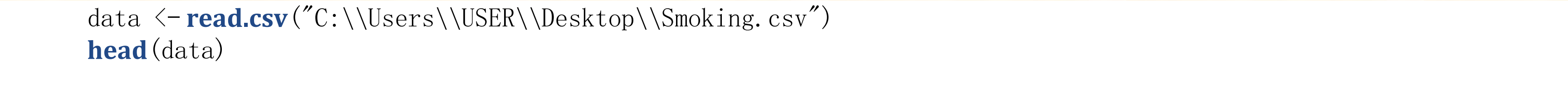
A Project by
Patha Pratim Das

- ❑ **Logistic Regression:** Logistic regression is used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature. That means Logistic regression is usually used for Binary classification problems. Binary Classification refers to predicting the output variable that is discrete in two classes. A few examples of Binary classification are Yes/No, Pass/Fail, Win/lose, Cancerous/Non-cancerous, etc. Although it is said Logistic regression is used for Binary Classification, it can be extended to solve multiclass classification problems.

- ❑ **Data Description:** We have data on 400 individuals of a city in America. Here data is available on chance of having cancer , Smoking level, Trouble in Breathing and Average smoking time per week(in hrs) for each individuals. We wants to know the factors that influence the decision of whether a person have a high/moderate/low chance of having cancer and apply multinomial logistic regression model.

The given data is as below,



```
data <-read.csv(”C:\\Users\\USER\\Desktop\\Smoking.csv”)
head(data)
```

##		Chance of Cancer	Smoke Level	Trouble in Breathing	Avg. smoking time per week(in hrs)
##	1	High	0	0	3.26
##	2	High	1	1	3.21
##	3	High	1	1	3.94
##	4	Moderate	1	0	2.81
##	5	Low	0	1	2.53
##	6	Low	0	1	2.59

## Summery of the data

summary(data)

##	Chance of Cancer	Smoke Level	Trouble in Breathing	Avg. smoking time per week(in hrs)
##	High :220	Min. :0.0000	Min. :0.0000	Min. :1.900
##	Moderate :140	1st Qu. :0.0000	1st Qu. :0.0000	1st Qu. :2.720
##	Low : 40	Median :0.0000	Median :0.0000	Median :2.990
##		Mean :0.1575	Mean :0.1425	Mean :2.999
##		3rd Qu. :0.0000	3rd Qu.:0.0000	3rd Qu. :3.270
##		Max. :1.0000	Max. :1.0000	Max. :4.000

## Dimension of the data

dim(data)

```
##    [1]     400        4
```

```
with(data, table(Smoke Level, Chance of Cancer ))
```

```
##Chance of Cancer

## Smoke Level    High       Moderate     Low
##   0             200         110         27
     1             20          30          13
```

```
with(data, table(Trouble in Breathing, Chance of Cancer))
```

```
##Chance of Cancer
## Trouble in Breathing     High     Moderate     Low
 ##            0             189       124        30
 ##            1              31        16        10
```

```
with(data, do.call(rbind, tapply(Avg smoking time per week(in hrs), Chance of  Cancer,
function(x) c(M = mean(x), SD = sd(x)))))
```

```
##            M          SD
## High       2.952136   0.4035940
## Moderate   3.030071   0.3893446
## Low        3.147250   0.3560322
```

Note :

Now we will proceed for analysis with Multinomial logistic regression.

❑ Multinomial logistic regression:

First, we need to choose the level of our outcome that we wish to use as our baseline and specify this in the relevel function. Then, we run our model using multinom. The multinom package does not include p-value calculation for the regression coefficients, so we calculate p-values using Wald tests (here z-tests).

```
library(nnet)

data$Level2 <- relevel(data$Level, ref = ”High”)
test <- multinom(Chance of Cancer~ Smoke Level + Trouble in breathing + Avg smoking time per week(in hrs) ,data = data)
```

```
## # weights: 15 (8 variable)
## initial  value 439.444915
## iter  10 value 357.012275
## final value 356.996982
## converged
```

```
## Summery  of the  test

summary(test)
```

```
## Call:
## multinom(formula = Chance of Cancer ~ Smoke Level + Trouble      in breathing + Avg smoking time per week(in hrs),
      data = data)

## Coefficients:
##                (Intercept)      Smoke Level     Trouble in Breathing     Avg. smoking time per week(in hrs)
## Moderate       -1.878955         0.9516492          -0.4188168                          0.4487486
## Low            -4.847763         1.3741555           0.360066                           0.9240376

## Std. Errors:
##                (Intercept)      Smoke Level      Trouble in Breathing     Avg. smoking time per week(in hrs)
## Moderate         0.863786          0.3170625           0.3432944                        0.2902058
 ## Low             1.449023          0.4221675           0.4434658                        0.4741715
##
##            Residual Deviance:               713.994
##            AIC: 729.994
```

```
z <- summary(test)$coefficients/summary(test)$standard.errors
z
```

```
##              (Intercept)     Smoke Level     Trouble in Breathing     Avg. smoking time per week(in hrs)

## Moderate     -2.175255        3.001456          -1.2199931                      1.546312
## Low          -3.345539        3.255001            0.81193664                    1.948741
```

```
# 2-tailed  z  test

p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
```

```
##              (Intercept)     Smoke Level     Trouble in Breathing     Avg. smoking time per week(in hrs)

# Moderate      0.0296110241    0.002686917        0.2224675                    0.12202928
# Low           0.0008212272    0.001133920        0.4168280                    0.05132636
```

```
exp(coef(test))
```

```
##              (Intercept)     Smoke Level     Trouble in Breathing     Avg. smoking time per week(in hrs)

# Moderate      0.152749700      2.589977           0.6578247                     1.566351
 # Low          0.007845912      3.951738           1.4334241                     2.519442
```

❑ Fitted values :

Finally, the predicted probabilities for each of our outcome levels are calculated below.

```
head(pp <- fitted(test))
```

```
##         High     Moderate        Low
## 1  0.5496900    0.3626072   0.08770283
## 2  0.3055654    0.5104748   0.18395981
## 3  0.2370200    0.3614384   0.40154168
## 4  0.6081586    0.3278200   0.06402147
## 5  0.6424002    0.3053903   0.05220947
## 6  0.6923306    0.2224163   0.08525308
```

Response Variable: Here the response variable is how much likely a person is likely to have Cancer duue to smoking, (say y).

Here the response variable y has 3 categories,

y=0; High chance to have Cancer

y=1; Moderate chance to have Cancer

y=2; Low chance to have Cancer

Predictors: Here number of predictor is 3.

i. Smoking Level: It is a categorical variable say X_1.

$$X_1 : \begin{cases} 1; & \text{if the person smokes at least 10 cigarettes per day.} \\ 0; & \text{if the person smokes less than 10 cigarettes per day.} \end{cases}$$

ii. Trouble in Breathing: It is also a categorical variable X_2.

$$X_2 : \begin{cases} 1; & \text{if the person has trouble in breathing after smoking.} \\ 0; & \text{if the person don't have trouble in breathing after smoking.} \end{cases}$$

iii. Average Smoking time per day (in hrs.): It is a continuous variable; let us denote it by X_3 .

Here our reference category of response variable is y=0 i.e. high chance to have Cancer.

Hence,

$$\log\left[\frac{P(y=i/\underline{X})}{P(y=0/\underline{X})}\right] = \beta_{i0} + \beta_{i1}I(X_1=1) + \beta_{i2}I(X_2=1) + \beta_{i3}X_3 + \varepsilon; \quad (i=1,2)$$

From the summary of the test we can observe that,

$$\hat{\beta}_{10} = -1.8789, \hat{\beta}_{11} = 0.951, \hat{\beta}_{12} = -0.42, \hat{\beta}_{13} = 0.45$$

$$\hat{\beta}_{20} = -4.85, \hat{\beta}_{21} = 1.37, \hat{\beta}_{22} = 0.36, \hat{\beta}_{23} = 0.924$$

Interpretations: $\hat{\beta}_{10}$ & $\hat{\beta}_{20}$ are intercept terms.

1. $\underline{e^{\hat{\beta}_{11}} = e^{0.951} = 2.5882966623}$: It means how the odds of chance of having Cancer being Moderate $(i=1)$ (relative to chance of having Cancer being High ($i=0$)). Changes for a person smoking at least 10 cigarettes $(i=1)$ per day versus a person smoking less than 10 cigarettes per day; holding trouble in Breathing & average time of smoking per day as constant. Here the odds of chance of having Cancer being Moderate (relative to chance of having Cancer being High) is 2.5882966623 times for a person smoking at least 10 cigarettes per day holding trouble in Breathing average time of smoking per day as constant.
2. $\underline{e^{\hat{\beta}_{12}} = e^{-0.42} = 0.6570468198}$: It says similarly that, the odds of chance of having Cancer being Moderate ($i=1$)(relative to chance of having Cancer being High ($i=0$)) for a person having trouble in Breathing after smoking is 0.6570468198 times than a person not having trouble in Breathing after smoking provided smoking level & average smoking time per day (in hours) is constant or fixed.
3. $\underline{\hat{\beta}_{13} = 0.45}$: The estimated value of the parameter is 0.45. Hence keeping all other variables constant if we increase X3 or average smoking time per day (in hours) by 1 amount,

$$\log\left(\frac{Chance\,of\;having\;Cancer\;Moderately}{Chance\,of\;having\;Cancer\;Highly}\right)$$ increases by an amount 0.45.

$\hat{\beta}_{21} = 1.37$

4. $\underline{\therefore e^{\hat{\beta}_{21}} = 3.9353506955}$ **:** The odds of chance of having Cancer being Low $(i = 2)$ (relative to chance of having Cancer being High $(i = 0)$)for a person smoking at least 10 cigarettes per day is 3.9353506955 times than a person smoking less than 10 cigarettes per day holding trouble in Breathing average time of smoking per day as constant.

5. $\underline{\hat{\beta}_{22} = 0.36; e^{\hat{\beta}_{22}} = e^{0.36} = 1.4333294146}$**:** The odds of chance of having Cancer being Low $(i = 2)$ (relative to chance of having Cancer being High$(i = 0)$) for a person having trouble in Breathing after smoking is 1.4333294146 times than a person who does not have problem in Breathing holding smoking level & average time of smoking per day as constant.

6.$\underline{\hat{\beta}_{23} = 0.924}$ **:** The estimated value of the parameter is 0.924. Hence, keeping all other variables fixed if we increase X_3 or average smoking time per day (in hours) by 1 amount;

$\log\left(\frac{\text{Chance of having Cancer being Low}}{\text{Chance of having Cancer being High}}\right)$ increases by an amount 0.924.

❑**<u>For Testing the parameters are actually=0 or not</u>**

By the 2-tailed z-test result, we can see that

A. **<u>For chance of having Cancer being Moderate relative to having Cancer being High</u>**:

1. **The P-value for** $H_0 : \beta_{11} = 0 \ \ vs \ \ H_1 : \beta_{11} \neq 0$ **is** $0.002686917 < 0.05$ **:** Hence we will reject H_0 in favour of H_1 ; Hence in the light of the given data there is significant effect of the smoking level $(X_1 = 1)$. On the response for the category (y=1) (the chance of having Cancer being Moderate) relative to the category y=0 (i.e. chance of having Cancer being High).
2. **The P-value for testing (** $H_0 : \beta_{12} = 0 \ \ vs \ \ H_1 : \beta_{12} \neq 0$ **) is** $0.2224675 > 0.05$**:** Hence we accept H_0 and in the light of the given data there is no such significant effect of the variable, trouble in Breathing $(X_2 = 1)$ on the response for the category (y=1) (i.e. chance of having Cancer being Moderate) relative to the category y=0 (i.e. chance of having Cancer being High).
3. **The P-value of testing (** $H_0 : \beta_{13} = 0 \ \ vs \ \ H_1 : \beta_{13} \neq 0$ **) is** $0.12202928 > 0.05$ **:** Hence here also we accept H_0 in favour of H_1 and there is no significant effect of the variable X_3 (i.e. average smoking time per day (in hours)) on the response for the category (y=1) (i.e. chance of having Cancer being Moderate) relative to the category y=0 (i.e. chance of having Cancer being High).

B. For chance of having Cancer being Low relative to that chance of being High:

4. The P-value for testing $H_0 : \beta_{21} = 0 \;\; vs \;\; H_1 : \beta_{21} \neq 0$ **is** $0.001133920 < 0.05$ **:** Hence H_0 is rejected in favour of H_1. So, we can conclude there is significant effect of the variable smoke level $(X_1 = 0)$ on the response for the category (y=2) (i.e. chance of having Cancer being Low) relative to the category y=0 (i.e. chance of having Cancer being High).

5. The P-value for testing $H_0 : \beta_{22} = 0 \;\; vs \;\; H_1 : \beta_{22} \neq 0$ **is** $0.4168280 > 0.05$**:** Hence H_0 is accepted in favour of H_1. So, we can conclude that there is no such significant effect of the variable 'Trouble in Breathing' on the response for the category (y=2) (i.e. chance of having Cancer being low) relative to the category y=0 (i.e. chance of having Cancer being High).

6. The P-value for testing $H_0 : \beta_{23} = 0 \;\; vs \;\; H_1 : \beta_{23} \neq 0$ **is** 0.05132636 **which is slightly more than** 0.05 **:** Hence we accept H_0 in favour of H_1. So, in the light of the given data, we can conclude that there is no significant effect of the variable X_3 (i.e. average smoking time per day (in hours)) on the response for the category (y=2) (i.e. chance of having Cancer being Low) relative to the category y=0 (i.e. chance of having Cancer being High).

❑ **Conclusion:** Hence what ever the category of the response y be (given y=0) only X_1 i.e. smoking level has significant effect on that response.

Acknowledgement and Bibliography:

I would like to express my heartfelt gratitude to **Professor Sudheesh Kattumannil** for his guidance and teaching, without which the project work would not have been a success.

Further, I have taken the help of the following books:

- ***Categorical Data Analysis*** by ***Alan Agresti.***
- ***Modern Statistical Methods for Astronomy with R applications*** by ***Eric D. Feigelson*** and ***G. Jogesh Babu***.

THANK YOU