6/11/2025

TEGEGN TISE MOKONA

University of Verona

Predicting House Prices Using Advanced Econometric Modeling in R

**Project Summary**

This project analyzes how various housing and regional characteristics influence self-reported house values using **Italian household survey data**. Employing advanced econometric techniques in **R**, the study explores the effects of variables such as **house age, size, number of bathrooms, and purchase price** on property value.

The analysis began with **OLS and log-linear models** and progressed to more complex estimations, including:

* **Quadratic and interaction terms** to capture nonlinear relationships.
* **White and RESET tests** for model diagnostics and specification checks.
* **Chow tests** to detect structural breaks across regions (North vs. South).
* **Instrumental Variable (IV) regression** using CPI-based instruments to correct for endogeneity in purchase price.
* **Probit model** to estimate the probability of negative returns on housing investment.

**Key Findings:**

* Larger homes have higher value, but marginal gains diminish as size increases.
* Purchase price significantly predicts current market value.
* Regional disparities exist: homes in Northern Italy tend to have different price dynamics than those in the South.
* The IV model confirms that ignoring endogeneity biases estimates.
* The Probit model achieved **77.26% prediction accuracy** for identifying negative return expectations.

This project demonstrates mastery of **applied econometric analysis**, **model testing**, and **causal inference**, bridging theory and real-world data in housing economics.

***Preliminary Set-up***

First, I load data into the R software by using following code:

A close up of a word

AI-generated content may be incorrect.

Then, I add some required libraries as by using following command

**>*library(car)***

**>*library(lmtest)***

**>*library(AER)***

**>*library(MASS)***

# ***EXERCISE 1***

1. **OLS Regression of valabit on Age, Bathrooms, and Size**

***valabit*  1  2*age*  3*bath*2  4*m*2  **

To answer this, I start by defining two new variables introduced here, which did not exist in the data before, as follows:

* To create “Age” variable, I used following command to deduct year of construction from survey year.

**> Dataset$age = 2014 - Dataset$ancostr**

* To create dummy variable “bath2”, I used following command

**> Dataset$bath2 = ifelse(Dataset$bagni = = 2, 1, 0)**

Then, I run first OLS regression named “Reg1”, and the command is following

**> Reg1 = lm(valabit ~ age + bath2 + m2, data = Dataset)**

**> summary (Reg1)**

Then, the summary of this regression is provided as follows:

A screenshot of a computer

AI-generated content may be incorrect.

The coefficient of age variable is 179.04, and the corresponding t-value is t-test = 2.093. The coefficient of age variable is significant because the absolute value of t-test is higher than critical value of 1.96 at 5% significance level. From this we can conclude that, Age contributes to determining the house value. In fact, keeping other things constant, each additional year of age increases the estimated house value by €179.

1. **Log-linear Model of valabit**

**ln *valabit*   1  2*age*  3*bath*2  4*m*2  **

I run this log-linear regression by using following command and named it as “Reg2”

***> Reg2 = lm(log(valabit) ~ age + bath2 + m2, data = Dataset)***

***> summary (Reg2)***

Then, the summary of this regression is provided as follows:

A screenshot of a computer

AI-generated content may be incorrect.

In the first regression (Reg1), the coefficient for the surface area of House (m²) is 1950.22, and the t-test value is 23.153, which significantly exceeds the critical t-statistic value of 196 in absolute terms. Thus, the effect of house surface (m²) is significant, and for every additional square meter increase in housing surface, the self-reported market value of the house increased by 1,950 euros, keeping other things constant.

***Growth Rate:*** In regression 2 (Reg2 log-linear model), the coefficient of surface of House(M2) is 0.00497 with a t-test value of 17.483, which is significantly higher than the critical value of t-statistic, 196, in absolute value. Again, the effect of house surface (M²) is significant, and holding other factors constant, one additional square meter increment in the housing surface is associated with an approximately 0.497% increase in the self-evaluated current value of the house.

1. **Including Log Purchase Price**

I run this log-linear regression by using following command and named it as “Reg3”

***> Reg3 = lm(log(valabit) ~ age + bath2 + m2 + log(impacq), data = Dataset)***

***> summary (Reg3)***

Then, the summary of this regression is provided as follows:

A screenshot of a computer

AI-generated content may be incorrect.

***Elasticity***: The coefficient of the newly added variable, log of the purchase price of the house, ***log(impacq)*** is 0.1102 with a T-test value = 10.214, which, in absolute value, is significantly higher than the critical value at the 5% significance level, 1.96. Thus, the effect of the purchase price of the house is significant, and keeping other factors constant, a 1% increase in the original purchase price of the house is associated with a 0.11% increase in the current self-evaluated value of the house.

1. **Hypothesis Testing**

In order to check if the effects of age and m2 are the same are the same, I run an F-test. The null and the alternative hypothesis are set as follow:

***Hypothesis Test 1: Are the effects of age and m2 the same?***

***Null Hypothesis (H₀): β\_age = β\_m2***

***Alternative Hypothesis (H₁): β\_age ≠ β\_m2***

To conduct this hypothesis testing, I have used the following command:

***>* linearHypothesis(Reg3, "age = m2")**

The result of this command is:

A screenshot of a computer

AI-generated content may be incorrect.

The result shows an F-value of 100.9 with a P-value of 2.2\*10-16, which is almost equal to Zero. The test follows Chi-squared distribution of parameter 1, whose critical value is 3.841. Since 100.9 > 3.841 and the p-value is smaller than 0.05, we can reject the null hypothesis and can conclude that the effects of age and m² are not the same.

***Hypothesis Test 2: Effect of Two Bathrooms***

In order to check this, I have conducted the significance of the coefficient of the newly added variable “bath2” as follows.

Null Hypothesis (H₀): β\_bath2 = 0

Alternative Hypothesis (H₁): β\_bath2 ≠ 0

To conduct this hypothesis testing, I have used the following code:

***>*linearHypothesis(Reg3, "bath2 = 0")**

The result of this command is:

A computer screen shot of a math equation

AI-generated content may be incorrect.

The result shows that we have an F-value of 68.894 and a P-value of 2.551\*10-16, which is almost equal to Zero. The test follows Chi-squared distribution of parameter 1, whose critical value is 3.841. Since 68.894 > 3.841 and the p-value is smaller than 0.05, we can reject the null hypothesis and can conclude that there is a significant difference in value for the houses with two bathrooms and with zero or one bathroom.

1. **Including Quadratic Term**

***ln valabit   1  2 age  3bath2  4 m2  5 m2^2 6 ln impacq  ***

From the above equation, the marginal effect of the house surface on the value of the house is partial derivative of the above equation w.r.t m2, and provided as follows:

Thus, to exact value of the marginal effect of the house surface on the value of the house, I need the value of the  **and .** To get the value of these parameters I run regression with the equation 4 format. First before I run this regression equation, I add the new variable “(M2)2 into data set, and named it as “m2\_Square” by using following command:

**>Dataset$m2\_square = Dataset$m2^2**

Then, I used the following command to run this regression equation and named it as “Reg4”

***>Reg4 = lm(log(valabit)~age + bath2 + m2 + m2\_square +log (impacq), data = Dataset)***

***>summary (Reg4)***

Then, the summary of this regression is provided as follows:A screenshot of a computer

AI-generated content may be incorrect.

As can be clearly seen in the above summary of Reg4, the value of   = 0.008005646 with t-value = 11.204 > 1.96, and  = -0.000007955 with t-value = |-4.890|>1.96.  These two coefficients are highly significant, as their t-value is significantly higher than the critical value of t at the 5% significance level. Then, I Compute the marginal effect of House Surface, when the house surface is 50 m² and 200 m² as follows:

***The effect\_When\_Surface\_50m2 = B4 + 2 \* B5 \* 50 = 0.008005646 + 2\*(-0.000007955\*50) = 0.00721013***

***The effect\_When\_Surface\_200m2 = B4 + 2\*B5\*200 = 0.008005646 + 2\*(-0.000007955\*200) = 0.004823581***

Thus, at 50 m², every additional increase in housing surface(M2) is associated with 0.72% increases in self-evaluated housing price, while at 200 m² this marginal effect dropped to only a 0.48% increment in housing price. While adding surface has a positive effect on housing price, The negative coefficient at (M2) shows that marginal benefit declines as the size of the house increases.

***EXERCISE 2***

On this exercise, we focus on the regression in Equation (4).

1. **Run a White test (approximate variant). What do you conclude?**

I used the following code to run the white test:

***>resi = residuals(Reg4)***

***>yhat = predict(Reg4)***

***>reg\_white = lm(I(resi^2) ~ yhat + I(yhat^2), data = Dataset)***

***>white\_stat = nrow(Dataset) \* summary(reg\_white)$r.squared***

***>white\_stat***

***>pchisq(white\_stat, df = 2, lower.tail = FALSE***

From this test, the resulting test statistic was 39.19 with a p-value of 3.09 × 10⁻⁹. Since the p-value is well below 0.05, we reject the null hypothesis of homoskedasticity and conclude that the model exhibits heteroskedasticity.

1. **Run a RESET test. What do you conclude?**

RESET Test – Hypothesis Statement

**Null Hypothesis (H₀)**: The model is correctly specified. That is, the regression includes all relevant variables in the correct functional form:

**H₀: β\_fitted² = β\_fitted³ = 0**

**Alternative Hypothesis (H₁)**: The model is mis specified — at least one nonlinear transformation of the fitted values helps explain the dependent variable:

**H₁: At least one of β\_fitted² or β\_fitted³ ≠ 0**

I used the following code to run this test for model specification.

***>resettest(Reg4, power = 2:3, type = "fitted")***

From this code, the RESET test statistic is 2.89 with a p-value of 0.056. Although, I can’t fully reject null hypothesis at 5% threshold for significance, I fully reject the null hypothesis at the 10% level, suggesting the model may suffer from misspecification.

1. **Run a Chow test comparing two groups of individuals living in the North, or in the Center/South respectively. What do you conclude?**

I have used the following series of commands to run this Chow comparing the difference in two regions.

First, I Created a dummy for region by using following command.

**>Dataset$region\_dummy = ifelse(Dataset$area3 == 1, 1, 0)**

Then, I have run following a model with interaction terms to allow different coefficients by region.

**reg5 = lm(log(valabit) ~ age + bath2 + m2 + m2\_square + log(impacq) +**

**region\_dummy + age:region\_dummy + bath2:region\_dummy + m2:region\_dummy + m2\_square:region\_dummy + log(impacq):region\_dummy, data = Dataset)**

Then, I have used linearHypothesis() (Chow test) to jointly test if interaction terms are zero (no structural break).

***>Chow\_Test = linearHypothesis(reg5, c("region\_dummy = 0", "age:region\_dummy = 0", "bath2:region\_dummy = 0","m2:region\_dummy = 0","m2\_square:region\_dummy = 0","log(impacq):region\_dummy = 0"), test = "Chisq")***

The summary of this test is provided as follows:

A white background with black text

AI-generated content may be incorrect.

As it can be seen in the result, The test returned a statistically significant result (Chi-squared = 18.89, p-value = 0.0043), providing strong evidence of a structural break between the two groups. I, therefore, conclude that there is a statistically significant difference in the relationship between housing characteristics and perceived house value across the two regional groups.

1. **Add Area Dummies — North and South, in Equation4 (Reg4)**

The Center is excluded from the equation to avoid the **dummy variable trap** (perfect multicollinearity). By doing so, it becomes the **reference group**, allowing the coefficients of the North and South dummies to be interpreted as differences relative to the Center.

1. **Test for the presence of area effects in the regression in Equation (5). Is this test equivalent to the Chow test you ran in 2c)?**

I used the following code to Create area dummies, and run new regression and named it as “Reg5” as follows:

***>Dataset$north = ifelse(Dataset$area3 == 1, 1, 0)***

***>Dataset$south = ifelse(Dataset$area3 == 3, 1, 0)***

***>Reg5 = lm(log(valabit) ~ age + bath2 + m2 + m2\_squared + log(impacq) + north + south, data = Dataset)***

I used the following code to test whether the north and south dummies are jointly significant in explaining house value.

***>linearHypothesis(Reg5, c("north = 0", "south = 0"))***

The result of this test:

A white background with black text

AI-generated content may be incorrect.

The joint F-test yielded a statistic of 41.63 with a p-value < 2.2e-16. From this result, we strongly reject the null hypothesis that both coefficients are zero, concluding that geographic area has a significant effect on perceived house value. This result is equivalent to the Chow test in Exercise 2(c).

**EXERCISE 3**

1. **Define the new dependent variable return and run the regression**

I have used the following code to create a new variable with given formula:

***>Dataset$return = (Dataset$valabit / Dataset$impacq) - 1***

Then, I run OLS regression on return as follows.

**>Reg6 = lm(return ~ age + bath2 + m2 + I(m2^2) + log(impacq) + north + south, data = Dataset)**

**>summary(Reg6)**

The summary of this regression is provided as follows:

A screenshot of a computer program

AI-generated content may be incorrect.

From this regression, the coefficient of the age variable is 0.0008785(nearly zero), with a t-value of 0.102<1.96. Since the t-value is less than the critical value of 1.96, this coefficient is not statistically significant.

1. **Instrumental Variable (IV) Setup: CPI as IV for log(impact)**

In this model, the endogenous variable is log(impacq) because it appears on both sides of the equation, making it potentially correlated with the error term. The exogenous variables include north, south, bath2, m2, m2², and age. The instruments used — cpi1 and cpi2 — are considered **relevant** because they are strongly correlated with the endogenous variable (log(impacq)), as they influence the purchase price through inflation and changes in the general price level. Moreover, they are considered **valid** instruments because they are unlikely to affect the return on investment directly and are assumed to be uncorrelated with the error term in the regression.

1. **Test for Instrument Relevance and Validity**

**I have used the following command to run IV model and check the relevance and validity of the instruments.**

***>*** ***Reg7\_iv\_model = ivreg(return ~ age + bath2 + m2 + m2\_squared + north + south + log(impacq) |age + bath2 + m2 + m2\_squared + north + south + cpi1 + cpi2, data= Dataset)***

***>summary(Reg7\_iv\_model, diagnostics = TRUE)***

The summary output of this code:

A computer screen shot of numbers and symbols

AI-generated content may be incorrect.

As can be seen in the above result, The first-stage regression confirms strong relevance of the instruments (cpi1 and cpi2) for log(impacq) (Weak instrument test p < 0.001). However, the Sargan test strongly rejects the validity of the instruments (p < 0.001), suggesting that one or both instruments may be correlated with the error term, undermining the exogeneity assumption. We therefore can conclude that the instruments are relevant, but at least one of them is not valid.

1. **Based on your previous answers, should Equation (6) be estimated with OLS or IV? Use an appropriate test if necessary.**

In order to decide, I have checked the result of a Wu-Hausman test was performed to test the endogeneity of the regressor log(impacq). The null hypothesis of exogeneity was strongly rejected (p-value < 0.001), indicating that log(impacq) is endogenous. We therefore can conclude that the use of Instrumental Variables (IV) estimation is appropriate rather than OLS to avoid biased and inconsistent estimates.

1. **Using the best model (OLS or IV), report the effect of *bath2*. Is the coefficient significant?**

**Since the appropriate model to get un**biased and consistent estimates is IV model, I have used the following code to re-run the model and report the effect of having 2 bathrooms(bath2) on the self-evaluated housing price.

# Run the IV regression (2SLS) using ivreg from the AER package

***>*** **Reg8\_iv\_model = ivreg(return ~ age + bath2 + m2 + m2\_squared + north + south + log(impacq) | age + bath2 + m2 + m2\_squared + north + south + cpi1 + cpi2,**

**data = Dataset)**

***>*** **summary(Reg8\_iv\_model)**

Then, I have used following command to Test significance of bath2 coefficient

***>linearHypothesis(Reg8\_iv\_model, "bath2 = 0")***

Summary of this IV-reg:

A screenshot of a computer program

AI-generated content may be incorrect.

By Using the preferred IV estimation method, the coefficient of bath2 is estimated at 2.112 with a corresponding t-statistic of 2.131, which is higher than the critical value of 1.96 at the 5% significance level. Thus, we can conclude that bath2 has a positive and statistically significant effect on the dependent variable at the 5% significance level.

**EXERCISE 4**

1. **Create rendneg and run OLS regression**

I have used the following command to create a new dummy variable.

***> Dataset$rendneg = ifelse(Dataset$varvalabit == 3, 1, 0)***

Then, I used the following command to run OLS regression and named it as “Reg7”:

**> Reg7 = lm(rendneg ~ age + bath2 + m2 + m2\_squared + log(impacq) + north + south, data = Dataset)**

**> summary(Reg7)**

The summary of this OLS regression is provided as follows.

A screenshot of a computer

AI-generated content may be incorrect.

From this result, the coefficient for the variable “North” is 0.0595 with t-value 2.016, which is higher than critical value of 1.96 at 5% significance level, and we can conclude that the effect is statistically significant and being located in the North significantly increases the likelihood of expecting lower future returns.

1. **Estimate Equation (7) using Probit**

I used the following command to run this probit model:

**> Reg7\_probit = glm(rendneg ~ age + bath2 + m2 + m2\_squared + log(impacq) + north + south, family = binomial(link = "probit"), data = Dataset)**

**> summary(Reg7\_probit)**

The summary of this Probit regression is provided as follows.

A screenshot of a computer

AI-generated content may be incorrect.

In this probit estimation, the estimated coefficient for north is 0.2020, with z-value is 2.015, and the p-value is 0.04386, which is less than 0.05 threshold. This means the effect of being in the North is statistically significant at the 5% level. Since the coefficient of the north is positive and significant, we can conclude that living in the North is associated with a higher likelihood of expecting negative future returns on investment.

1. **Test for the significance of the house surface in the probit model.**

To assess the significance of **the house surface** in the probit model, I have used a **joint hypothesis test.**

**Hypotheses**

* **Null hypothesis (H₀): The coefficients on m2 and I(m2^2) are both equal to zero**
* **Alternative hypothesis (H₁): At least one of the coefficients is not equal to zero**

Then, I have used the following command to run this hypothesis testing:

**> linearHypothesis(Reg7\_probit, c("m2 = 0", "m2\_squared = 0"))**

The result of this command is provided as follows:

A screenshot of a computer

AI-generated content may be incorrect.

The test yields a Chi-squared statistic of 4.8072 with a p-value of 0.09039, which is greater than the 5% significance level. Therefore, we fail to reject the null hypothesis and conclude that house surface (measured by m2 and its square) is not statistically significant in explaining the likelihood of expecting negative returns on investment.

1. **Test if the effects of North and South are identical.**

To test whether the coefficients on the dummy variables north and south are statistically equal in the probit model, I have used Wald test as follows:

**Hypotheses**

* *Null hypothesis (H₀): The coefficients on North and South are the same*
* *Alternative hypothesis (H₁): The coefficients on North and South are the different.*

To run this hypothesis, I have used the following command:

***> linearHypothesis(Reg7\_probit, "north = south")***

The result of this command is provided as follows:

A math equations on a white background

AI-generated content may be incorrect.

The test yields a Chi-squared statistic of 1.0744 with a p-value of 0.30, which is **greater** than the 5% significance level. Since the p-value (0.30) is greater than 0.05, we fail to reject the null hypothesis. Therefore, we conclude that there is no statistically significant difference between the effects of living in the North and South on the probability of expecting negative returns on investment.

1. **How many of the observations are correctly predicted in the probit model?**

I have used the following command to check this

***> probs = predict(Reg7\_probit, type = "response")***

***> predicted\_class = ifelse(probs > 0.5, 1, 0)***

***> conf\_matrix = table(Predicted = predicted\_class, Actual = Dataset$rendneg)***

***> accuracy = sum(diag(conf\_matrix)) / sum(conf\_matrix)***

***> cat("Proportion correctly predicted:", round(accuracy, 4), "\n")***

Thus, the result shows that **77.26%** of the observations were correctly predicted by the probit model.