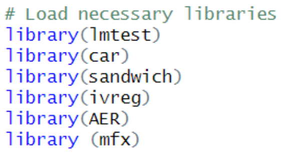
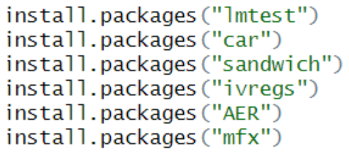
Dataset number – 14

Date – 15- 05 -2025

HOMEWORK – 1



```
# Load necessary libraries
library(lmtest)
library(car)
library(sandwich)
library(ivreg)
library(AER)
library (mfx)
```

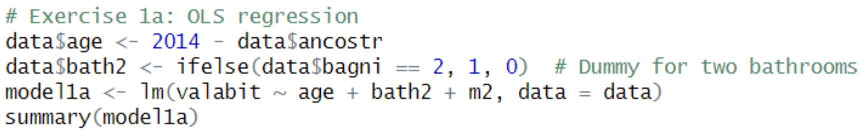


```
install.packages("lmtest")
install.packages("car")
install.packages("sandwich")
install.packages("ivregs")
install.packages("AER")
install.packages("mfx")
```

STEP-1 Before starting the analysis we have to lode these libraries in R. and if the libraries are not installed in your R, then we must install it CHECK THE STEP - 2

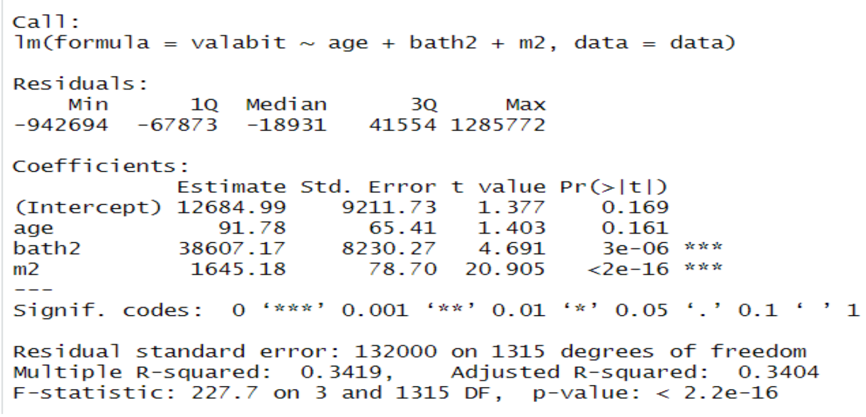
EX – 1 (A)

How can we run the OLS regression? So, to run the OLS regression We must give the command in R see down below.



```
# Exercise 1a: OLS regression
data$age <- 2014 - data$ancostr
data$bath2 <- ifelse(data$bagni == 2, 1, 0)  # Dummy for two bathrooms
model1a <- lm(valabit ~ age + bath2 + m2, data = data)
summary(model1a)
```

After running the regression here, we can see the out came.

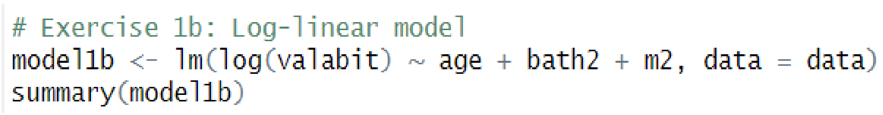


```
Call:
lm(formula = valabit ~ age + bath2 + m2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-942694  -67873  -18931   41554 1285772 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12684.99    9211.73   1.377    0.169    
age            91.78      65.41   1.403    0.161    
bath2       38607.17    8230.27   4.691    3e-06 ***
m2           1645.18      78.70  20.905   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 132000 on 1315 degrees of freedom
Multiple R-squared:  0.3419,	Adjusted R-squared:  0.3404 
F-statistic: 227.7 on 3 and 1315 DF,  p-value: < 2.2e-16
```

The coefficient for age is positive (+91.78), which means each year of age corresponds with an increase of about 91.78 euros in house value. But however, the p-value is 0.161, which is not statistically significant at normal levels, just like (5%). So that means that there is no strong evidence in the data that the age of the house influences its value. And for that reason, we cannot conclude that age contributes to determining house value based on this model.

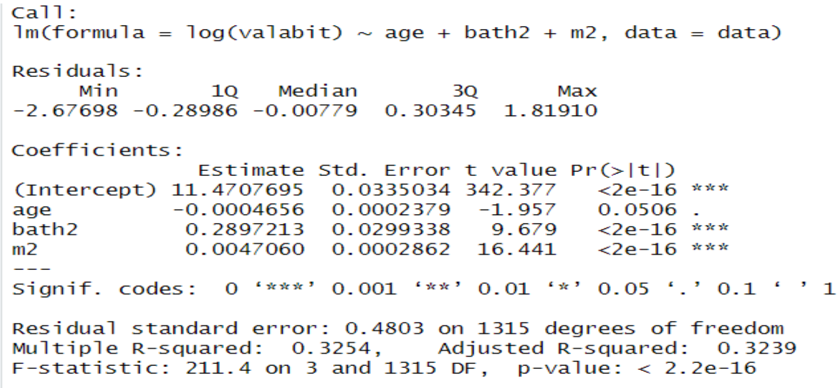
EX- 1 (B)

Convert the dependent variable in log we are using function **(log)** in the R**.**



```
# Exercise 1b: Log-linear model
model1b <- lm(log(valabit) ~ age + bath2 + m2, data = data)
summary(model1b)
```

Here as you can see that we put the log function to that dependent variable

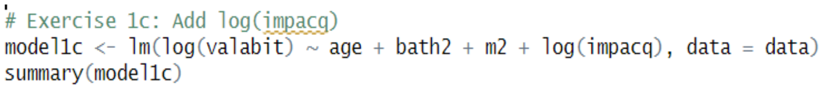


```
Call:
lm(formula = log(valabit) ~ age + bath2 + m2, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.67698 -0.28986 -0.00779  0.30345  1.81910 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.4707695  0.0335034 342.377   <2e-16 ***
age         -0.0004656  0.0002379  -1.957   0.0506 .  
bath2        0.2897213  0.0299338   9.679   <2e-16 ***
m2           0.0047060  0.0002862  16.441   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4803 on 1315 degrees of freedom
Multiple R-squared:  0.3254,	Adjusted R-squared:  0.3239 
F-statistic: 211.4 on 3 and 1315 DF,  p-value: < 2.2e-16
```

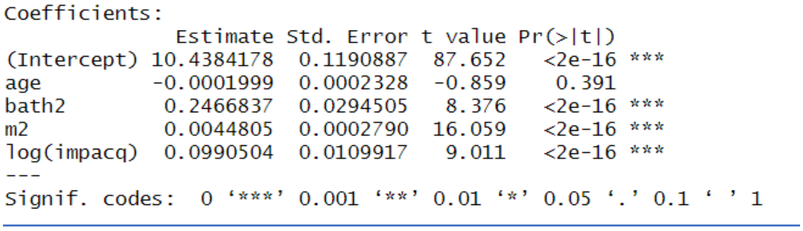
So, for the **linear regression modal** (1), the coefficients of m2=1645.18 (highly significant, p<0.001) so that means each additional square meter increases the house values by 1,645 units. "Other factors constant" and **log linear regression model** (2) the coefficient of m2:0.0047 (highly significant, p<0.001) which mean each additional square meter increase the log of house values by 0.0047% for house value par extra square meter ("because in log models' coefficients are in percentage changes.) here both models show that the size of the house(m2) is a highly significant and important to show of the values.

EX- 1 (C)

To add the log purchase price of the house in the regression we must use ***log (impacq)*** in R.



```
# Exercise 1c: Add log(impacq)
model1c <- lm(log(valabit) ~ age + bath2 + m2 + log(impacq), data = data)
summary(model1c)
```

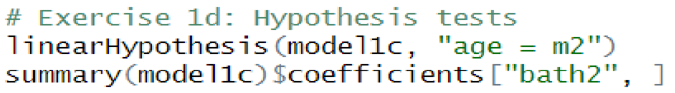


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.4384178  0.1190887  87.652   <2e-16 ***
age         -0.0001999  0.0002328  -0.859    0.391    
bath2        0.2466837  0.0294505   8.376   <2e-16 ***
m2           0.0044805  0.0002790  16.059   <2e-16 ***
log(impacq)  0.0990504  0.0109917   9.011   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

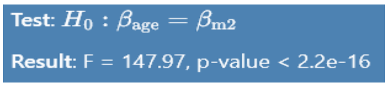
When we include the new variable **log(impacq),** which is the purchase price of the house, into the regression, so for that the coefficient is 0.099, with a very small error with a large value of 9.011 and a p -value less than 0.001. That means that the coefficient is highly statistically significant. Also, for example, when we test the null hypothesis that the coefficient is zero, that means no effect on house values.in another way, it's different from zero, since the p-value is very small (0.05) **we reject the null hypothesis**. So, to conclude that a 1% increase in the **(impacq)** 0.099% increase in house value shows a positive impact on this model.

EX-1(D)

To check the hypothesis that the effects of age and m2 are the same and test the hypothesis that the effect is two bathrooms and one or no bathrooms, we must run the linear hypothesis test in R. ("we have to run an F-test"). Also to run the hypothesis, we must use a library (car).

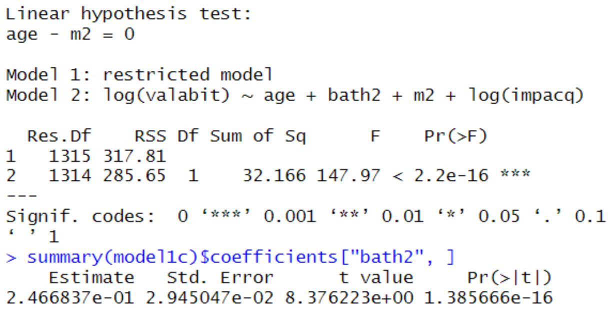


```
# Exercise 1d: Hypothesis tests
linearHypothesis(model1c, "age = m2")
summary(model1c)$coefficients["bath2", ]
```



Test: $H_0 : \beta_{\mathrm{age}} = \beta_{\mathrm{m2}}$

Result: F = 147.97, p-value < 2.2e-16

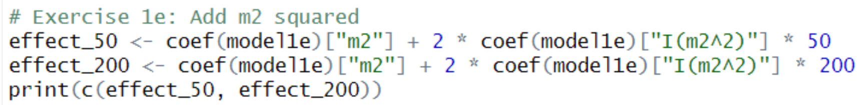


```
Linear hypothesis test:
age - m2 = 0

Model 1: restricted model
Model 2: log(valabit) ~ age + bath2 + m2 + log(impacq)

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1   1315 317.81                                  
2   1314 285.65  1    32.166 147.97 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1
> summary(model1c)$coefficients["bath2", ]
    Estimate   Std. Error      t value     Pr(>|t|) 
2.466837e-01 2.945047e-02 8.376223e+00 1.385666e-16 
```

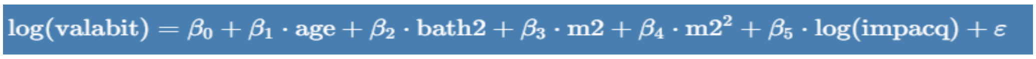
Here, I tested the effect of age and m2 on house value are the same. The F-test resulted a value of 147.97 and a p-value of less than 2.2e-16, which is highly significant. And for that reason, **we reject the null hypothesis** and conclude that the effect of age on houses is different from the effect of size m2.

Also, the impact of having two bathrooms compared to one or none. Here the coefficient of variable bath2 is 0.2467, which is highly significant. With p -value<2e16) that means that with the two bathrooms values are 24.7% higher than market value. So, **to conclude this, having two bathrooms increases the value of a house.**

EX- 1(E)

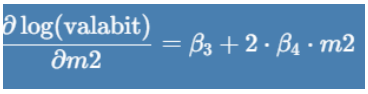


```
# Exercise 1e: Add m2 squared
effect_50 <- coef(model1e)["m2"] + 2 * coef(model1e)["I(m2^2)"] * 50
effect_200 <- coef(model1e)["m2"] + 2 * coef(model1e)["I(m2^2)"] * 200
print(c(effect_50, effect_200))
```



$$\log(\mathbf{valabit}) = \beta_0 + \beta_1 \cdot \mathbf{age} + \beta_2 \cdot \mathbf{bath2} + \beta_3 \cdot \mathbf{m2} + \beta_4 \cdot \mathbf{m2}^2 + \beta_5 \cdot \log(\mathbf{impacq}) + \varepsilon$$

To find the margin effect for surface at specific value we must use this formula

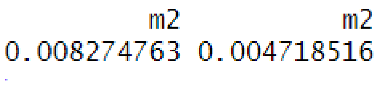


$$\frac{\partial \log(\mathrm{valabit})}{\partial m2} = \beta_3 + 2 \cdot \beta_4 \cdot m2$$

The house surface when it is equal to 50 m2

0.0827+2. (0.00472).50 = 0.00827+0.472 = 0.47185

0.00827+0.47185=0.480126



```
         m2          m2 
0.008274763 0.004718516 
```

The house surface when it is equal to 200 m2

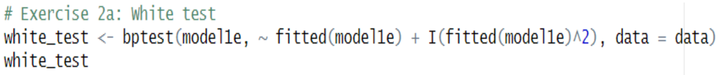
0.0827+2. (0.00472).200 = 0.00827+0.472 = 1.8874

0.0827476+1.8874=1.8956

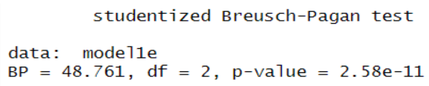
At 50 m2 a small increase in surface 48% and 200 m2 190%.**so the larger the house stronger impact.**

Ex – 2(A)

So, to run white test I am using **Breusch Pagen test** and to use that test you need to install or ("only if you already install it") called library **(lmtest).**



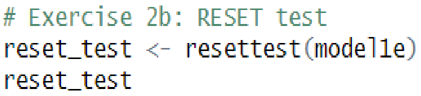
```
# Exercise 2a: White test
white_test <- bptest(model1e, ~ fitted(model1e) + I(fitted(model1e)^2), data = data)
white_test
```



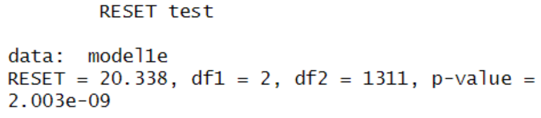
```
        studentized Breusch-Pagan test

data:  model1e
BP = 48.761, df = 2, p-value = 2.58e-11
```

The white test result (BP = 48.76, p-value = 2.58e-11 which is < 0.05) it shows a strong **heteroskedasticity** in the regression model. That means that residuals do not have continuous variance. And which unaccepted for one of the OLS assumptions. And for that we should use **the robust standard error to analyse**. there for **we reject the null hypothesis.**

Ex – 2(B)



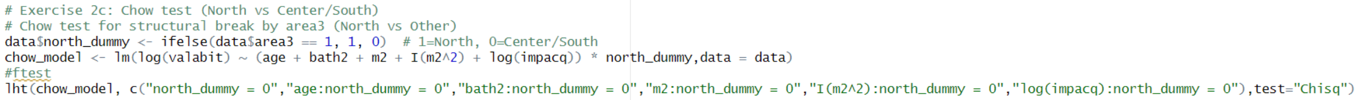
```
# Exercise 2b: RESET test
reset_test <- resettest(model1e)
reset_test
```



```
        RESET test

data:  model1e
RESET = 20.338, df1 = 2, df2 = 1311, p-value =
2.003e-09
```

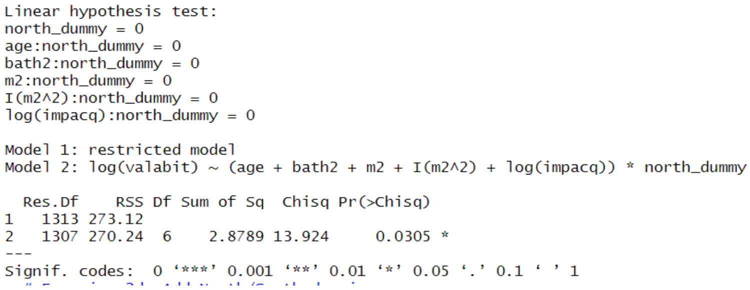
Reset =20.338 and the degrees of freedom for the test df1=2 and df2 = 1311 we compute the F-distribution. P-value = 2.003e-09<0.05) is smaller than the common significance. **We reject the null hypothesis.** That means that we must improve the model in the direction of **non-linear relationship or variables**.

Ex - 2(C)



```
# Exercise 2c: Chow test (North vs Center/South)
# Chow test for structural break by area3 (North vs Other)
data$north_dummy <- ifelse(data$area3 == 1, 1, 0)  # 1=North, 0=Center/South
chow_model <- lm(log(valabit) ~ (age + bath2 + m2 + I(m2^2) + log(impacq)) * north_dummy,data = data)
#ftest
lht(chow_model, c("north_dummy = 0","age:north_dummy = 0","bath2:north_dummy = 0","m2:north_dummy = 0","I(m2^2):north_dummy = 0","log(impacq):north_dummy = 0"),test="Chisq")
```

The chow test is used to check if the relationship between the independent variables and the dependent variables is the same in the two groups,

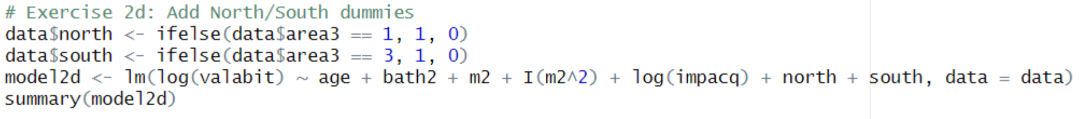


```
Linear hypothesis test:
north_dummy = 0
age:north_dummy = 0
bath2:north_dummy = 0
m2:north_dummy = 0
I(m2^2):north_dummy = 0
log(impacq):north_dummy = 0

Model 1: restricted model
Model 2: log(valabit) ~ (age + bath2 + m2 + I(m2^2) + log(impacq)) * north_dummy

  Res.Df    RSS Df Sum of Sq  Chisq Pr(>Chisq)
1   1313 273.12
2   1307 270.24  6    2.8789 13.924     0.0305 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the test = 13.92 and the p – value of the test is 0.0305 and the p =value is below 0.05 **we reject the null hypothesis** and the regression coefficient are the equal for both groups, so the effect of the age, bathroom, house, size(m2) and the price **(impacq)** on house value are different between the north and the centre/south regions.

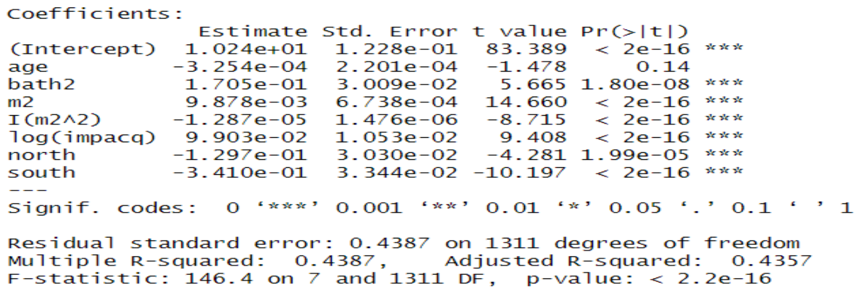
Ex – 2(D)

Here we use the dummy variables to represent different group north/south.



```
# Exercise 2d: Add North/South dummies
data$north <- ifelse(data$area3 == 1, 1, 0)
data$south <- ifelse(data$area3 == 3, 1, 0)
model2d <- lm(log(valabit) ~ age + bath2 + m2 + I(m2^2) + log(impacq) + north + south, data = data)
summary(model2d)
```

The model’s intercept represents log price **(log(valabit)** for a home in the centre.

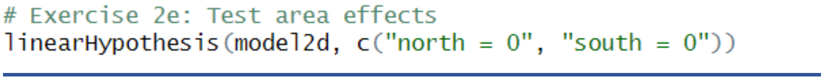


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.024e+01  1.228e-01  83.389  < 2e-16 ***
age         -3.254e-04  2.201e-04  -1.478     0.14    
bath2        1.705e-01  3.009e-02   5.665 1.80e-08 ***
m2           9.878e-03  6.738e-04  14.660  < 2e-16 ***
I(m2^2)     -1.287e-05  1.476e-06  -8.715  < 2e-16 ***
log(impacq)  9.903e-02  1.053e-02   9.408  < 2e-16 ***
north       -1.297e-01  3.030e-02  -4.281 1.99e-05 ***
south       -3.410e-01  3.344e-02 -10.197  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4387 on 1311 degrees of freedom
Multiple R-squared:  0.4387,	Adjusted R-squared:  0.4357 
F-statistic: 146.4 on 7 and 1311 DF,  p-value: < 2.2e-16
```

Here north = -0.129 that means that homes in the north are expected to have a **12.97% lower price** then homes in the centre.

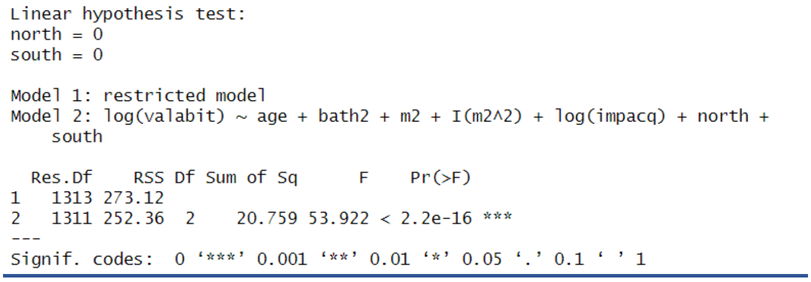
And in south = -0.3410 are 34.10**% lower** price then homes in the centre. And the reason why the dummy variable for centre is not included in the equation because **it's the reference category.**

Ex-2(E)



```
# Exercise 2e: Test area effects
linearHypothesis(model2d, c("north = 0", "south = 0"))
```

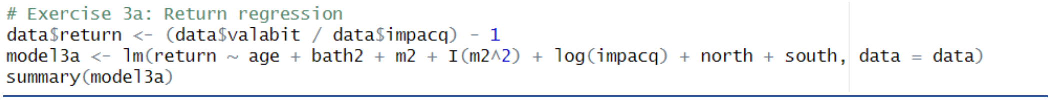
So here linear hypothesis test (f-test) here checks if they are dummies improve the regression. and the chow test separate regressions (north vs centre/south) and this test breaks between groups.



```
Linear hypothesis test:
north = 0
south = 0

Model 1: restricted model
Model 2: log(valabit) ~ age + bath2 + m2 + I(m2^2) + log(impacq) + north + 
    south

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1   1313 273.12                                  
2   1311 252.36  2    20.759 53.922 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

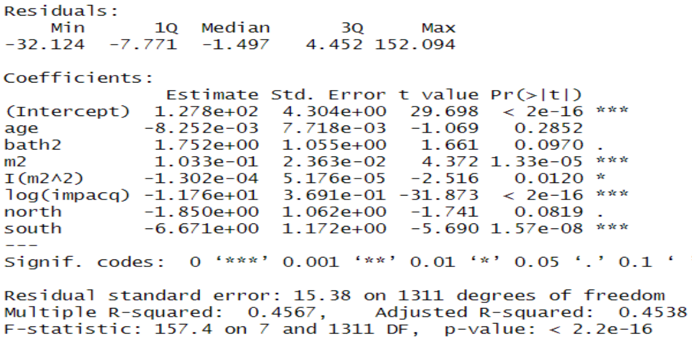
The null hypothesis for both the region north and south is zero which means that the region has no effect on the housing values. the outcome of the test shows f-value =53.92 with a p-value of<2.2e-16 and this is very small that means that **we reject the null hypothesis.** The test related to the chow test that we performed in Ex 2 (c). that shoes a **structural breck** between two groups by comparing separate regression for each group. On the other hand, the linear hypothesis test in Ex-2 (e) checks whether dummy variables for regions improve a single pooled model. In conclusion the chow test checks for entire model changes across groups, the f-test on dummies checks for level differences. **So, yes – in the sense that both tests are testing for group difference, but they test different things. And indeed, the test is not equivalent to chow test in 2c.**

Ex -3(A)



```
# Exercise 3a: Return regression
data$return <- (data$valabit / data$impacq) - 1
model3a <- lm(return ~ age + bath2 + m2 + I(m2^2) + log(impacq) + north + south, data = data)
summary(model3a)
```

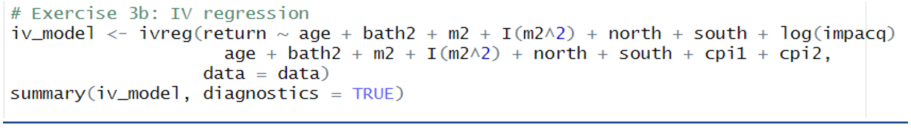
We defined a new dependent variable which represent the return on investment in the house values relative to the purchase price.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-32.124  -7.771  -1.497   4.452 152.094 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.278e+02  4.304e+00  29.698  < 2e-16 ***
age         -8.252e-03  7.718e-03  -1.069   0.2852    
bath2        1.752e+00  1.055e+00   1.661   0.0970 .  
m2           1.033e-01  2.363e-02   4.372 1.33e-05 ***
I(m2^2)     -1.302e-04  5.176e-05  -2.516   0.0120 *  
log(impacq) -1.176e+01  3.691e-01 -31.873  < 2e-16 ***
north       -1.850e+00  1.062e+00  -1.741   0.0819 .  
south       -6.671e+00  1.172e+00  -5.690 1.57e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 1311 degrees of freedom
Multiple R-squared:  0.4567,	Adjusted R-squared:  0.4538 
F-statistic: 157.4 on 7 and 1311 DF,  p-value: < 2.2e-16
```

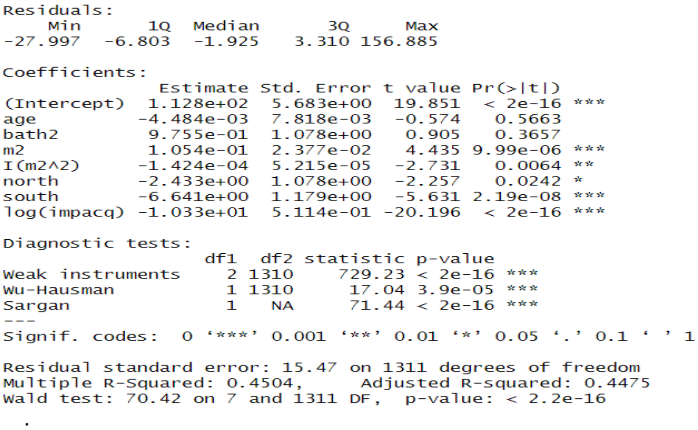
In the regression we see the dependent variable is defined as the return on investment **(Valabit/impacq)-1**. For that the coefficient on the house age is estimated at -0.825 which means that each additional year of the house age is associated with a decrease in return of approximately 0.825%. however, this effect is not that much significant(p=0.285). So that we do not have enough evidence to prove that that age of the house has reliable impact on the return model.

Ex – 3(B)



```
# Exercise 3b: IV regression
iv_model <- ivreg(return ~ age + bath2 + m2 + I(m2^2) + north + south + log(impacq)
                     age + bath2 + m2 + I(m2^2) + north + south + cpi1 + cpi2,
                  data = data)
summary(iv_model, diagnostics = TRUE)
```

In order to run this model you have to **install the package call "ivreg".**in this model we are trying to explain **return = (valabit/impacq)-1** as a function of house characteristics and log **(impaq)**.so here endogenous variable is **"log(impacq)"** exogenous variables is " **age","bath","m2","north","south"** and instruments **"cpi1","cpi2".**



```
Residuals:
    Min      1Q  Median      3Q     Max 
-27.997  -6.803  -1.925   3.310 156.885 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.128e+02  5.683e+00  19.851  < 2e-16 ***
age         -4.484e-03  7.818e-03  -0.574   0.5663    
bath2        9.755e-01  1.078e+00   0.905   0.3657    
m2           1.054e-01  2.377e-02   4.435 9.99e-06 ***
I(m2^2)     -1.424e-04  5.215e-05  -2.731   0.0064 ** 
north       -2.433e+00  1.078e+00  -2.257   0.0242 *  
south       -6.641e+00  1.179e+00  -5.631 2.19e-08 ***
log(impacq) -1.033e+01  5.114e-01 -20.196  < 2e-16 ***

Diagnostic tests:
                  df1  df2 statistic p-value    
Weak instruments    2 1310    729.23 < 2e-16 ***
Wu-Hausman          1 1310     17.04 3.9e-05 ***
Sargan              1   NA     71.44 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.47 on 1311 degrees of freedom
Multiple R-Squared: 0.4504,	Adjusted R-squared: 0.4475 
Wald test: 70.42 on 7 and 1311 DF,  p-value: < 2.2e-16 
  .
```

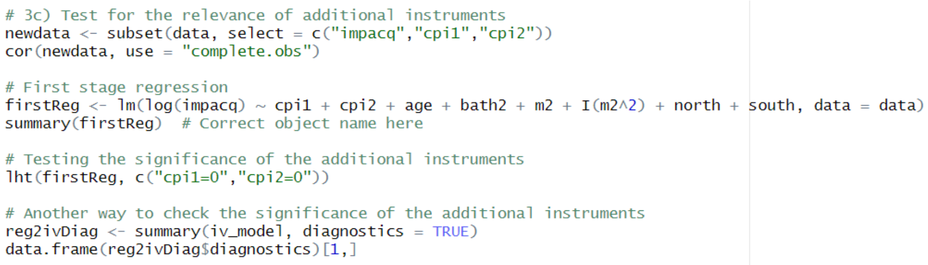
The model analysing house returns potentially suffers from endogeneity because the purchase price is included for both sides on the equation, directly as a regressor and indirectly thorough the dependent variable and the instrument variable represent the consumer price index at the year of purchase and the year before housing price by inflation levels at the time of purchase. Regression results show (F-stat = 729.23) and indicate that endogeneity is indeed present (Wu Hausman p = 3.9e-05)) and however, the Sargan test show as the concern **about instrument validity (p < 2e-16) which means that in further the exogeneity may be necessary. Overall, the IV model is justified in tums of endogeneity.**

Ex – 3(C)

$$\log(\text{impacq}) = \beta_0 + \beta_1 cpi1 + \beta_2 cpi2 + \ldots + \epsilon$$

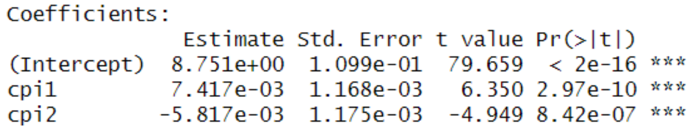
$$H_0 : \beta_{cpi1} = 0 \quad \text{and} \quad \beta_{cpi2} = 0$$

Test for the relevance and validity of the instruments.



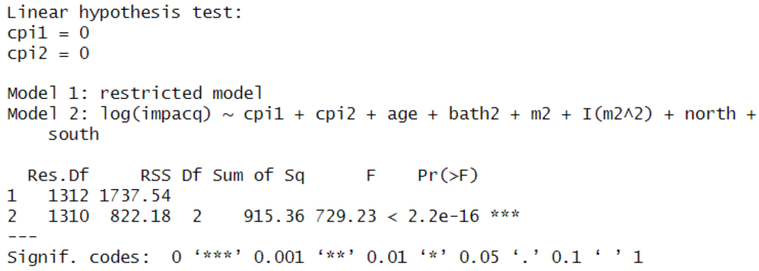
```
# 3c) Test for the relevance of additional instruments
newdata <- subset(data, select = c("impacq","cpi1","cpi2"))
cor(newdata, use = "complete.obs")

# First stage regression
firstReg <- lm(log(impacq) ~ cpi1 + cpi2 + age + bath2 + m2 + I(m2^2) + north + south, data = data)
summary(firstReg)  # Correct object name here

# Testing the significance of the additional instruments
lht(firstReg, c("cpi1=0","cpi2=0"))

# Another way to check the significance of the additional instruments
reg2ivDiag <- summary(iv_model, diagnostics = TRUE)
data.frame(reg2ivDiag$diagnostics)[1,]
```

So here we are testing whether additional instruments (“cpi1”,” cpi2”) are relevant for (IV) regression model.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.751e+00  1.099e-01  79.659  < 2e-16 ***
cpi1         7.417e-03  1.168e-03   6.350 2.97e-10 ***
cpi2        -5.817e-03  1.175e-03  -4.949 8.42e-07 ***
```

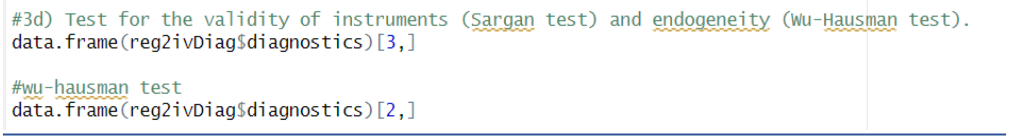
Both cpi1 and cpi2 coefficient are statistically significant at very high level. And for that these variables strongly predict **(log(impacq))** and they are endogenous regressor.in join significance test where we must use **lht()** (“Lenier hypothesis”)



```
Linear hypothesis test:
cpi1 = 0
cpi2 = 0

Model 1: restricted model
Model 2: log(impacq) ~ cpi1 + cpi2 + age + bath2 + m2 + I(m2^2) + north +
    south

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)    
1   1312 1737.54                                  
2   1310  822.18  2    915.36 729.23 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

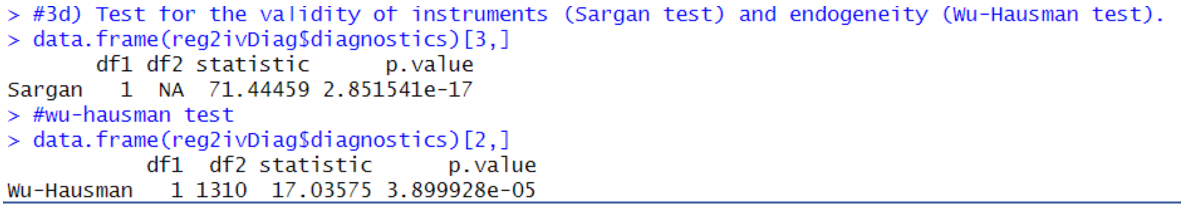
Here that both instrument coefficients are zero. And the result of F-statistic is very high (729.3) with p-value <2.2e-16 and **for that reject the null hypothesis**. And this conforms **that instruments are highly relevant for this model**. So overall these results are provided as very strong evidence that cpi1 and cpi2 are valid and powerful instruments and suitable for use in the IV estimation.

Ex – 3(D)



```
#3d) Test for the validity of instruments (Sargan test) and endogeneity (Wu-Hausman test).
data.frame(reg2ivDiag$diagnostics)[3,]

#wu-hausman test
data.frame(reg2ivDiag$diagnostics)[2,]
```

The Sargan test checks the validity of the instruments for more specific whether the instruments are uncorrelated with the error. And the WU-test use for the endogenous regressor.

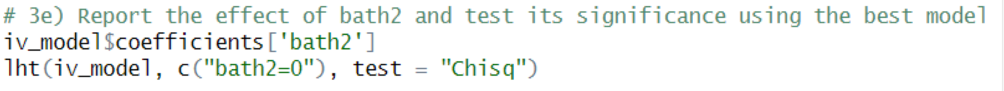


```
> #3d) Test for the validity of instruments (Sargan test) and endogeneity (Wu-Hausman test).
> data.frame(reg2ivDiag$diagnostics)[3,]
       df1 df2 statistic      p.value
Sargan   1  NA  71.44459 2.851541e-17
> #wu-hausman test
> data.frame(reg2ivDiag$diagnostics)[2,]
           df1  df2 statistic      p.value
Wu-Hausman   1 1310  17.03575 3.899928e-05
```

So, the test results for Sargan test are = 71.44 with very small p-value =2.84e-17 which means that **null hypothesis is rejected** and indeed that at least one instrument may be invalid or correlated with the error term.

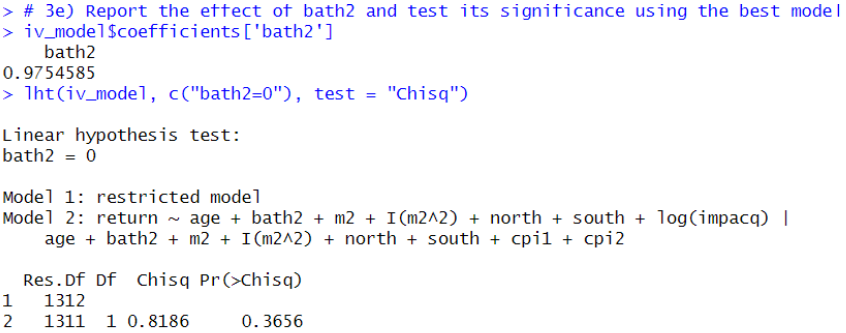
And the Wu-Hausman test 17.04 with p-value-3.9e-05 which means that the **null hypothesis rejected** and indeed endogeneity is present, and the OLS estimates would be biased and inconsistent. T**here for OLS is not appropriate. And Equation (6) should be estimated using IV.**

Ex – 3(E)



```
# 3e) Report the effect of bath2 and test its significance using the best model
iv_model$coefficients['bath2']
lht(iv_model, c("bath2=0"), test = "Chisq")
```

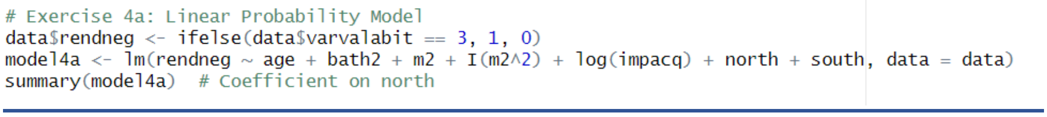
We are using chi-square test.



```
> # 3e) Report the effect of bath2 and test its significance using the best model
> iv_model$coefficients['bath2']
    bath2
0.9754585
> lht(iv_model, c("bath2=0"), test = "Chisq")

Linear hypothesis test:
bath2 = 0

Model 1: restricted model
Model 2: return ~ age + bath2 + m2 + I(m2^2) + north + south + log(impacq) |
    age + bath2 + m2 + I(m2^2) + north + south + cpi1 + cpi2

  Res.Df Df  Chisq Pr(>Chisq)
1   1312
2   1311  1 0.8186     0.3656
```

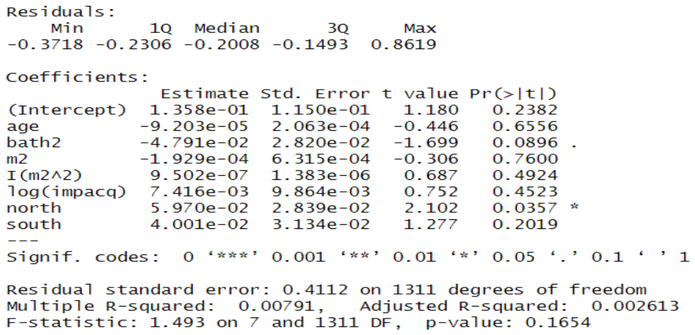
So, the coefficient for the bath2 = 0.975, which means one unit increase in bath2 is increase 0.975 unites in the dependent variable (return). So here we performed the **chi-square test** and the and the chi-square test =0.819 with corresponding p-value 0.366 and since the p-value is much higher then (0.05) **We fail to reject the null hypothesis**. and this model so the effect of bath2 is not significant in this IV modal and for that reason we don't have sufficient evidence that bath2 has a meaningful effect on return.

Ex – 4(A)



```
# Exercise 4a: Linear Probability Model
data$rendneg <- ifelse(data$varvalabit == 3, 1, 0)
model4a <- lm(rendneg ~ age + bath2 + m2 + I(m2^2) + log(impacq) + north + south, data = data)
summary(model4a)  # Coefficient on north
```

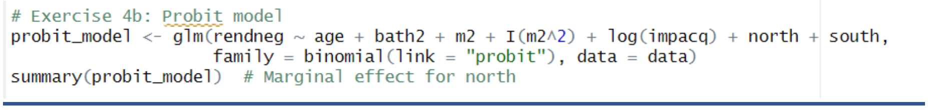
The OLS estimator coefficient on north in the regression model predicting **rendneg** a dummy variable shows whether the future returns on investment are negative.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-0.3718 -0.2306 -0.2008 -0.1493  0.8619 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  1.358e-01  1.150e-01   1.180   0.2382  
age         -9.203e-05  2.063e-04  -0.446   0.6556  
bath2       -4.791e-02  2.820e-02  -1.699   0.0896 .
m2          -1.929e-04  6.315e-04  -0.306   0.7600  
I(m2^2)      9.502e-07  1.383e-06   0.687   0.4924  
log(impacq)  7.416e-03  9.864e-03   0.752   0.4523  
north        5.970e-02  2.839e-02   2.102   0.0357 *
south        4.001e-02  3.134e-02   1.277   0.2019  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4112 on 1311 degrees of freedom
Multiple R-squared:  0.00791,	Adjusted R-squared:  0.002613 
F-statistic: 1.493 on 7 and 1311 DF,  p-value: 0.1654
```

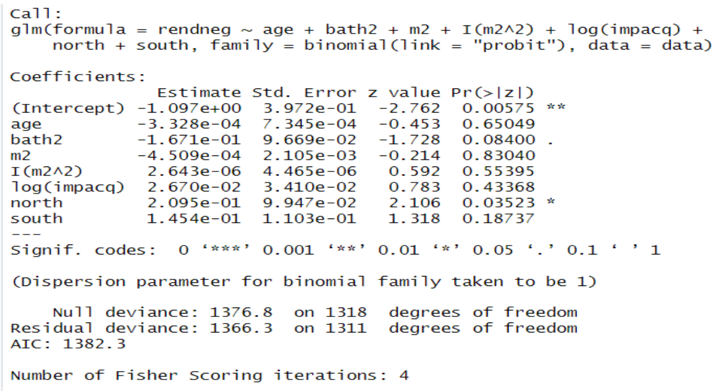
So, coefficient for the north = 0.057 with std 0.0284 the t-value 2.10 and the p-value0.0357 which is less than 0.05 and that means that being in the north region is associated with increase of about 5.97% in the point that there is a probability that future returns on investment will be negative. And since the **rendneg** is a binary variable (0 or 1) the coefficient can be interpreted as the marginal effect on the negative returns.so according to this data we can say that there is high probability in the futures that being un the north gives you negative returns.

Ex- 4(B)



```
# Exercise 4b: Probit model
probit_model <- glm(rendneg ~ age + bath2 + m2 + I(m2^2) + log(impacq) + north + south,
                    family = binomial(link = "probit"), data = data)
summary(probit_model)  # Marginal effect for north
```

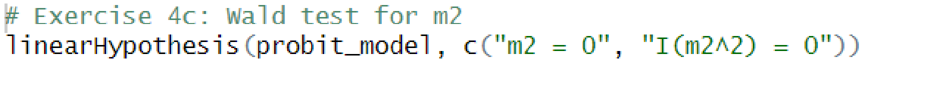
Probit model



```
Call:
glm(formula = rendneg ~ age + bath2 + m2 + I(m2^2) + log(impacq) + 
    north + south, family = binomial(link = "probit"), data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)   
(Intercept) -1.097e+00  3.972e-01  -2.762  0.00575 **
age         -3.328e-04  7.345e-04  -0.453  0.65049   
bath2       -1.671e-01  9.669e-02  -1.728  0.08400 . 
m2          -4.509e-04  2.105e-03  -0.214  0.83040   
I(m2^2)      2.643e-06  4.465e-06   0.592  0.55395   
log(impacq)  2.670e-02  3.410e-02   0.783  0.43368   
north        2.095e-01  9.947e-02   2.106  0.03523 * 
south        1.454e-01  1.103e-01   1.318  0.18737   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1376.8  on 1318  degrees of freedom
Residual deviance: 1366.3  on 1311  degrees of freedom
AIC: 1382.3

Number of Fisher Scoring iterations: 4
```

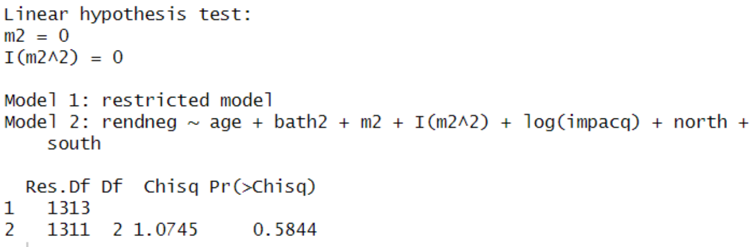
So, here using the probit model the coefficient on the variable north is 0.0209 with Std 0.0995 and z – value 2.106 and a p – value 0.035. and this indicated that being in the north region has a positive effect and that means the future returns on investment will be negative. The positive sign of the coefficient means the north location increases the latent propensity so the binary outcome **rendnge = 1 and the test reject the null hypothesis** that the effect of the north is zero. Also, the alternative that **(H1):** The coefficient on north is not zero

Ex – 4(C)



```
# Exercise 4c: Wald test for m2
linearHypothesis(probit_model, c("m2 = 0", "I(m2^2) = 0"))
```

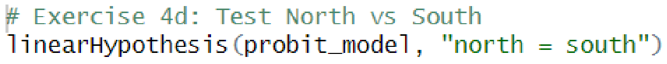
Here we test whether house surface is (m2) and its square (m2^2) have significant effect on the properties of negative investments returns.



```
Linear hypothesis test:
m2 = 0
I(m2^2) = 0

Model 1: restricted model
Model 2: rendneg ~ age + bath2 + m2 + I(m2^2) + log(impacq) + north +
    south

  Res.Df Df  Chisq Pr(>Chisq)
1   1313
2   1311  2 1.0745     0.5844
```

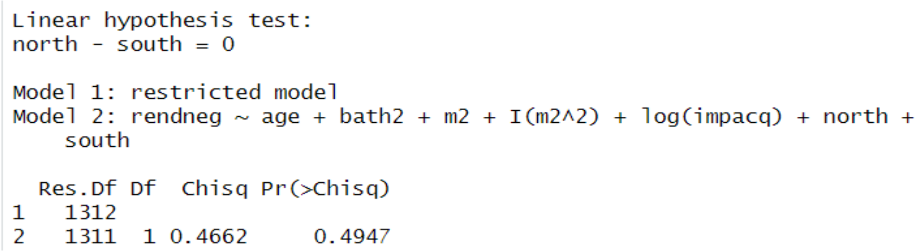
And this test we can solve by using the joint hypothesis test where both coefficients set equal to zero. And the chi-squared test =1.0745 with 2 degrees of freedom and the p-value = 0.5844 and **the p – values is very large We fail to reject the null hypothesis**. This means that house surface does not have effect on the likelihood of negative investment returns in the model.

Ex – 4(D)



```
# Exercise 4d: Test North vs South
linearHypothesis(probit_model, "north = south")
```

The test whether the effects of living in the north and south identify. Here we conducted a linear hypothesis test with the null hypothesis north – south = 0.

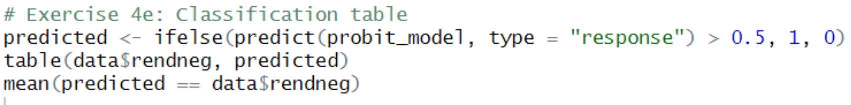


```
Linear hypothesis test:
north - south = 0

Model 1: restricted model
Model 2: rendneg ~ age + bath2 + m2 + I(m2^2) + log(impacq) + north +
    south

  Res.Df Df  Chisq Pr(>Chisq)
1   1312
2   1311  1 0.4662     0.4947
```

The chi -squared test is = 0.4662 with 1 degree of freedom and the p-value = 0.04947, since the p-values is higher than (0.05) **We fail to reject the null hypothesis**.

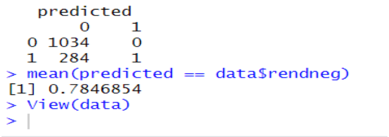
That means that there is no significant difference between the effects of living in the north and south on the probability of expecting the negative return.

Ex – 4(E)



```
# Exercise 4e: Classification table
predicted <- ifelse(predict(probit_model, type = "response") > 0.5, 1, 0)
table(data$rendneg, predicted)
mean(predicted == data$rendneg)
```

Classification



```
   predicted
       0    1
  0 1034    0
  1  284    1
> mean(predicted == data$rendneg)
[1] 0.7846854
> View(data)
> 
```

In this model, a total of 1,035 out of 1,319 observations were correctly predicted where the model correctly identified an outcome of 0, and casa 1 where it correctly predicted. `Modal correctly classification accuracy 78.47%.