# Assignment2:Logistic regression

Me!

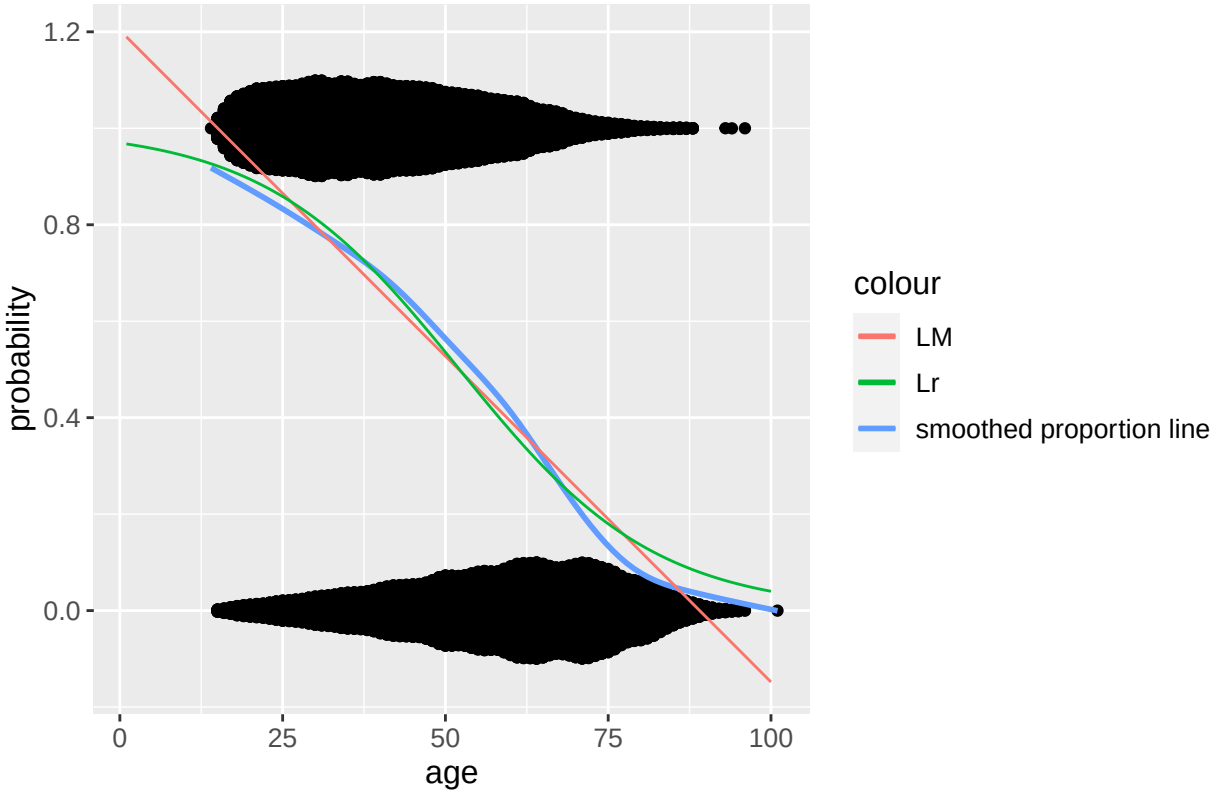
2/11/2021

## Problem 1: Age and internet use:

The internet use variable is binary variable (freqnet) measured with 1 for frequent use , while the age variable (agea) is measured in years. the internet use has become a social necessity. it is used by all ages for network and information seeing especially by young ages who are living in a contemporary world with new technologies so I hypothesize that there is a significant relationship between age and internet use. older ages (+ 60) are less likely to adopt internet comparing to other ages.

Figure 1 shows the relationship between age and frequent internet usage in Europe in the raw data, where the respondents grouped by age. it is clear from the slope that there is a negative relationship between age and frequent internet use. As the age increases the usage for internet decreases and this supports my hypothesis. To predict the internet use as a function of age, I run a linear probability model (red line) and logistic model (green line) with one binary outcome (freqnet) and one explanatory variable (agea) and included it in figure. By comparing the lines in graph, we can see that the blue line referring to logistic is closer to the true data (blue) but with longer ends while the green is straight from top to down. the results are shown in Table1.

Figure1: frequent use of internet by age



In the linear probability column (1)/Table1, the constant shows 101 % probability when age variable is 0 to be a frequent internet user and with each one increase in age the probability decreases by 13,5 %. also shows 24 % variance in data while the logistic probability column (2) is hard to interpret in this case and the age coefficient need to be exponentiated: exp(coef(m2)). model 2 implies that Older people have nearly 0.5 times higher odds of using internet in a dialy basis or several times per week compared to those in younger ages. while the outcomes of both models a little bit different but it is still support my hypothesis.

## Problem 2: Education and internet use:

I think that the more educated individuals are the more likely to use internet in a daily basis concerning their studies, researches and information seeking so I think there is a positive relationship between these two variables. So I hypothesize that individuals with high level are more likely to use internet in a daily or weekly basis.

The variable related to education (eisced) is ordinal measured by a scale from 1-7 where 7 refers to MA level. Table 2 shows the percentage of frequent and non frequent uses of internet. we see that individuals with most education level like BA and MA level have higher share of frequent internet user and this supports my hypothesis. chi squared refers to a significant relationship.

I run linear probability and logistic regression models for both variables freqnet and eisced as a factor. we can see the results of these models in table3. in col(1) that refers to the linear probability model, we see that there is a significant relationsship between education variables in all levels and frequent use of internet. also we notice that the higher level of education, the more likely to be frequent internet user, for example, for MA -BA level, the probability to be frequent internet user is between 62-66%. these results support my hypothesis.

As I calculate predictions for both models (figure 2), I find results are similar i.e people with lower level of education have lower probabilty 16% to use internet in frequent basis and we can see that findings and numbers are compatible with our findings in table1

```
##     model               eisced          pred
##  Length:14          Min.   :1.00   Min.   :0.1612
##  Class :character   1st Qu.:2.25   1st Qu.:0.4328
##  Mode  :character   Median :4.00   Median :0.5517
##                     Mean   :4.00   Mean   :0.5485
##                     3rd Qu.:5.75   3rd Qu.:0.7390
##                     Max.   :7.00   Max.   :0.8277
```

#Problem 3: Determinants of internet use.

There are various socio-economic factors that may impact the use of internet for example, demographic variables such as (gender= gndr) if the user is male or female may differentiate in the level of frequent using of internet. Educated female individuals who have other responsibilities at house or with kids are less likely to be frequent internet users. Also another variable is the household income decile (hinctnta). People who have higher income decile are more likely to be frequent internet users as they will be able to pay invoices and pay internet packages or to have an open net connection as illustrated in figure 4. Third variable is the iscomajor (Occupational groups). we find a strong connection between the level of education and the type of work.(iscomajor) is an ordinal variable measured on scale from 0 to 10 where 10 is Never worked.

To examine the relationship between these variables, I run linear probability and logistic models using two variables (hinctnta) & (gndr) as shown on table 3-col(3) we find significant relationship between these variable and the frequent use of internet. also we find a slight decrease in the probabilities for educated people to be frequent internet users but still reasonable probabilities. for example worker with BA level of education has probability 52% to be frequent internet user. also we find a negative significant relationship between gender(female) and freqnet. it is less likely for females to be frequent internet users by 4,5% comparing to

males. it is hard to estimate the results of logistic regression as shown in col(4) but by producing predictions we can have a better image: for people from both genders with low levels of education and moderate income have low probability 20% of being frequent internet users. All results still from both models still support my hypothesis.

Figure 5 illustrates the relationship between income decile and level of education with the frequent use of internet divided by gender. we see there is non linear relationship between these variables (red line). we can see that higher level of income connected with being a frequent user but with more probability for males. also we see the green line with represent the logistic model is in line with the red one but with longer tail at the beginning. we see the results in figure 4 is similar to our findings in table 1.

```
##
## Table1: LPM and LR regression models of frequent Internet use by age
## ================================================
##                          Dependent variable:
##                 ---------------------------------
##                                freqnet
##                          OLS              logistic
##                          (1)                 (2)
## ------------------------------------------------
## Age/years            -0.135***           -0.664***
##                       (0.001)             (0.008)
##
## Constant             1.014***            2.534***
##                       (0.005)             (0.030)
##
## ------------------------------------------------
## Observations          37,825              37,825
## Adjusted R2           0.242
## Log Likelihood                          -21,079.260
## ================================================
## Note:                        *p<0.1; **p<0.05; ***p<0.01
##                 Source: European Social Survey (ESS)2010
```

Table 1: Table2: Frequent Internet Use by level of education

| | Internet use | | |
|---|---|---|---|
| Level of Education | Non Frequent User | Frequent Internet User | Total |
| 1 | 83.9% (3455) | 16.1% (664) | 100.0% (4119) |
| 2 | 58.7% (3684) | 41.3% (2587) | 100.0% (6271) |
| 3 | 50.6% (2964) | 49.4% (2891) | 100.0% (5855) |
| 4 | 44.8% (3830) | 55.2% (4713) | 100.0% (8543) |
| 5 | 38.9% (1938) | 61.1% (3045) | 100.0% (4983) |
| 6 | 17.2% (619) | 82.8% (2974) | 100.0% (3593) |
| 7 | 21.8% (974) | 78.2% (3487) | 100.0% (4461) |
| Total | 46.2% (17464) | 53.8% (20361) | 100.0% (37825) |

*Note:* Chi-squared = 5189, df = 6, p-value < 0.001 Source: European Social Survey (ESS)2010

```
##
## Table3: LPM and LR regression models of frequent Internet use by the level of education
```

| | Dependent variable: | | | |
|---|---|---|---|---|
| | freqnet | | | |
| | OLS | logistic | OLS | logistic |
| | (1) | (2) | (3) | (4) |
| lower secondary | 0.251*** | 1.296*** | 0.201*** | 1.119*** |
| | (0.009) | (0.050) | (0.009) | (0.051) |
| lower tier upper secondary | 0.333*** | 1.624*** | 0.239*** | 1.263*** |
| | (0.009) | (0.050) | (0.009) | (0.051) |
| upper tier upper secondary | 0.390*** | 1.857*** | 0.297*** | 1.515*** |
| | (0.009) | (0.048) | (0.009) | (0.049) |
| advanced vocational, sub-degree | 0.450*** | 2.101*** | 0.333*** | 1.669*** |
| | (0.010) | (0.051) | (0.010) | (0.053) |
| BA level | 0.667*** | 3.219*** | 0.516*** | 2.705*** |
| | (0.011) | (0.061) | (0.011) | (0.063) |
| MA level | 0.620*** | 2.925*** | 0.442*** | 2.261*** |
| | (0.010) | (0.056) | (0.010) | (0.058) |
| hinctnta | | | 0.047*** | 0.224*** |
| | | | (0.001) | (0.005) |
| Female | | | -0.045*** | -0.231*** |
| | | | (0.005) | (0.023) |
| Constant | 0.161*** | -1.649*** | 0.041*** | -2.297*** |
| | (0.007) | (0.042) | (0.008) | (0.049) |
| Observations | 37,825 | 37,825 | 37,825 | 37,825 |
| Adjusted R2 | 0.137 | | 0.200 | |
| Log Likelihood | | -23,325.230 | | -21,933.210 |
| Note: | | | *p<0.1; **p<0.05; ***p<0.01 | |
| | Source: European Social Survey (ESS)2010 | | | |

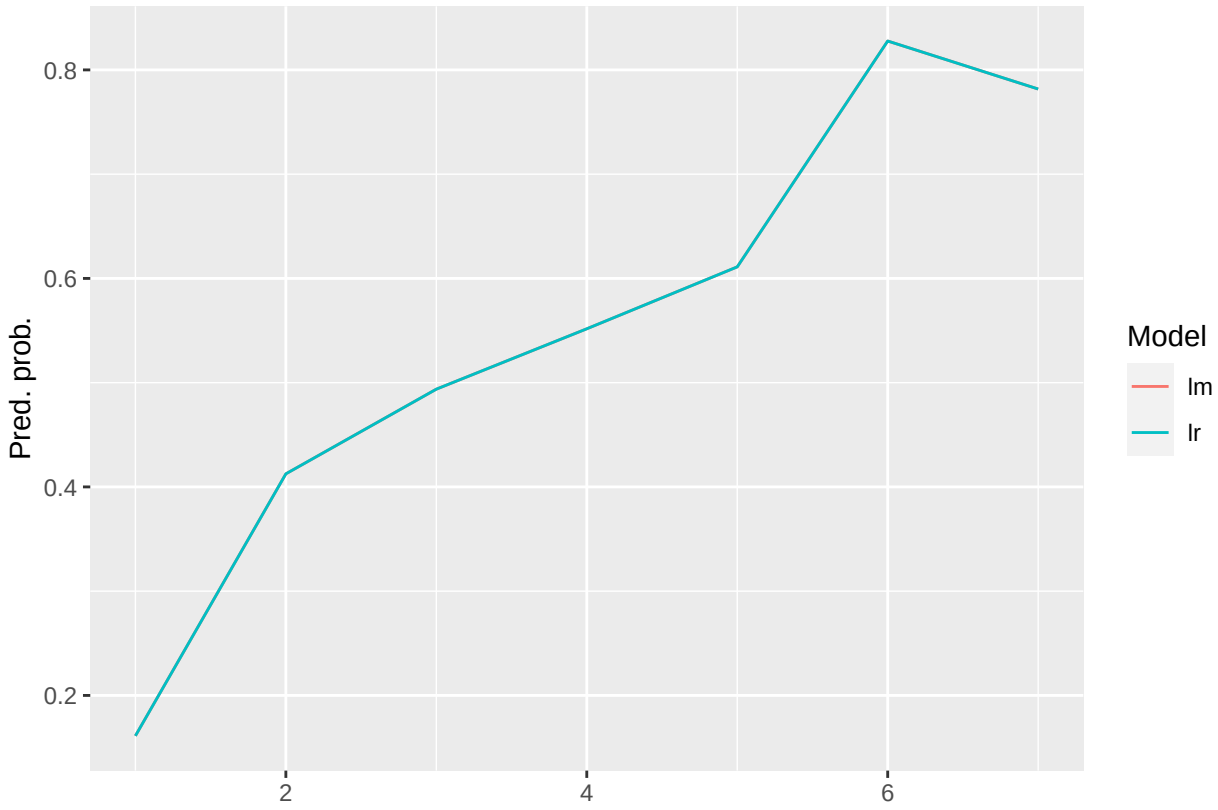



Figure2: Level of Education

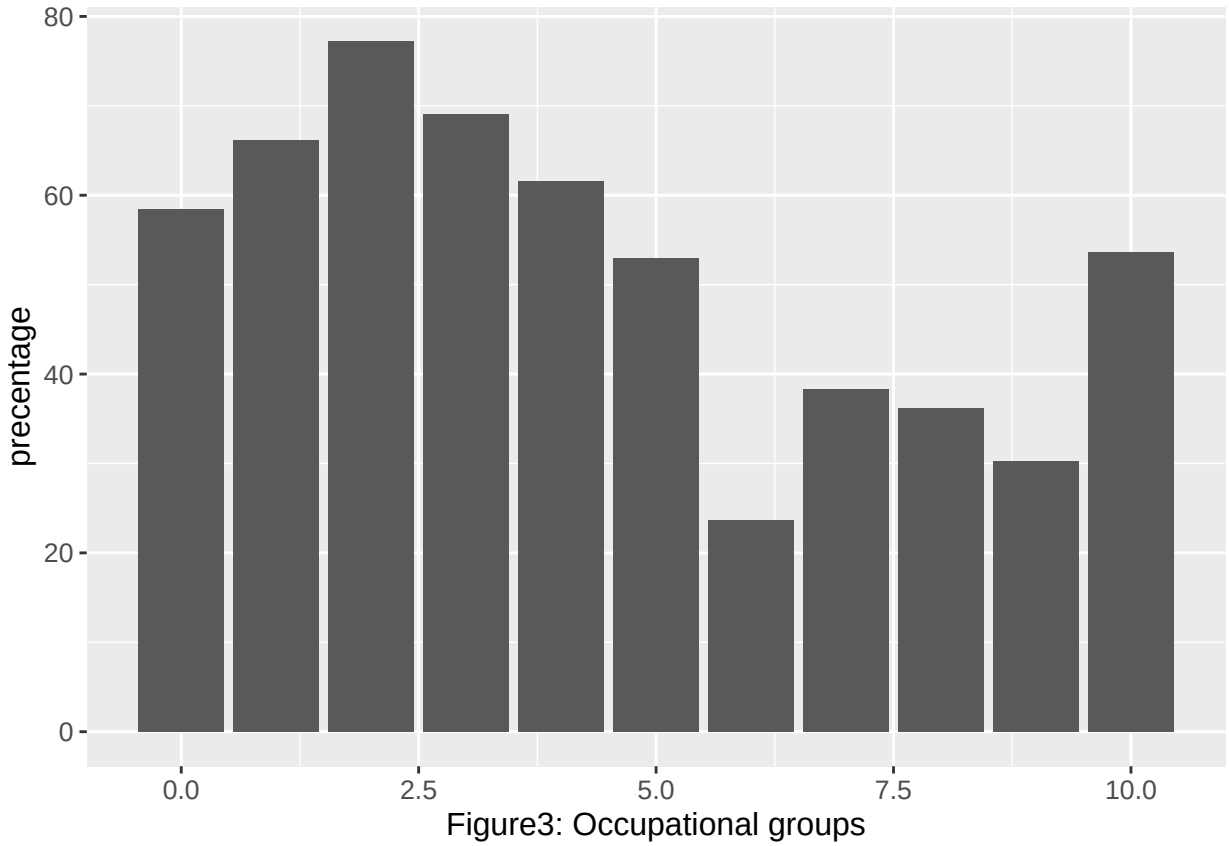



Figure3: Occupational groups

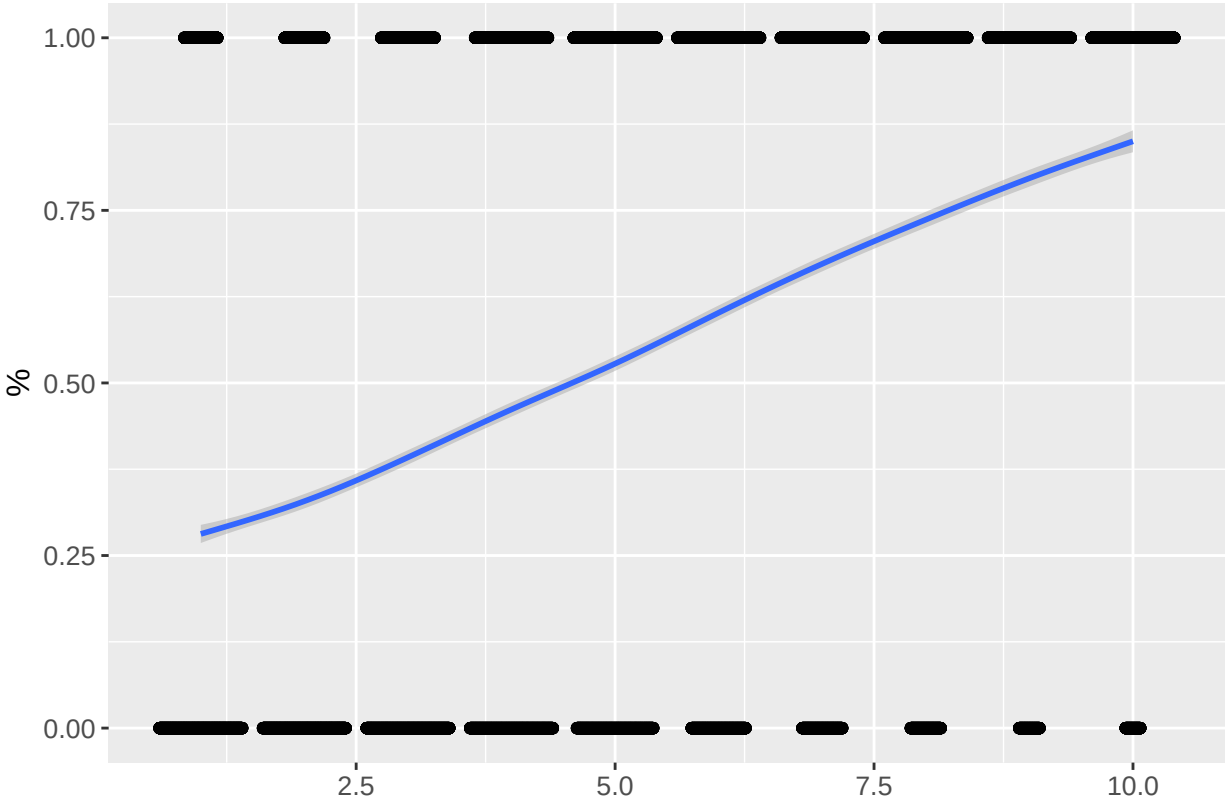



Figure4: household income decile

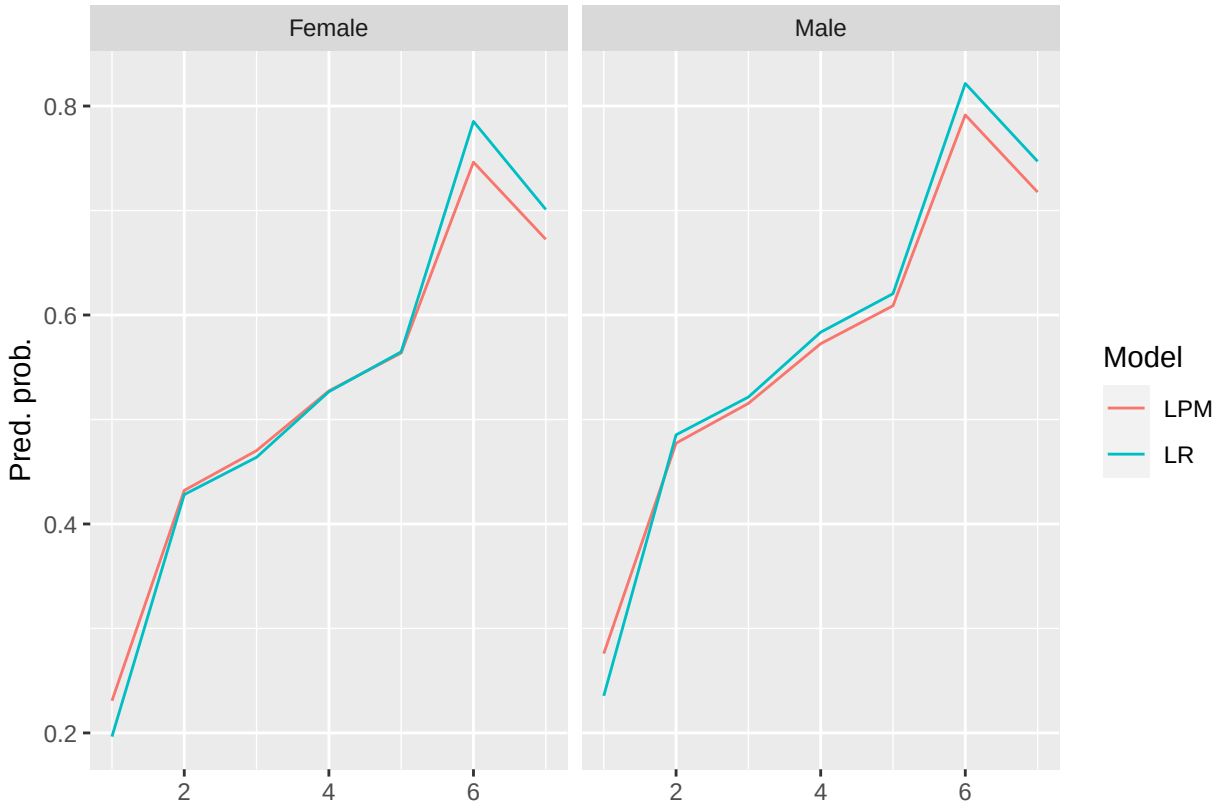



Figure5: Frequnet internet users by income decile