Financial Econometrics in R/Python
Group Assignment 2

Group 3
The Business School, Imperial College London
Rakesh Bali (CID: 02478023)
Rui Moreno (CID: 02532521)
Siyu Liu (CID: 02479121)
Zhiyu Chen (CID: 02517659)
Zhuofan Chen (CID: 02151338)

26-11-2023

Contents

Preparation of R and Data

Loading R packages

Load the packages required.

```
library(quantmod)
library(dplyr)
library(readxl)
library(moments)
library(lmtest)
library(sandwich)
library(MASS)
library(ggplot2)
library(knitr)
library(margins)
library(jtools)
library(e1071) # For Naive Bayesian Classifier in (h)
library(tree) # For dicision tree in (h)
library(randomForest) # For random Forest in (h)
library(class) # For k-NN in (h)
```

Loading Data

Load `growthdata.xlsx` into a data frame.

```
employment_data <- read_excel("employment_08_09.xlsx")
# clean data
data <- na.omit(employment_data)
```

Question a

What fraction of workers in the sample were employed in April 2009? Use your answer to compute a 95% confidence interval for the probability that a worker was employed in April 2009, conditional on being employed in April 2008

```
number_of_workers_employed_2008 <- length(employment_data$employed)

workers_employed_2009 <- employment_data %>% filter(employed == 1)

number_of_workers_employed_2009 <-
  length(workers_employed_2009$employed)

fraction_of_workers_employed_2009 <-
  round(number_of_workers_employed_2009 / number_of_workers_employed_2008,
        4)

standard_error_fraction_employed_2009 <-
  sqrt((
    fraction_of_workers_employed_2009 * (1 - fraction_of_workers_employed_2009)
  ) / number_of_workers_employed_2008)

confidence_level <- 0.95
margin_of_error <-
  qnorm((1 + confidence_level) / 2) * standard_error_fraction_employed_2009

confidence_interval <-
  c(
    fraction_of_workers_employed_2009 - margin_of_error,
    fraction_of_workers_employed_2009 + margin_of_error
  )

print(confidence_interval)
```

```
## [1] 0.8667041 0.8842959
```

In April 2009, 87.55% of workers were employed. On a 95% confidence level, the confidence interval is computed for the probability that a worker was employed in April 2009 to be [0.867, 0.884].

Question b

Regress Employed on Age and Age2 , using a linear probability model

```
# required linear probability model
lpm_b <-  lm(employed ~ age + I(`age`^2), data = data)
summary(lpm_b)
```

```
##
## Call:
## lm(formula = employed ~ age + I(age^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91854  0.08342  0.09933  0.13867  0.28625
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.218e-01  5.705e-02   5.640  1.8e-08 ***
## age          2.746e-02  2.894e-03   9.491  < 2e-16 ***
## I(age^2)    -3.159e-04  3.479e-05  -9.080  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3274 on 4770 degrees of freedom
## Multiple R-squared:  0.01956,    Adjusted R-squared:  0.01915
## F-statistic: 47.59 on 2 and 4770 DF,  p-value: < 2.2e-16
```

Question b-i)

Based on this regression, was the age a statistically significant determinant of employment in April 2009

```
# t test - statistical significance of coefficients
coeftest(lpm_b, vcov = vcovHC(lpm_b), type = "HC1")
```

```
##
## t test of coefficients:
##
##               Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  3.2177e-01  6.9629e-02  4.6212 3.916e-06 ***
## age          2.7463e-02  3.4474e-03  7.9662 2.030e-15 ***
## I(age^2)    -3.1592e-04  4.1022e-05 -7.7014 1.626e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *age* variable is a statistically significant determinant of the 2009 April Employment rate. The linear probability model with *age* and squared *age* regressed on *employed* gives a relationship of

$$employed = 0.3218 + 0.02746 \times age - 0.0003159 \times age^2$$

.

The t-value of the *age* variable at 9.491 is significant enough to reject the null hypothesis at a significance level of 1%, i.e. there is no statistical evidence that the coefficient of *age* in the probability model is different from zero. The p-value is less than

$$2 \times 10^{-16}$$

, which is considerably smaller than the significance level, further supporting that *age* is a statically significant predictor of *employed*. This variable stays significant even after adjustment for heteroskedasticity with a t-vale of 7.9662.

Question b-ii)

Is there evidence of a nonlinear effect of age on the probability of being employed?

The above linear probability model indicates an inverted parabola in the relationship between *squared age* and employment rate in April 2009. Any one unit of age rise is expected to decrease the probability of being employed by

$$3.159 \times 10^{-4}$$

unit. A p-value of less than

$$2 \times 10^{-16}$$

of the *squared age* variable suggests there is strong evidence against the null hypothesis that the true coefficient of *squared age* is zero. With a statistically significant t-value of -9.08, the quadratic term is proved to be contributing to the model and thereon a non-linearity of *age* on the probability of being employed is detected.

In order to further prove the non-linearity, i.e. to prove the significance of the squared term of *age*, a linear probability model *lpm_b_1* with *age* as the only regressor is run below. With a much higher F-statistics at 47.59, the main regression *lpm_b* in question b-i carries better overall fitting goodness of the dependent variable *employed*, and *squared age* thereon should be included in a linear regression to model it.

```
# regression to detect the linearity of the variable 'age' only
lpm_b_1 <- lm(employed ~ age, data = data)
waldtest(lpm_b, lpm_b_1)
```

```
## Wald test
##
## Model 1: employed ~ age + I(age^2)
## Model 2: employed ~ age
##   Res.Df Df      F    Pr(>F)
## 1   4770
## 2   4771 -1 82.439 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taking the main regression *lpm_b* as the fitted objective, a Wald test has been performed on it against the hypothesis matrix of *lpm_b_1* regression coefficients. The null hypothesis that the coefficients of these two models are the same can be rejected at a significance level of 1% with a F-statistics of 82.439.

Therefore, *squared age* is of statistically significant explanatory power over *employed* due to the fact that including it in the linear probability model will cause a material difference in the regression. The importance of the squared term proves against the linearity of *age*.

Question b-iii)

Compute the predicted probability of employment for a 20-year-old worker, a 40year-old worker, and a 60-year-old worker

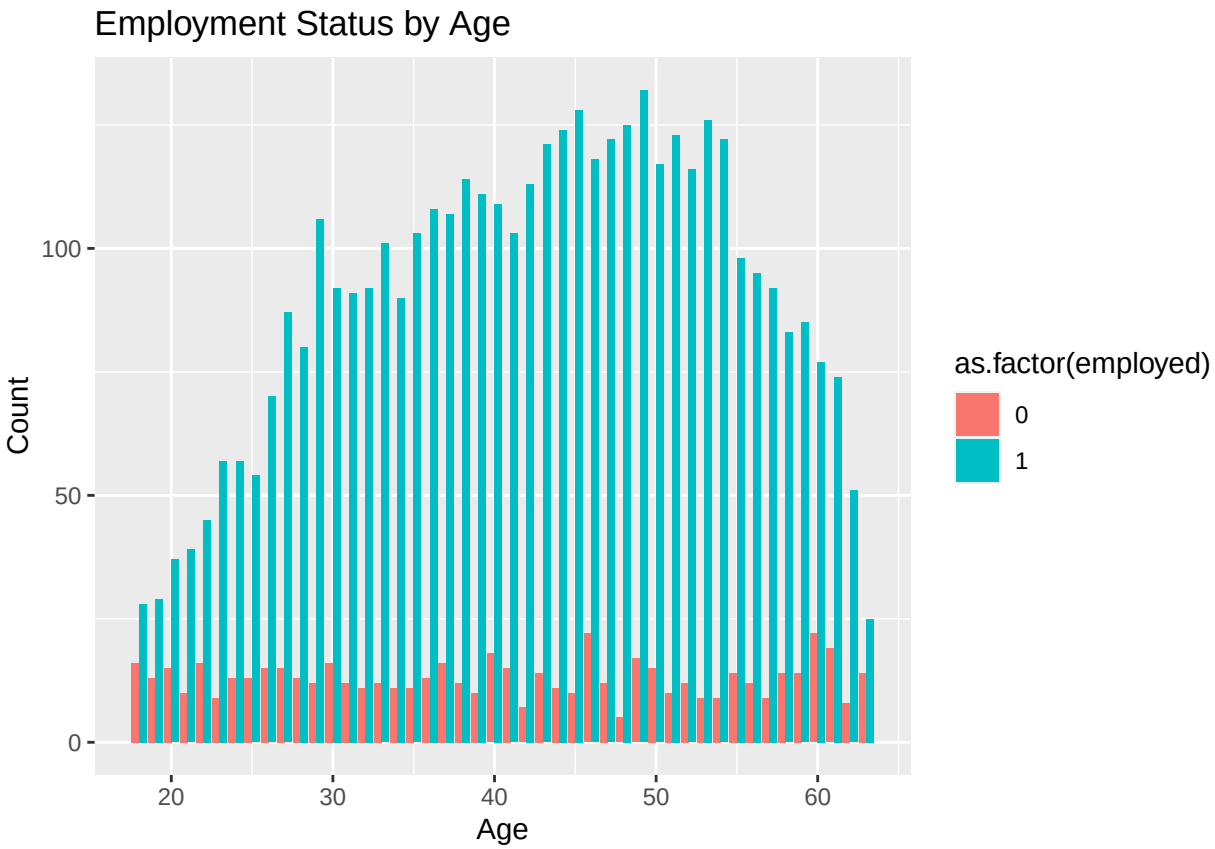
```
# predictions for sets of values of three models above
new_data <- data.frame(age = c(20, 40, 60),
                       age_sqr = c(400, 1600, 3600))
lpm_prediction <- predict(lpm_b, new_data, type = "response")
names(lpm_prediction) <- c("Age 20", "Age 40", "Age 60")
print(list(lpm_prediction = lpm_prediction))
```

```
## $lpm_prediction
##    Age 20    Age 40    Age 60
## 0.7446626 0.9148167 0.8322326
```

Based on the linear probability model, a new data set is created given the specified values of the predictors *age* and *squared age*. The prediction list states the estimated probabilities of being employed at the age of 20, 40 and 60 are approximately 74.47%, 91.48% and 83.22% respectively under the assumption of a linear relationship of both the linear and the quadratic forms of *age* on *employed*.

```
age <- seq(18, 63)

ggplot(data, aes(x=age, fill= as.factor(employed))) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Employment Status by Age", x = "Age",y = "Count")
```

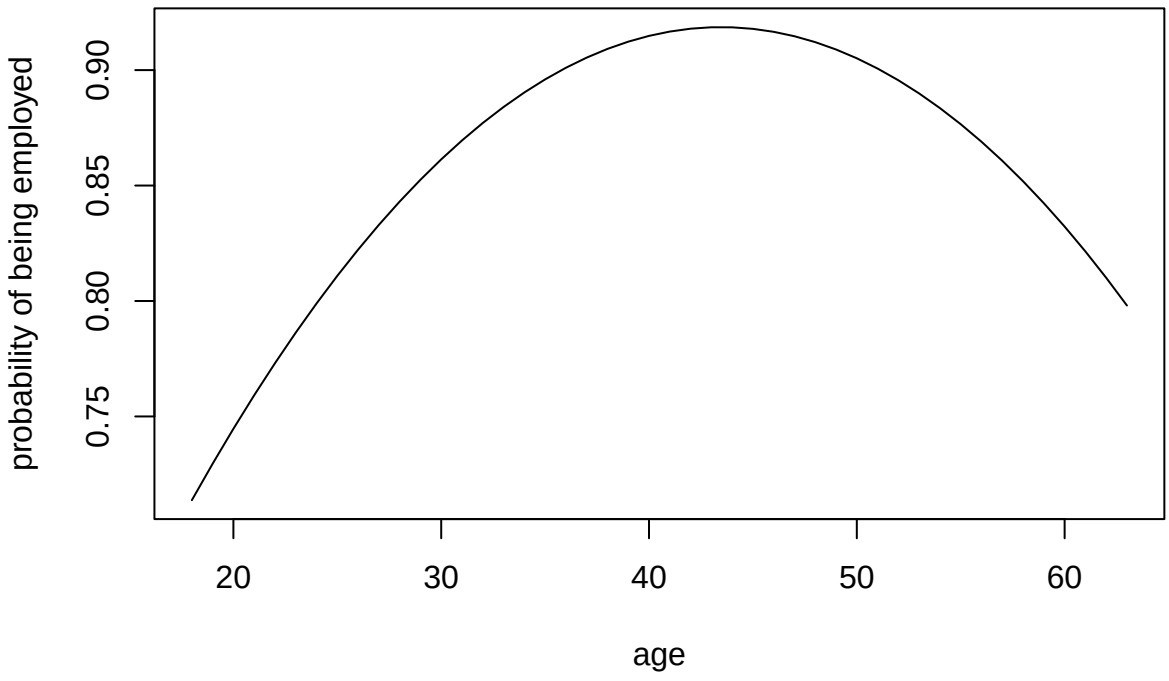


A sample *age* sequence from 18 to 63 was further created in order to detect the movement of *employed*. The absolute value of the *unemployed* variable stays within a consistent range across different age levels. On the other hand, the increasing trend of the employment probability inverted in the *age* range around 40 to 50 to decrease against age level.

This is consistent with what happens in the real world after an economic recession. Younger individuals often have less work experience and may be employed in entry-level positions more susceptible to economic fluctuations. Individuals in the middle-age bracket may have accumulated a moderate level of experience, potentially making them more resilient to job losses during economic downturns. Job security for older individuals decreases post-recession, possibly due to industry restructuring, technological advancements, or age-related biases.

```
a = lpm_b$coefficients[3]
b = lpm_b$coefficients[2]
c = lpm_b$coefficients[1]
f = function(x) {
  a * x ^ 2 + b * x + c
}

plot(age, f(age), type = 'l', ylab = "probability of being employed")
```




```
print(paste("The age with the highest probability of being employed in 2009 is",
            round(- b / (2 * a), 1)))
```

```
## [1] "The age with the highest probability of being employed in 2009 is 43.5"
```

The above curve suggests that the probability of being employed is lower at the extremes of *age*, both young and old, and higher in the middle range. The peak of the employment probability appears at the age of 43.5. The inverted parabola-shape of the curve again emphasizes the non-linearity of the linear probability model as the increasing trend does not continue after the age of 43.5 but turned to decrease instead, which is a characteristic of the negative quadratic relationship.

Question c

Repeat (b) using a probit regression

```
probit_c <- glm(employed ~ age + I(`age` ^ 2),
                family = binomial(link = "probit"), data)
summary(probit_c)
```

```
##
## Call:
## glm(formula = employed ~ age + I(age^2), family = binomial(link = "probit"),
##     data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.1949915  0.2573300  -4.644 3.42e-06 ***
## age          0.1180873  0.0133545   8.843  < 2e-16 ***
## I(age^2)    -0.0013641  0.0001619  -8.425  < 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: 3594.2  on 4772  degrees of freedom
## Residual deviance: 3511.3  on 4770  degrees of freedom
## AIC: 3517.3
##
## Number of Fisher Scoring iterations: 4
```

Question c-i)

Based on this regression, was the age a statistically significant determinant of employment in April 2009

```
# t test - statistical significance of coefficients
coeftest(probit_c, vcov = vcovHC(probit_c), type = "HC1")
```

```
##
## z test of coefficients:
##
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -1.19499145  0.26393735 -4.5276 5.967e-06 ***
## age          0.11808727  0.01380293  8.5552 < 2.2e-16 ***
## I(age^2)    -0.00136412  0.00016804 -8.1178 4.748e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of the coefficient of age is less than

$$2.2 \times 10^{-16}$$

, which is close enough to zero. At a significance level of 1%, the null hypothesis can be rejected. Thus, based on our Probit regression, it can be concluded that *age* is a statistically significant determinant of *employed* in April 2009.

```r
margins(probit_c)
```

```
## Average marginal effects

## glm(formula = employed ~ age + I(age^2), family = binomial(link = "probit"),     data = data)

##       age
##  0.001168
```

The z-value of the *age* variable at 8.555 is a further evidence to reject the null hypothesis at a significance level of 1%, i.e. there is statistical evidence that the coefficient of *age* in the probability model is different from zero.

Question c-ii)

Is there evidence of a nonlinear effect of age on the probability of being employed?

The p-value of the coefficient of *squared age* is 4.748e-16, which is nearly zero as well. As it is less than our significance level, we could reject the null hypothesis. Thus, based on our probit regression, we could say *squared age* is also a statistically significant determinant of employment in April 2009. The quadratic term is proved to be contributing to the model and therefore a non-linearity of *age* on the probability of being employed is detected.

In order to further prove the non-linearity, i.e. to prove the significance of the squared term of *age*, a probit model with *age* as the only regressor is run below. And then waldest is performed on both the models i.e. one model with *squared age* and the other without. The result of the waldtest confirms that both the models are different at less than 1% significance. Hence, it is confirmed that the *squared age* is significant and it explains the nonlinear effect of age on the probability of being employed.

```r
# regression to detect the linearity of the variable 'age' only
probit_c_1 <- glm(employed ~ age,
                  family = binomial(link = "probit"), data)
waldtest(probit_c, probit_c_1)
```

```
## Wald test
##
## Model 1: employed ~ age + I(age^2)
## Model 2: employed ~ age
##   Res.Df Df      F    Pr(>F)
## 1   4770
## 2   4771 -1 70.987 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By looking at the Wald test between the LPM model and the Probit model, a p-value of less than 2.2*10^(-16) is observed, which is nearly zero. As the p-value is less than the significance level (even considering a 1% here), the null hypothesis could be rejected which states that the two models are the same. Therefore, the Probit model is different from the Probit model with only* age* as the variable, and the non-linearity of *age* on the probability of being employed is reinforced.

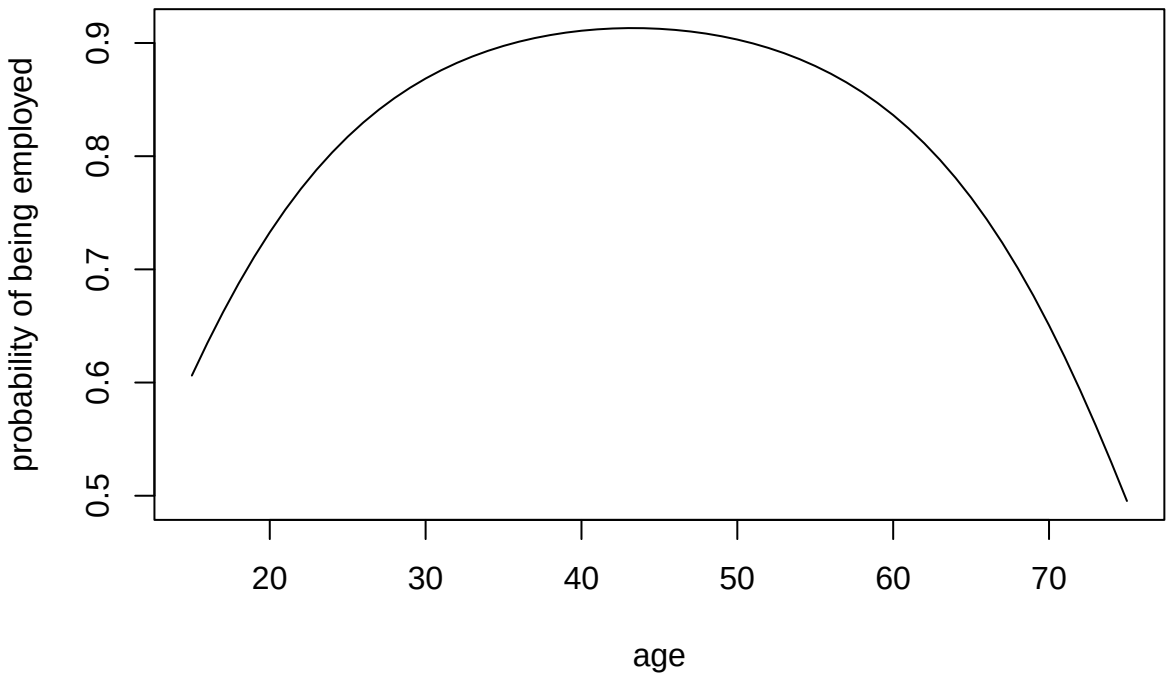
Question c-iii)

Compute the predicted probability of employment for a 20-year-old worker, a 40year-old worker, and a 60-year-old worker

```
# predictions for sets of values of three models above
new_data <- data.frame(age = c(20, 40, 60),
                       age_sqr = c(400, 1600, 3600))
probit_prediction <- predict(probit_c, new_data, type = "response")
names(probit_prediction) <- c("Age 20", "Age 40", "Age 60")
print(list(probit_prediction = probit_prediction))
```

```
## $probit_prediction
##    Age 20    Age 40    Age 60
## 0.7327351 0.9108339 0.8363122
```

```
a = probit_c$coefficients[3]
b = probit_c$coefficients[2]
c = probit_c$coefficients[1]
f = function(x) {
  pnorm(a * x^2 + b * x + c)
}
age <- seq(15, 75)
plot(age, f(age), type = 'l', ylab = "probability of being employed")
```



```r
print(paste("The age with the highest probability of being employed in 2009 is",
            round(- b / (2 * a), 1)))
```

```
## [1] "The age with the highest probability of being employed in 2009 is 43.3"
```

Based on the Probit model, a new data set is created given the specified values of the predictors *age* and *squared age*. The prediction list states the estimated probabilities of being employed at the age of 20, 40 and 60 are approximately 73.27%, 91.08% and 86.63% respectively. This means, for example, a 20-year-old worker is predicted to be employed in April 2009 with a probability of 0.7327351, based on the historical data we have.

The lowest predicted employment probability at the *age* of 20 might be due to the fact that people at a younger age of 20 have less experience and therefore face greater exposure to unemployment risk in an economic recession. And the old people may face some physical restrictions while working, which may also have a significant impact on their chances of getting employed. Middle-aged people are the main workforce in society with their better physical and experiential conditions. Such a trend can also be shown in our plot of the probability of being employed under the Probit model.

Question d

Repeat (b) using a logit regression

```
# required linear probability model
logit_d <- glm(employed ~ age + I(`age` ^ 2),
               family = binomial(link = "logit"), data)
summary(logit_d)
```

```
##
## Call:
## glm(formula = employed ~ age + I(age^2), family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.3733020  0.4569256  -5.194 2.06e-07 ***
## age          0.2187055  0.0240930   9.078  < 2e-16 ***
## I(age^2)    -0.0025338  0.0002937  -8.627  < 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: 3594.2  on 4772  degrees of freedom
## Residual deviance: 3510.5  on 4770  degrees of freedom
## AIC: 3516.5
##
## Number of Fisher Scoring iterations: 4
```

Question d-i)

Based on this regression, was the age a statistically significant determinant of employment in April 2009

```
# t test - statistical significance of coefficients
coeftest(logit_d, vcov = vcovHC(logit_d), type = "HC1")
```

```
##
## z test of coefficients:
##
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -2.37330198  0.46795737 -5.0716 3.944e-07 ***
## age          0.21870551  0.02486199  8.7968 < 2.2e-16 ***
## I(age^2)    -0.00253379  0.00030442 -8.3235 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The logistic regression is a statistical method used for modelling the probability that a dependent variable is categorical with two levels at 0 and 1, in this case, the *employed* binary variable. The sigmoid function is used to model the relationship of *age* and *squared age* against the employment probability.

Based on the above result, the *age* variable is a statistically significant determinant of the April 2009 Employment probability rate with over 99% confidence level as the model p-value is less than 0.01.

```r
margins(logit_d)
```

```
## Average marginal effects

## glm(formula = employed ~ age + I(age^2), family = binomial(link = "logit"),     data = data)

##       age
##  0.001216
```

The z-value of the *age* variable at 8.7968 is significant enough to reject the null hypothesis at a significance level of 1%, i.e. there is no statistical evidence that the coefficient of *age* in the probability model is different from zero. The p-value is less than

$$2 \times 10^{-16}$$

, which is way smaller than the significance level, further supporting that *age* is a statically significant predictor of *employed*. One-unit change in *age* shall lead to 0.001215823 units of change in the dependent variable *employed*.

Question d-ii)

Is there evidence of a nonlinear effect of age on probability of being employed?

The logit model *logid_d* above reveals a distinctive concave shape in the relationship between *squared age* and *employed* in April 2009. The statistical significance of the t-value at -8.3235 associated with the quadratic term validates its substantial impact on the model, indicating a pronounced non-linear effect of *age* on the likelihood of being employed.

To further establish the significance of the non-linear relationship, a logit model *logit_d_1* with *age* as the only regressor is created below. Subsequently, a Wald test is conducted on both models: one incorporating the *squared age* term and the other without. The results of the Wald test provide compelling evidence that the two models differ significantly at a significance level of less than 1%. This again confirms the importance of the *squared age* term, substantiating its role in explaining the non-linear impact of *age* on the probability of being employed.

```r
# regression to detect the linearity of the variable 'age' only
logit_d_1 <- glm(employed ~ age,
                 family = binomial(link = "logit"), data)
waldtest(logit_d, logit_d_1)
```

```
## Wald test
##
## Model 1: employed ~ age + I(age^2)
## Model 2: employed ~ age
##   Res.Df Df      F    Pr(>F)
## 1   4770
## 2   4771 -1 74.422 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question d-iii)

Compute the predicted probability of employment for a 20-year-old worker, a 40year-old worker, and a 60-year-old worker

```
# predictions for sets of values of three models above
new_data <- data.frame(age = c(20, 40, 60),
                       age_sqr = c(400, 1600, 3600))
logit_prediction <- predict(logit_d, new_data, type = "response")
names(logit_prediction) <- c("Age 20", "Age 40", "Age 60")
print(list(logit_prediction = logit_prediction))
```

```
## $logit_prediction
##    Age 20    Age 40    Age 60
## 0.7285524 0.9105890 0.8358088
```

Based on the *logit_d* model, a new data set is created given the specified values of the predictors *age* and *squared age*. With another similar tendency, the logit prediction list states the estimated probabilities of being employed at the age of 20, 40 and 60 are approximately 72.86%, 91.06% and 83.58% respectively.

```
a = logit_d$coefficients[3]
b = logit_d$coefficients[2]
c = logit_d$coefficients[1]
f = function(x) {
  1 / (1 + exp(-a * x ^ 2 - b * x - c))
}
age <- seq(15, 75)
plot(age, f(age), type = 'l', ylab = "probability of being employed")
```


```
print(paste("The age with the highest probability of being employed in 2009 is",
            round(- b / (2 * a), 1)))
```

```
## [1] "The age with the highest probability of being employed in 2009 is 43.2"
```

Again the sample *age* sequence from 18 to 63 was created to plot the employment probability curve with a similar depicted curve tendency and a similar *age* peak at 43.2 years old. Beyond this peak point, the rising employment probability starts to decline, illustrating a characteristic quadratic relationship. The value of *employed* diminishes at both tails to underscore the non-linear nature of the logistic regression model.

Question e

Are there important differences in your answers to (b)-(d)? Explain

```
model_list <-
  list(summary(margins(lpm_b)), summary(margins(probit_c)), summary(margins(logit_d)))

model_df <- do.call(rbind, model_list)
model_names <- c("LPM", "Probit", "Logit")

model_df <- cbind(Model = model_names, model_df)
model_df
```

```
##    Model factor         AME           SE        z          p        lower
## 1    LPM    age 0.000891404 0.0004116039 2.165684 0.03033533 8.467523e-05
## 2 Probit    age 0.001168486 0.0005355142 2.181988 0.02911042 1.188970e-04
## 3  Logit    age 0.001215823 0.0005598945 2.171521 0.02989181 1.184496e-04
##         upper
## 1 0.001698133
## 2 0.002218074
## 3 0.002313196
```

The probabilities of being employed will increase with *age* by 0.089%, 0.11%, and 0.12%, observed from the Linear Probability Model (LPM), Probit, and Logit models, respectively. In addition, the margins and the age that is predicted to have a highest probability of being unemployed are all similar among the three models. Thus, there is not a significant difference between the answers to questions (b) to (d).

However, these three models are very different in definition. LPM is a regression model used for binary outcomes, where the dependent variable is binary (usually 0 or 1). Coefficients in LPM can be interpreted as the change in the probability of the event for a one-unit change in the corresponding independent variable. However, there is a limitation of LPM - the predicted probabilities can fall outside the usual probability [0, 1] range, and it may suffer from heteroscedasticity.

Probit is a type of regression model used for binary outcomes, similar to the LPM. However, it uses the cumulative distribution function of the standard normal distribution (Probit function) to model the relationship. Unlike LPM, coefficients in the probit model represent the change in the z-score of the latent variable for a one-unit change in the corresponding independent variable.

Consider a Logit model where y is the dependent variable and the xs are the independent variables, the log odds of the probabilities (log(P(x)/1-P(x)) is modeled as a linear combination of the predictor variables. Different from LPM, increasing the i-th x by one unit changes the log odds by units of the amount of the corresponding coefficient *beta-i*. Equivalently, the odds change in percentage by an exponential of *beta-i*. Both Probit and Logic models will have the predicted probabilities lying in the usual probability range of [0,1].

Regarding the limitations of the Probit and Logit models, they are both logistic regression models and have a non-linear relationship between the independent variables and the probability of the event occurring. While this is often more realistic than the linear probability model, it can make interpretation more complex.

Thus, the three models are different from each other by definition, even though they show similar results in our analysis of the effect of age on employment in question (b) to (d).

The data set includes variables measuring the workers' educational attainment, sex, race, marital status, region of the country, and weekly earnings in April 2008.

Question f

```
lpm_f <- lm(employed ~ age + I(age^2) + educ_lths + educ_hs + educ_somecol +
              educ_aa + educ_bac + educ_adv + female + married + race +
              ne_states + so_states + ce_states + we_states + earnwke,
            data = data)
summary(lpm_b)
```

```
##
## Call:
## lm(formula = employed ~ age + I(age^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91854  0.08342  0.09933  0.13867  0.28625
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.218e-01  5.705e-02   5.640  1.8e-08 ***
## age          2.746e-02  2.894e-03   9.491  < 2e-16 ***
## I(age^2)    -3.159e-04  3.479e-05  -9.080  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3274 on 4770 degrees of freedom
## Multiple R-squared:  0.01956,    Adjusted R-squared:  0.01915
## F-statistic: 47.59 on 2 and 4770 DF,  p-value: < 2.2e-16
```

```
summary(lpm_f)
```

```
##
## Call:
## lm(formula = employed ~ age + I(age^2) + educ_lths + educ_hs +
##     educ_somecol + educ_aa + educ_bac + educ_adv + female + married +
##     race + ne_states + so_states + ce_states + we_states + earnwke,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99075  0.06617  0.10387  0.14472  0.39438
##
## Coefficients: (2 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.700e-01  6.209e-02   5.959 2.72e-09 ***
## age           2.459e-02  3.041e-03   8.087 7.71e-16 ***
## I(age^2)     -2.872e-04  3.611e-05  -7.954 2.24e-15 ***
## educ_lths    -8.385e-02  2.398e-02  -3.496 0.000476 ***
## educ_hs      -2.238e-02  1.734e-02  -1.291 0.196923
```

```
## educ_somecol -1.767e-03  1.814e-02  -0.097 0.922375
## educ_aa       6.571e-03  2.010e-02   0.327 0.743724
## educ_bac     -1.304e-02  1.702e-02  -0.766 0.443693
## educ_adv             NA         NA      NA       NA
## female       -5.206e-03  9.972e-03  -0.522 0.601672
## married      -7.354e-04  1.042e-02  -0.071 0.943731
## race         -1.015e-02  8.388e-03  -1.210 0.226186
## ne_states     1.366e-02  1.418e-02   0.963 0.335557
## so_states     1.836e-02  1.299e-02   1.413 0.157708
## ce_states     4.087e-02  1.359e-02   3.007 0.002653 **
## we_states            NA         NA      NA       NA
## earnwke       3.481e-05  9.727e-06   3.579 0.000349 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3258 on 4758 degrees of freedom
## Multiple R-squared:  0.03157,    Adjusted R-squared:  0.02872
## F-statistic: 11.08 on 14 and 4758 DF,  p-value: < 2.2e-16
```

Question f-i)

By adding those covariates to the linear probability model regression of point (b), investigate whether the conclusions on the effect of Age on employment from (b) are affected by omitted variable bias

The comparison between the initial linear regression model and the adjusted model reveals a small difference in the coefficients for *age* and *squared age*. In the initial model *lpm_b*, *age* and *squared age* exhibited significant effects in different directions on employment probability. In the adjusted model *lpm_f*, while the coefficients for *age* and *squared age* remain significant at a significance level of 1% with t-values at 8.087 and -7.954 respectively, the coefficient values have changed slightly. This difference suggests that the additional variables have impacted the interpretation of the *age*-related effects on *employed*. Namely, the introduction of new variables such as regional factors, weekly earnings and *educ_lths* - the highest level of education is less than a high school graduate - points out potential omitted variable bias in the initial model.

Additionally, the statistical significance of certain new variables including educational factors, regional factors and weekly earnings in the adjusted model highlights the importance of considering a broader set of factors. Moreover, the higher adjusted R-squared of the adjusted model indicates an increase in explanatory power which in turn suggests the added covariates contribute to a better understanding of the variability in employment status.

To conclude, while these results might suggest the presence of some omitted variable bias in the initial model, the small coefficient difference for *age* and *squared age* between the two models and the fact they are still extremely significant in the adjusted model leads us to think there is not an impactful omitted variable bias. Nevertheless, it is important to point out that the adjusted model better explains the changes due to the added covariates.

Question f-ii)

Use the regression results to discuss the characteristics of workers who were hurt the most by the 2008 financial crisis

The regression results provide insights into the characteristics of workers particularly affected by the 2008 financial crisis.

Firstly, the negative coefficient for the variable representing workers with less than a high school education - *educ_lths* - suggests that individuals with lower educational attainment faced a more significantly negative

impact on their employment status, i.e. they were more likely to be unemployed in April 2009. This fact aligns with the usual economic literature which defends a higher vulnerability of less-educated workers during economic downturns.

Additionally, the positive coefficient for the variable indicating workers from central states - *ce_states* - implies that individuals residing in central regions of the country experienced a more significantly negative effect on their employment probability compared to those in other regions. One possible explanation for the geographical disparity in the impact of the financial crisis could be attributed to variations in industry composition across different regions. This aligns with the fact that central states are deeply connected with the automotive industry (e.g. Michigan) and with manufacturing and exports, both of which were severely affected by the Great Recession.

Moreover, the positive and statistically significant coefficient for weekly earnings - *earnwke* - indicates that workers with higher pre-crisis earnings experienced a relatively less severe negative impact on their employment status. This underscores the financial resilience of individuals with higher earnings, suggesting that economic downturns disproportionately affected those with lower income levels. One possible explanation is the tendency of large corporate companies to heavily downsize their workforce during a financial crisis to reduce costs but not its higher levels - or at least not to the same extent. Subsequently, this leads to the correlation between lesser weekly earnings - *workforce* - and a higher probability of being let go during a financial crisis. Another possible explanation comes from the fact higher weekly earners are usually more qualified. For example, if we considered someone who became unemployed after the data was collected, the more qualified they were the higher the chance they could find another position before April 2009. In turn, this would again substantiate the positive coefficient for weekly earnings.

Finally, the positive coefficient for age suggests that, on average, older workers experienced more favourable outcomes regarding employment in the aftermath of the crisis. This finding is consistent with the idea that older workers possess valuable experience, skills, and industry knowledge, making them more valuable to a company and, consequently, more resilient to economic downturns. However, the negative coefficient for *squared age* indicates a diminishing positive effect with age increase and a decline in this effect for the oldest workers. This diminishing effect could be due to factors such as age-related discrimination, and changing labour market dynamics such as technological changes or retirement, all of which would be exacerbated during a financial crisis and associated cost-cutting policies.

Overall, the analysis suggests that the 2008 financial crisis disproportionately affected younger workers, those with lower educational attainment and less weekly earnings, and those residing in specific geographical regions.

Question (g)

In this question, we examine the in-sample accuracy using the three models established in question b to d, i.e., the linear probability model, the probit model, and the logit model that regress Employed on Age and Age^2. The methodology is to make predictions using the three models separately, but using the original data rather than new data. Based on the prediction result, we make binary classification. If the prediction value is larger than 0.5, we classify it as 1, which means the individual will be employed. Otherwise, if the prediction value is less than 0.5, we classify it as 0, which means that the individual will not be employed. Next, we compare the classified prediction result on employment with the real employment case. The total percentage that prediction employment status is equal to the the real employment case will be the in-sample accuracy of the models.

```
# Predictions for models b, c, d (In-sample)
predictions_b_raw <- predict(lpm_b,
                             data, type = "response")  # For linear probability model
predictions_c_raw <- predict(probit_c, data, type = "response")  # For probit model
predictions_d_raw <- predict(logit_d, data, type = "response")  # For logit model

# Make binary classification
predictions_b <- ifelse(predictions_b_raw > 0.5, 1, 0)
predictions_c <- ifelse(predictions_c_raw > 0.5, 1, 0)
predictions_d <- ifelse(predictions_d_raw > 0.5, 1, 0)

# Get actual classes
actual_employed <- data$employed  # Assuming 'Employed' is the actual employment status

# Evaluate in-sample accuracy
lpm_b_accuracy <- sum(predictions_b == actual_employed) / length(actual_employed)
lpm_c_accuracy <- sum(predictions_c == actual_employed) / length(actual_employed)
lpm_d_accuracy <- sum(predictions_d == actual_employed) / length(actual_employed)

print(paste("The proportion of correctly assigned classes of model b is ",
            lpm_b_accuracy*100, "%"))
```

```
## [1] "The proportion of correctly assigned classes of model b is  87.5130944898387 %"
```

```
print(paste("The proportion of correctly assigned classes of model c is ",
            lpm_c_accuracy*100, "%"))
```

```
## [1] "The proportion of correctly assigned classes of model c is  87.5130944898387 %"
```

```
print(paste("The proportion of correctly assigned classes of model d is ",
            lpm_d_accuracy*100, "%"))
```

```
## [1] "The proportion of correctly assigned classes of model d is  87.5130944898387 %"
```

From the in-sample test, the proportion of correctly assigned classes for models b-d is 87.513% for all models. Although there are differences in raw in-sample predictions among these three models, the binary classification of the three models' prediction results are actually identical.

```r
if(identical(predictions_b_raw, predictions_c_raw) &&
   identical(predictions_b_raw, predictions_d_raw))  {
  print("The in-sample prediction of the three models' predictions are identical.")
} else {
  print("The in-sample prediction of the three models' predictions are not identical.")
}
```

```
## [1] "The in-sample prediction of the three models' predictions are not identical."
```

```r
if(identical(predictions_b, predictions_c) &&
   identical(predictions_b, predictions_d))  {
  print("The binary classification of the three models' predictions are identical.")
} else {
  print("The binary classification of the three models' predictions are not identical.")
}
```

```
## [1] "The binary classification of the three models' predictions are identical."
```

This result implies that although the three models have differences in regression principles, they are consistent in showing the strong correlation between the prediction target employment status and predictors *age* and *squared age*. On the other hand, it shows that predicting employment status through *age* and *squared age* is not comprehensive enough as there is over 12% error in all three models. The introduction of other predictors such as regional factors, weekly earnings and *educ_lths* - the highest level of education is less than a high school graduate - are also necessary for comparing.

```r
confusion_matrix_LPM <- table(Actual = data$employed, Predicted = predictions_b)
confusion_matrix_logit  <- table(Actual = data$employed, Predicted = predictions_c)
confusion_matrix_probit <- table(Actual = data$employed, Predicted = predictions_d)
print(confusion_matrix_LPM)
```

```
##       Predicted
## Actual    1
##      0  596
##      1 4177
```

```r
print(confusion_matrix_logit)
```

```
##       Predicted
## Actual    1
##      0  596
##      1 4177
```

```r
print(confusion_matrix_probit)
```

```
##       Predicted
## Actual    1
##      0  596
##      1 4177
```

Question (h)

(1) Naïve Bayes Classifier

```
nb_model <- naiveBayes(as.factor(employed) ~ age + age^2, data)
nb_predictions <- predict(nb_model, data, type = "class")
nb_accuracy <- sum(nb_predictions == actual_employed) / length(actual_employed)
print(paste("In-sample accuracy for Naive Bayes Classifier is ", nb_accuracy))
```

```
## [1] "In-sample accuracy for Naive Bayes Classifier is  0.875130944898387"
```

The Naïve Bayes classifier operates on the principle of Bayes' theorem with a simplifying assumption that features are independent. Despite this naive assumption, it effectively computes the probability of a data point belonging to a certain class by evaluating the conditional probabilities of each feature given the class label. Naive Bayes excels in processing large datasets due to its computational efficiency and general robustness. It has been proven remarkably effective even when the independence assumption doesn't hold perfectly in real-world datasets.

(2) Linear Discriminant Analysis

```
lda_model <- lda(as.factor(employed) ~ age + I(`age`^2), data)
lda_predictions <- predict(lda_model, data)$class
lda_accuracy <- sum(lda_predictions == actual_employed) / length(actual_employed)
print(paste("In-sample accuracy for Linear Discriminant Analysis is ", lda_accuracy))
```

```
## [1] "In-sample accuracy for Linear Discriminant Analysis is  0.875130944898387"
```

Linear Discriminant Analysis (LDA) is a dimensionality reduction and classification technique used in machine learning and statistics. It aims to find a linear combination of features that characterizes or separates two or more classes in the data. LDA identifies the axes (linear discriminants) that maximize the separation between classes while minimizing the variance within each class. By projecting the data onto these discriminative axes, LDA transforms the original features into a lower-dimensional space, making it easier to visualize and classify the data. It's particularly effective when the classes are well-separated and the assumptions of normally distributed classes with equal covariance matrices hold, enabling it to efficiently classify new data points based on their projected positions in this reduced space.

The accuracy of LDA in this scenario is equal to that of LPM, probit and logit models. The reason could be that the dimension in this model is particularly low, only consisting of ages. In scenarios where the data has a lower number of dimensions and the linear separability assumption holds, LDA might perform equally well as linear models. LDA is particularly effective in reducing dimensionality while preserving discriminatory information.

However, it is important to acknowledge that although Naive Bayes Classifier results in the same outcome of LDA in this particular scenario, the scenarios where both models perform equally well might be limited as they make different underlying assumptions about the data. Naive Bayes tends to perform well even with relatively less data and when the independence assumption isn't severely violated, while LDA's effectiveness relies more on the linear separability of the classes and the adherence to its assumptions.

(3) Quadratic Discriminant Analysis

```r
qda_model <- qda(as.factor(employed) ~ age + I(`age`^2), data)
qda_predictions <- predict(qda_model, data)$class
qda_accuracy <- sum(qda_predictions == actual_employed) / length(actual_employed)
print(paste("In-sample accuracy for Quadratic Discriminant Analysis is ", qda_accuracy))
```

```
## [1] "In-sample accuracy for Quadratic Discriminant Analysis is  0.864655353027446"
```

Quadratic Discriminant Analysis (QDA) is a classification technique similar to Linear Discriminant Analysis (LDA) but without the assumption of equal covariance matrices across classes. QDA calculates separate covariance matrices for each class, allowing for more flexibility in modelling the decision boundary between classes. By considering different covariance structures, QDA can capture nonlinear decision boundaries, making it more adaptable to datasets where classes have distinct and non-linear separations. QDA can be particularly powerful when the assumption of equal covariance matrices across classes, which is required by LDA, doesn't hold in the dataset, enabling better adaptation to the true underlying data distribution and potentially leading to improved classification accuracy.

However, QDA's flexibility comes at the cost of increased model complexity, requiring the estimation of more parameters and potentially being more prone to overfitting with smaller datasets compared to LDA. In this scenario, we only set parameters through ages and do not provide complex parameters that QDA needs. And that could be the reason why QDA performs worse than the LDA and all the models prior to it.

(4) Decision trees

```r
tree_model <- tree(as.factor(employed) ~ age + I(`age`^2), data)
tree_predictions <- predict(tree_model, data, type = "class")
tree_accuracy <- sum(tree_predictions == actual_employed) / length(actual_employed)
print(paste("In-sample accuracy for Decision Tree is ", tree_accuracy))
```

```
## [1] "In-sample accuracy for Decision Tree is  0.875130944898387"
```

Decision trees are versatile, non-parametric supervised learning models used for both classification and regression tasks. They hierarchically partition the feature space into segments based on the data's features, creating a tree-like structure where each internal node represents a feature test, each branch represents an outcome of that test, and each leaf node represents a class label or a predicted value. By sequentially splitting the data based on the most informative features, decision trees aim to maximize homogeneity within each resulting subset regarding the target variable. Their interpretability, ability to handle both numerical and categorical data and resistance to outliers make them popular.

The weakness of decision trees is that they are prone to overfitting, especially with complex structures or noisy data, which can be mitigated through techniques like pruning or using ensemble methods like random forests or gradient boosting. In this scenario, we do not require a decision tree to predict complex structures but only require it to predict simple regression on age, squared age and predict employment. So decision tree here could reach the same level of accuracy as LPM and other models.

(5) Random forests

```r
set.seed(123) # Setting a seed for reproducibility
rf_model <- randomForest(as.factor(employed) ~ age + I(`age`^2), data, num.trees= 100)
rf_predictions <- predict(rf_model, data)
rf_accuracy <- sum(rf_predictions == actual_employed) / length(actual_employed)
print(paste("In-sample accuracy for Random Forest (100 trees) is ", rf_accuracy))
```

```
## [1] "In-sample accuracy for Random Forest (100 trees) is  0.875130944898387"
```

Random Forest is based on decision tree classifiers that build multiple decision trees and merge their predictions to improve accuracy and reduce overfitting. It operates by creating a diverse set of decision trees through random sampling of both data points (bootstrap aggregating or "bagging") and features at each split. By aggregating predictions from these trees, typically through averaging for regression tasks or voting for classification, Random Forest reduces the variance and tends to yield more robust and accurate predictions than a single decision tree.

The optimisation allows random forests to handle high-dimensional data, maintain good performance with default parameters, and provide insights into feature importance. In this scenario, we do not provide a high-dimensional environment and cannot show the advantage that random forest improves at complex models. The accuracy of low dimensional age-employment prediction of random forest remains the same as of decision trees.

(6) K-Nearest Neighbours

```r
# First scale the data since kNN is sensitive to the scale of the data
modified_data <- data %>%
  mutate(square_age = age*age)
scaled_data <- scale(modified_data[, c("age", "square_age")])
# Define the number of neighbours
k <- 5
knn_predictions <- knn(train = scaled_data, test = scaled_data,
                       cl = as.factor(data$employed), k = k)
knn_accuracy <- sum(knn_predictions == actual_employed) / length(actual_employed)
print(paste("In-sample accuracy for KNN (k=5, 5 nearest neighbours) is ", knn_accuracy))
```

```
## [1] "In-sample accuracy for KNN (k=5, 5 nearest neighbours) is  0.875130944898387"
```

K-Nearest Neighbors (KNN) is a simple yet effective supervised learning algorithm used for both classification and regression tasks. It works by memorizing the entire training dataset and classifying new instances based on their similarity to the known data points. The algorithm determines the class of a new data point by looking at the majority class among its K nearest neighbours, where the "closeness" is typically computed using distance metrics like Euclidean or Manhattan distance in the feature space.

KNN's simplicity and flexibility make it easy to implement, especially for small to medium-sized datasets, but it might be computationally expensive with larger datasets due to its need to store and compute distances for all training instances. Additionally, choosing the right value of K and appropriate distance metrics is crucial for optimal performance. In this small-sized dataset scenario, KNN performs with similar accuracy between different K values and remains the same as previous models like LPM because the data is sparse and doesn't have a large number of dimensions.

–END–