Sample Answer Lab05: Logistic & Poisson Regression

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 work is higher than that of those wives with 5-year-old or younger children.	$H_0: \beta_{K618} \geq 0$ $H_1: \beta_{K618} < 0$
age	Negative direction. When persons get older, they may not have enough energy and their skills may have become rusty and therefore	$H_0: \beta_{age} \geq 0$ $H_1: \beta_{age} < 0$

	they are not as attractive to employers anymore. So the probability decreases as the age increases.	
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).
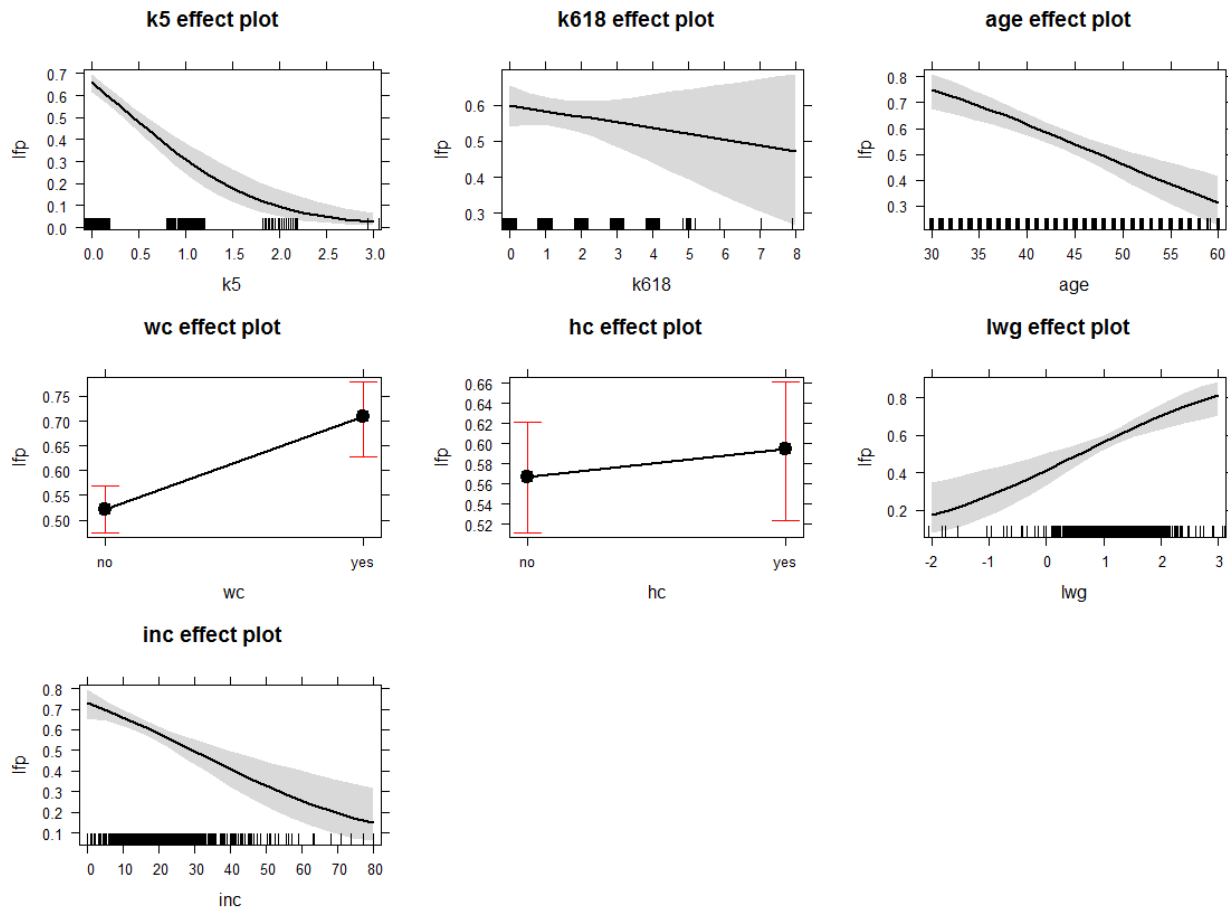
Comments: The estimated probability for the observation can be expressed by $\widehat{\pi_i} = \frac{1}{1+\exp\left(-x_i^t * b\right)}$. So the interpretation in terms of probability is not the same as that of the linear regression. The slope of the logistic curve for a given probability $\widehat{\pi_i}$ is $b_k \cdot \widehat{\pi_i} \cdot (1 - \widehat{\pi_i})$. 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 year old, the probability she works is 60%. When the number of 6 to 18 year 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 two 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 $H_0: \beta_{k618} = \beta_{hcyes} = 0$. 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"))
```
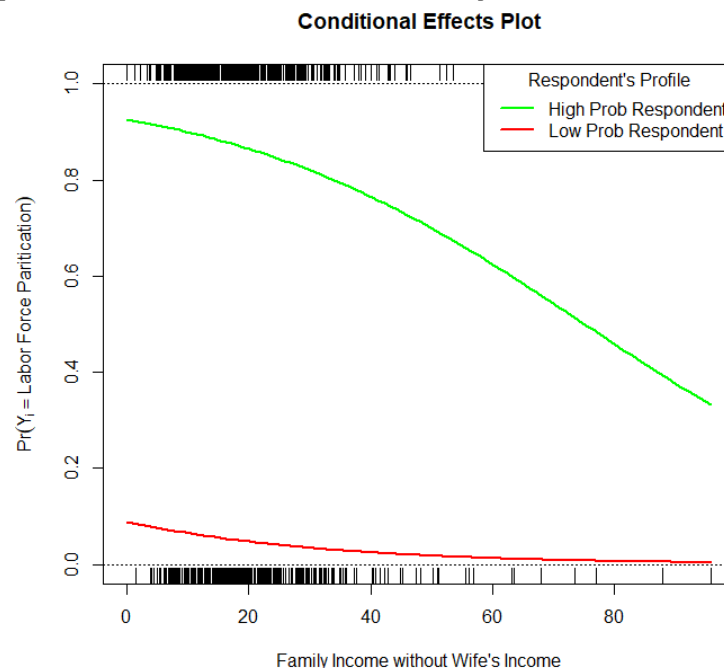

Conditional Effects Plot

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: 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**.

```
> glm01 <- glm(L_WM_P2_CN~URBRUR+RAD_MD+I(RAD_MD^2)+TOBACCO, offset=log(L_WM_P2_EX),
+              family=poisson, data=cancer)
> summary(glm01)

Call:
glm(formula = L_WM_P2_CN ~ URBRUR + RAD_MD + I(RAD_MD^2) + TOBACCO,
    family = poisson, data = cancer, offset = log(L_WM_P2_EX))

Deviance Residuals:
     Min       1Q    Median       3Q       Max
-23.3141   -3.7661   -0.9355   2.1832   26.2712

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.2038740  0.0087113  -23.40   <2e-16 ***
URBRURurban  0.2904292  0.0028756  101.00   <2e-16 ***
RAD_MD      -0.2329147  0.0064603  -36.05   <2e-16 ***
I(RAD_MD^2)  0.0340182  0.0019232   17.69   <2e-16 ***
TOBACCO      0.0093650  0.0003904   23.99   <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: 28494  on 507  degrees of freedom
Residual deviance: 12765  on 503  degrees of freedom
AIC: 16978

Number of Fisher Scoring iterations: 4
```
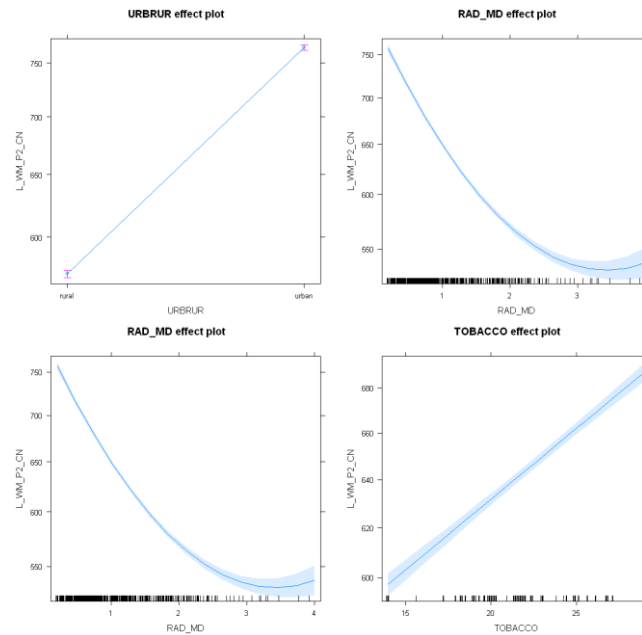
```
> plot(allEffects(glm01))
```


The expected number of cases, given the observed age-distribution, is given by the variable L_WM_P2_EX. It will become the offset. The log-transformation needs to be applied on this offset to match the link function of the Poisson model.

Parameter interpretation:

- Urban areas have higher number of cases than expected.
- Counterintuitively low levels of radon exposure lead to a number of cases above the offset. The quadratic relationship only changes direction above a radon level of 3.3. This is known as the Cohen's linear no-threshold controversy (see http://www.radonleaders.org/sites/default/files/Bond.pdf, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3315166/ and https://en.wikipedia.org/wiki/Bernard_Cohen_(physicist))
- Tobacco consumption increases the risk for lung cancer.

The deviance dropped by over 50%.

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 points]

Select as independent variables **~URBRUR+RAD_MD+I(RAD_MD^2)+TOBACCO**.

```
> ## Logistic regression
> cancer$LWMP2RT <- cancer$L_WM_P2_RT/100000
> glm02 <- glm(LWMP2RT~URBRUR+RAD_MD+I(RAD_MD^2)+TOBACCO,
+              family=binomial, weights=POPATRISK1982, data=cancer)
> summary(glm02)

Call:
glm(formula = LWMP2RT ~ URBRUR + RAD_MD + I(RAD_MD^2) + TOBACCO,
    family = binomial, data = cancer, weights = POPATRISK1982)

Deviance Residuals:
    Min      1Q   Median      3Q      Max
-3.8178  -0.3578   0.0773   0.4937   2.5034

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -7.622966   0.050541 -150.829  < 2e-16 ***
URBRURurban  0.010744   0.015993    0.672 0.501726
RAD_MD      -0.213245   0.036447   -5.851 4.89e-09 ***
I(RAD_MD^2)  0.039250   0.011055    3.551 0.000384 ***
TOBACCO      0.023262   0.002222   10.471  < 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: 423.94  on 507  degrees of freedom
Residual deviance: 279.09  on 503  degrees of freedom
AIC: 2785.7

Number of Fisher Scoring iterations: 4
```
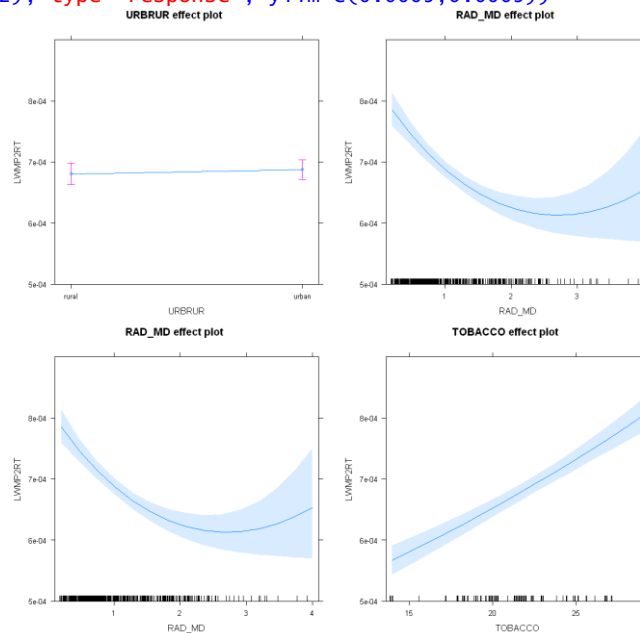
Dividing the per 100,000 mortality rate by 100,000 gives probability per year of dying of lung cancer.

```
> plot(allEffects(glm02), type="response", ylim=c(0.0005,0.0009))
```


Compared to the Poisson model evaluating lung cancer the logistic model is not as significant (see standard errors and width of the confidence intervals) anymore. The urban-rural effect is even insignificant now. The deviance dropped only by 35%.

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]

See task 6. Declare `type="response"` to obtain predictions on the probability scale rather than the logit-scale.

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]

```
> glm02 <- glm(LWMP2RT~URBRUR+RAD_MD+I(RAD_MD^2)+TOBACCO,
+              family=quasibinomial, weights=POPATRISK1982, data=cancer)
> summary(glm02)

Call:
glm(formula = LWMP2RT ~ URBRUR + RAD_MD + I(RAD_MD^2) + TOBACCO,
    family = quasibinomial, data = cancer, weights = POPATRISK1982)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-3.8178 -0.3578  0.0773  0.4937  2.5034

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -7.622966   0.037490 -203.335  < 2e-16 ***
URBRURurban  0.010744   0.011863    0.906    0.366
RAD_MD      -0.213245   0.027035   -7.888 1.93e-14 ***
I(RAD_MD^2)  0.039250   0.008200    4.787 2.23e-06 ***
TOBACCO      0.023262   0.001648   14.116  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.5502278)

    Null deviance: 423.94  on 507  degrees of freedom
Residual deviance: 279.09  on 503  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```
This model indicates that we observe under-dispersion of 0.55. Note: If as population at risk `MalePop=POP1980/2` instead of `POPATRISK1982` would have been used, over-dispersion would be around 4. For either over or under-dispersion the regression coefficients are not changing, however, depending on the degree of dispersion the standard errors and thus the t-values are changing. For under-dispersion the standard errors shrink, whereas, for over-dispersion they increase.

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 8: Estimate the basic gravity model $E\left(m_{ij}\right) = \beta_0 \cdot p_i^{\beta_1} \cdot p_j^{\beta_2} \cdot d_{ij}^{\beta_3}$ with Poisson regression and transforming the right-hand-side of the equation into a linear equation in the unknown regression coefficients $\beta_0, \beta_1, \beta_2$ and β_3 by applying the logarithm. [1 point]

```
> mig1.glm <- glm(MIJ~log(PI)+log(PJ)+log(DIJ), data=upfing, family=poisson,
subset=(I!=J))
> summary(mig1.glm)

Call:
glm(formula = MIJ ~ log(PI) + log(PJ) + log(DIJ), family = poisson,
    data = upfing, subset = (I != J))

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 ***
log(PI)      0.7142001  0.0009334   765.2   <2e-16 ***
log(PJ)      0.5838710  0.0009029   646.7   <2e-16 ***
log(DIJ)    -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
```

Task 9: 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 points]

Comment: The regression coefficients of origin and destination populations are positive whereas the regression coefficient of distance is negative. Consequently, with increasing population sizes the migration flow between an origin-destination pair increases. Because the magnitude of the regression coefficient associated **log(PI)** is larger than that **log(PJ),** large origin populations more emissive compare to the attractions based on large destination population sizes. The inter-regional distance becomes a deterrence of migration. The further distance between two regions, the less migration flows are expected.