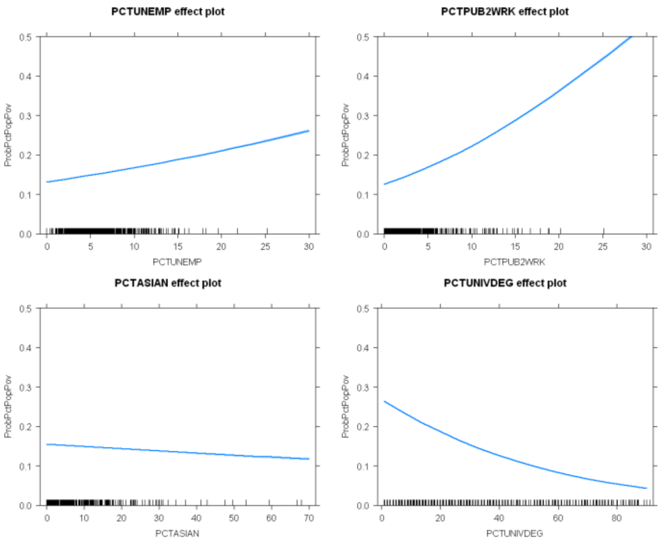
Task 6: Run a logistic regression model of the transformed **PCTPOPPOV** with the proper weights setting using the independent variables **PCTUNEMP+PCTPUB2WRK+PCTASIAN+PCTUNIVDEG**. [0.5 points]

```
> glm1 <- glm(ProbPctPopPov~PCTUNEMP+PCTPUB2WRK+PCTASIAN+PCTUNIVDEG, data=tractShp,
+              weights=tractShp$NIGHTPOP, family=binomial(link = "logit"))
> summary(glm1)

Call:
glm(formula = ProbPctPopPov ~ PCTUNEMP + PCTPUB2WRK + PCTASIAN +
    PCTUNIVDEG, family = binomial(link = "logit"), data = tractShp,
    weights = tractShp$NIGHTPOP)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-40.603   -8.912   -1.011    7.684   38.812
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.3652686  0.0050156 -272.21   <2e-16 ***
PCTUNEMP     0.0283250  0.0005068   55.89   <2e-16 ***
PCTPUB2WRK   0.0686421  0.0004461  153.87   <2e-16 ***
PCTASIAN    -0.0046790  0.0002686  -17.42   <2e-16 ***
PCTUNIVDEG  -0.0233273  0.0001175 -198.54   <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: 234085  on 526  degrees of freedom
Residual deviance:  94974  on 522  degrees of freedom
AIC: 99202

Number of Fisher Scoring iterations: 4
```

Task 7: Interpret the model by using a default conditional effects plot. [1.0 point]

```
> plot(allEffects(glm1), type="response", ylim=c(0,0.5), ask=FALSE)
```



As expected, if in a census tract the unemployment rate and the percentage of population who take public transportation to work increases, so does the percentage of the population living in poverty. In contrast, if the percentage of Asian and percentage with a university degree increases, the percentage living in poverty decreases.