Project in multiple regression and house pricing

Esben Høg *Aalborg Universitet*

Introduction

In the 5th semester the project deals with statistics, primarily the so-called *linear models*. Typically aims are to use methods from statistical theory to do estimation and more generally statistical inference based on real data.

Ultimately you should solve a real problem based on analyzing a dataset, i.e. interpret and apply statistics (and probability) theory on the real world problem in question.

Linear models and statistical inference

A simplified decription of the situation is as follows:

You have a so-called **response variable**, i.e. a variable of primary interest, which is sought to be explained by a number of other variables, the so-called **explanatory variables**. Another often used term for the response variable is the **dependent variable**. The relation between the variables is something like

$$y = f(x_1, x_2, \ldots, x_k).$$

Here y denotes the response and $x_1, \ldots x_k$ denote some k explanatory variables. Theoretically, one might expect the response variable to be a function of the explanatory variables as above, but such a relation is only an approximation. There will also be "noise".

If the functional relationship is linear and we add noise, based on the data, we might have a relationship like the following

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i,$$

where the x_i's and y_i's are observed values of relevant variables in the data set, and ε_i is the i'th (unobserved) noise term. We assume here that we have, say, n observations of the variables, i.e. $i = 1, 2, \ldots, n$. Typically the noise terms are normal distributed random variables with mean zero and of course some variance, and possibly other assumptions which naturally could be discussed in a project.

Econometrics

The special type of linear model, given above, from which we will start here, is more generally called **the multiple linear regression model**. It is a special case of a general linear model in statistics. In economics (or econometrics) the multiple linear regression model is sort of the foundation of **econometrics** or **an econometric model**.

Mathematically and statistically, there are a lot of interesting problems and properties associated with such models, such as described above (and it is, among other things, your task to analyze it more closely).

Price formation on the housing market

The concrete project proposed in this document deals with price formation in the Danish housing market. Your task is to construct, estimate and interpret a model that can be used to *predict* the selling price of a house based on information about the characteristics of the house and possibly information on locational characteristics as well. Such models are very important for the real estate industry, the authorities and the financial institutions that provide loans for the home.

Hedonic pricing

There is a special name for pricing models for houses/homes, namely a so-called hedonic pricing model or hedonic regression. Hedonic pricing is a direct application of multiple regression and econometrics.

Data

If you want to work with this concrete project you have access to a data set originally provided by the Danish real-estate organization HOME.

The data set contains a very large number of sales prices for homes in the four largest cities (*kommuner*) in Denmark together with a lot of other variables that describe various characteristics of the homes. It contains virtually all trades via HOME for houses (villas) in these cities during the period from 2010 to the summer of 2022.

The idea is that the theory of linear models, multiple regression and econometrics all come together with data, so that you can do an analysis of the price formation. Is it possible to predict price appropriately based on characteristics?

Possible aims

- Identify which variables have an effect on the price and how important these effects are.
- Do some diagnostic checking, also called model control: Are the assumptions used to estimate reasonably fulfilled?
- Interpret these effects in a meaningful (i.e. economic) way
- Investigate the differences w.r.t. price formation between the cities
- Verify that the effects of explanatory variables have the "correct" signs
- Is there any Covid19 effect on the housing market?