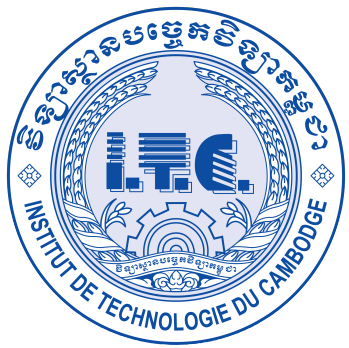
Regression Analysis

Chapter 01
Introduction

Phauk Sokkhey
phauk.sokkhey@itc.edu.kh
Nhim Malai
nhim.malai@itc.edu.kh

Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia

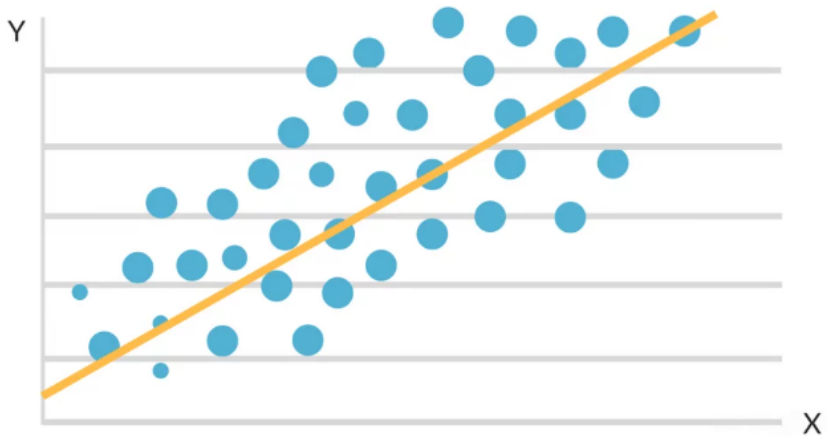


Contents

Introduction

What is Regression?

Regression is a statistical technique used to study the relationship between independent and dependent variables.
In machine learning, regression analysis is a fundamental concept that consists of a set of machine learning methods that *predict a continuous outcome variable* (y) based on the value of one or multiple predictor variables (x).



Introduction

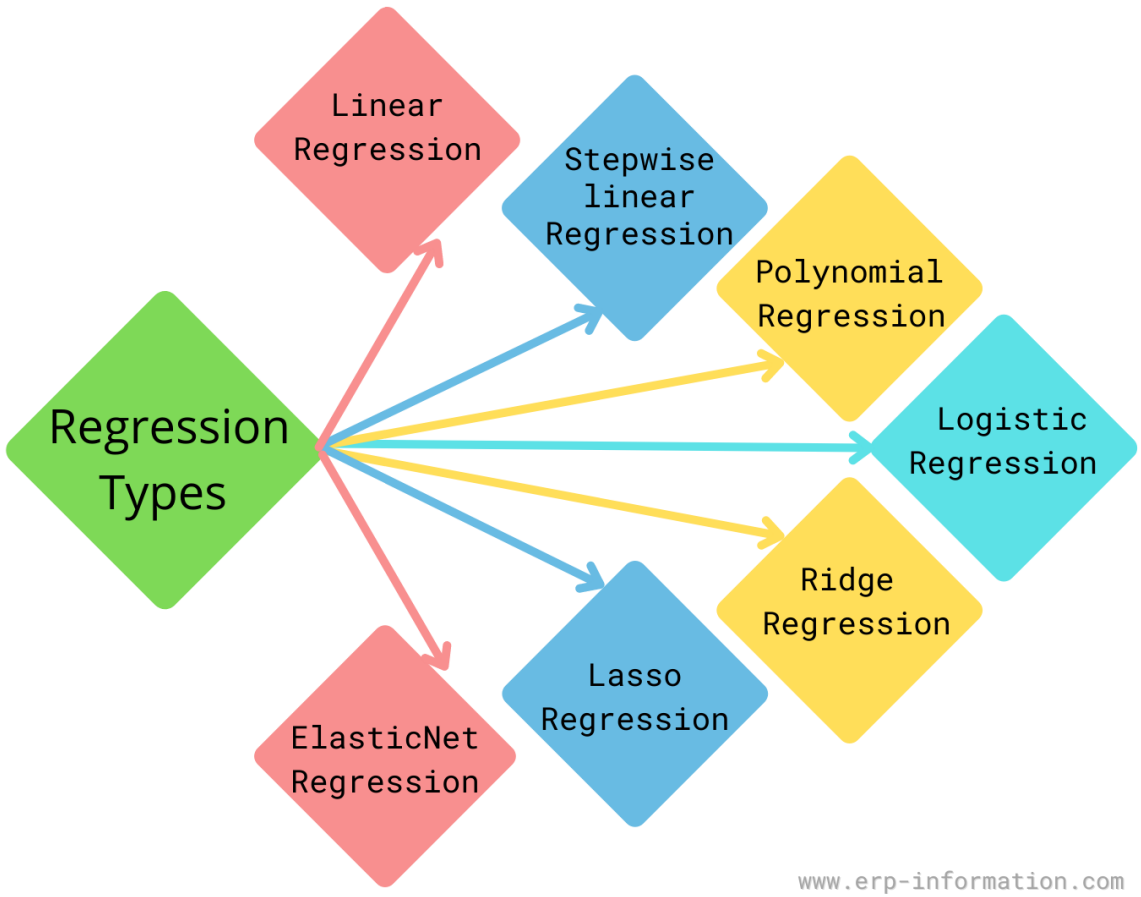
Examples of Regression

Regression models are widely used today. For example, regression analysis can predict a house's price given its features, predict the impact of SAT/GRE scores on college admissions, predict sales based on input parameters, and predict the weather.
Furthermore, there are different types of regression algorithms, such as linear regression, regression trees, lasso regression, and multivariate regression, that can help with the following:

- Sales and promotions forecasting
- Weather analysis and prediction
- Time series forecasting
- Indicating whether the stock prices of a company will increase in the future

Introduction

There are several types of regression in machine learning but just to name a few:



www.erp-information.com

Contents

History

- Sir Francis Galton (1822-1911) collected extensive data illustrating body height of parents and their grown children.
- He examined the *relationship* between body heights of the children and the average body height of both parents.
- To adjust for the natural height differences across gender, the body height of women was multiplied by a factor of 1.08.

Galton's Data

Table 1.1 Galton heredity data: contingency table between the height of 928 adult children and the average height of their 205 set of parents

Height of children	Average height of parents											
	64.0	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	73.0	Total
73.7	0	0	0	0	0	0	5	3	2	4	0	14
73.2	0	0	0	0	0	3	4	3	2	2	3	17
72.2	0	0	1	0	4	4	11	4	9	7	1	41
71.2	0	0	2	0	11	18	20	7	4	2	0	64
70.2	0	0	5	4	19	21	25	14	10	1	0	99
69.2	1	2	7	13	38	48	33	18	5	2	0	167
68.2	1	0	7	14	28	34	20	12	3	1	0	120
67.2	2	5	11	17	38	31	27	3	4	0	0	138
66.2	2	5	11	17	36	25	17	1	3	0	0	117
65.2	1	1	7	2	15	16	4	1	1	0	0	48
64.2	4	4	5	5	14	11	16	0	0	0	0	59
63.2	2	4	9	3	5	7	1	1	0	0	0	32
62.2	–	1	0	3	3	0	0	0	0	0	0	7
61.7	1	1	1	0	0	1	0	1	0	0	0	5
Total	14	23	66	78	211	219	183	68	43	19	4	928

The unit of measurement is inch which has already been used by Galton (1 inch corresponds to 2.54 cm)
Source: Galton (1889)

Galton's Discoveries

- Column-wise, i.e, for given average heights of the parents, the heights of the adolescents approximately follow a normal distribution.
- The normal distributions in each column have a common variance.
- When examining the relationship between the height of the children and the average height of the parents, an approximate linear trend was found with a slope of 2/3.
  - A slope with a value less than one led Galton to the conclusion that children of extremely tall (short) parents are usually shorter (taller) than their parents.
  - In either case there is a tendency towards the population average, and Galton referred to this as *regression* towards to the mean.

Scatter Plot of Galton's Data

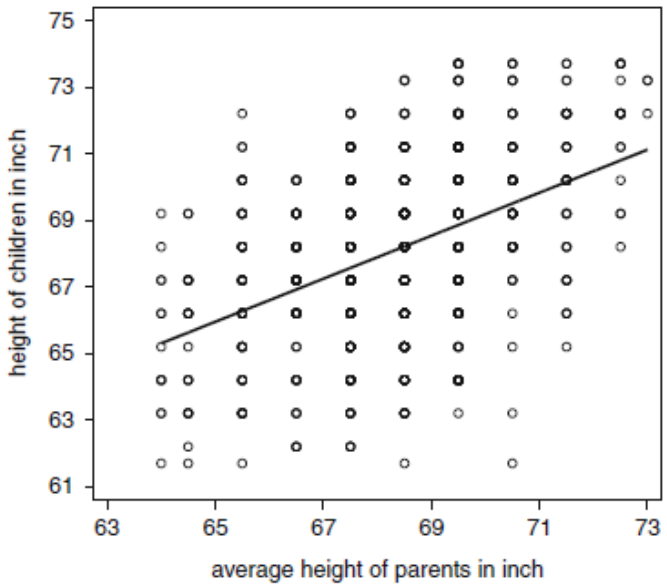


Fig. 1.1 Galton heredity data: scatter plot including a regression line between the height of children and the average height of their parents

- Galton visually added the trend or the *regression line*, which provides the average height of children as (average) parent height is varied to the scatter plot.

Galton's Simple Linear Regression

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- $x =$ "average size of parents" known as an explanatory variable, independent variable, regressor, or covariate.
- $y =$ "height of grown-up children" known as a response variable, outcome variable, or dependent variable.
- Galton determined the parameters β_0 and β_1 of the regression line in an ad hoc manner.
- Nowadays, these regression parameters are estimated via the *method of least square*.

Contents

Munich Rent Index Dataset

- A study of rent index in Munich, Germany in 1999.
- the goal is to model the impact of explanatory variables (living area, year of construction location, etc.) in the response variable of net rent or net rent per square meter.
- The data of 3,082 apartments were collected by representative random samplings on 9 variables.
  - `rent`: the monthly net rent per month (in Euro)
  - `rentsqm`: the net rent per month per square meter (in Euro)
  - `area`: living area in square meters
  - `yearc`: year of construction
  - `location`: quality of location: 1 - average location, 2 - good location, and 3 - top location
  - `bath`: quality of bathroom: 0 - standard, and 1 - premium
  - `kitchen`: quality of kitchen: 0 - standard, and 1 - premium
  - `cheating`: central heating: 0 - without central heating, 1 - with central heating
  - `district`: districts in Munich

Univariate Distributions of the Variables

- The first step when conducting a regression analysis is to get an overview of the variables
  - Summary and exploration of distribution of the variables
  - Identification of extreme values and outliers
  - Identification of incorrect variable coding
- Continuous variables
  - Descriptive statistics: mean, median, standard deviation, minimum and maximum
  - Graphical visualization: histogram and box plots, smooth nonparametric density estimators (kernel densities), and more.
- Categorical variables
  - Frequency table
  - Graphical visualization: bar graphs

Descriptive Statistics

Table 1.2 Munich rent index: description of variables including summary statistics

Variable	Description	Mean/ frequency in %	Std.- dev.	Min/max
rent	Net rent per month (in Euro)	459.43	195.66	40.51/1,843.38
rentsqm	Net rent per month per square meter (in Euro)	7.11	2.44	0.41/17.72
area	Living area in square meters	67.37	23.72	20/160
yearc	Year of construction	1,956.31	22.31	1918/1997
location	Quality of location according to an expert assessment			
	1 = average location	58.21		
	2 = good location	39.26		
	3 = top location	2.53		
bath	Quality of bathroom			
	0 = standard	93.80		
	1 = premium	6.20		
kitchen	Quality of kitchen			
	0 = standard	95.75		
	1 = premium	4.25		
cheating	Central heating			
	0 = without central heating	10.42		
	1 = with central heating	89.58		
district	District in Munich			

Histograms

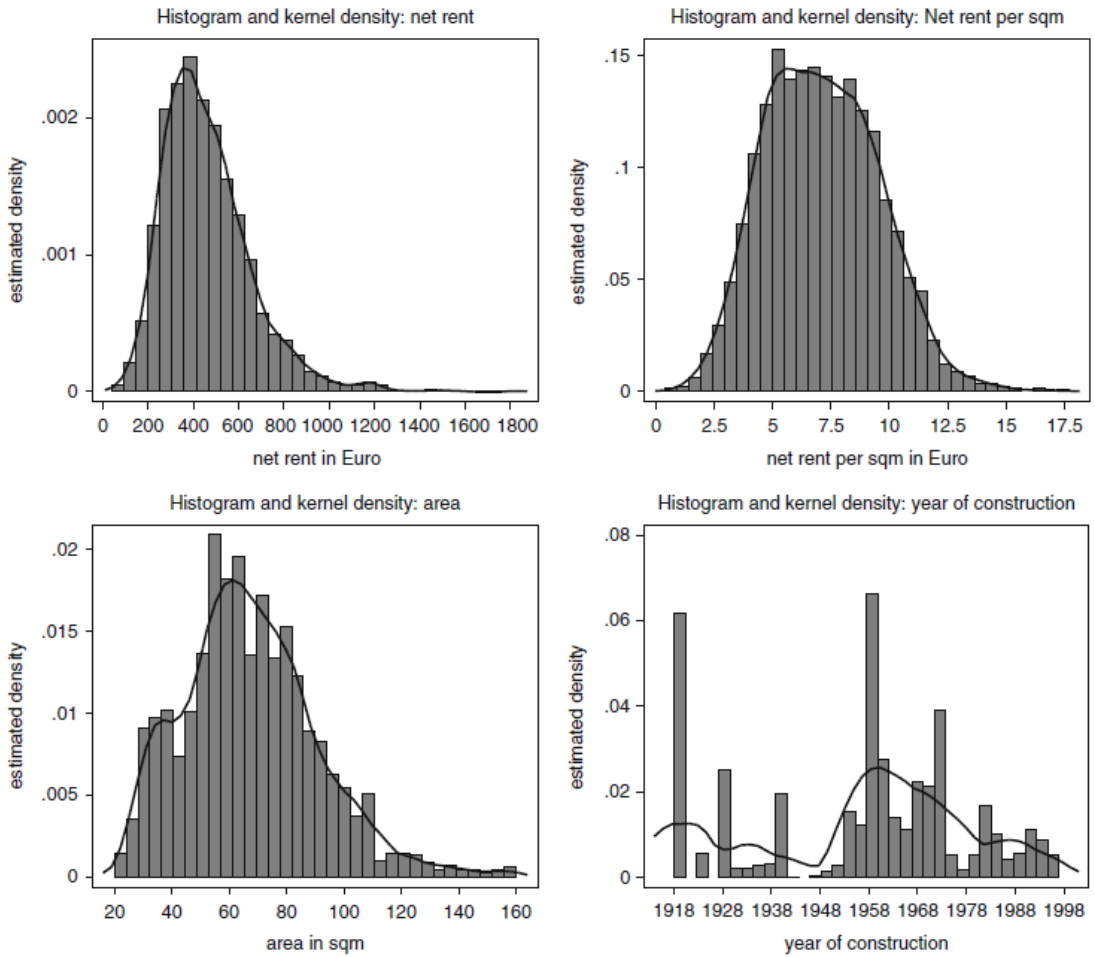
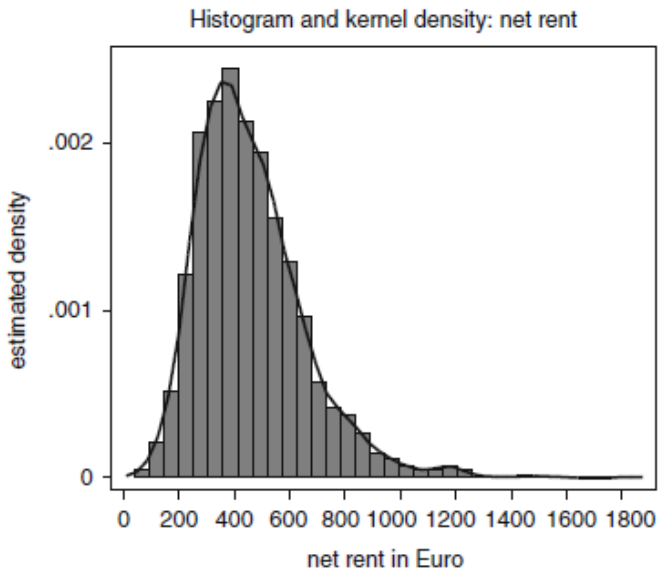


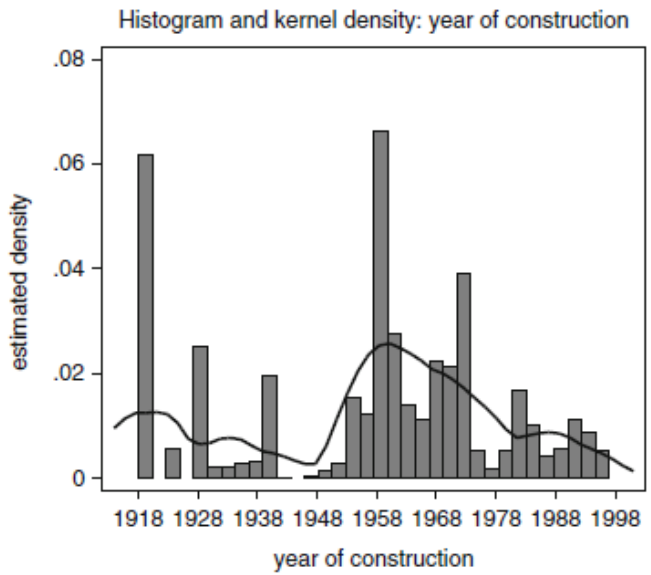
Fig. 1.3 Munich rent index: histograms and kernel density estimators for the continuous variables *rent*, *rentsqm*, *area* and *yearc*

Interpretations



The monthly net rent roughly varies between 40 and 1,843 Euro with an average of approximately 459 Euro. For the majority of apartments, the rent varies between 50 and 1200 Euro. For only a few apartments the monthly rent is higher than 1,200 Euro. This implies that any inference from a regression analysis regarding expensive apartments is comparable uncertain, when compared to the smaller and more model sized apartments. Generally, the distribution of the monthly net rent is asymmetric and skewed towards the right.

Interpretations



The distribution of the year of construction is highly irregular and multimodal, which is in part due to historical reasons. Whereas the data basis for apartments for the years of the economic crises during the Weimar Constitution and the Second World War is rather limited, there are much more observations for the later years of reconstruction (mode near 1960). Starting in the mid-1970s the construction boom stopped again. Altogether the data range from 1918 until 1997.

Graphical Association Analysis

- In the second step, we can graphically investigate the relationship between the response variable and explanatory variables
  - Type (e.g., linear versus nonlinear)
  - Strength
- Continuous variables: scatter plot
- Categorical variables: histograms, box plots, and (kenel) density estimators

Scatter Plots - Continuous Variables

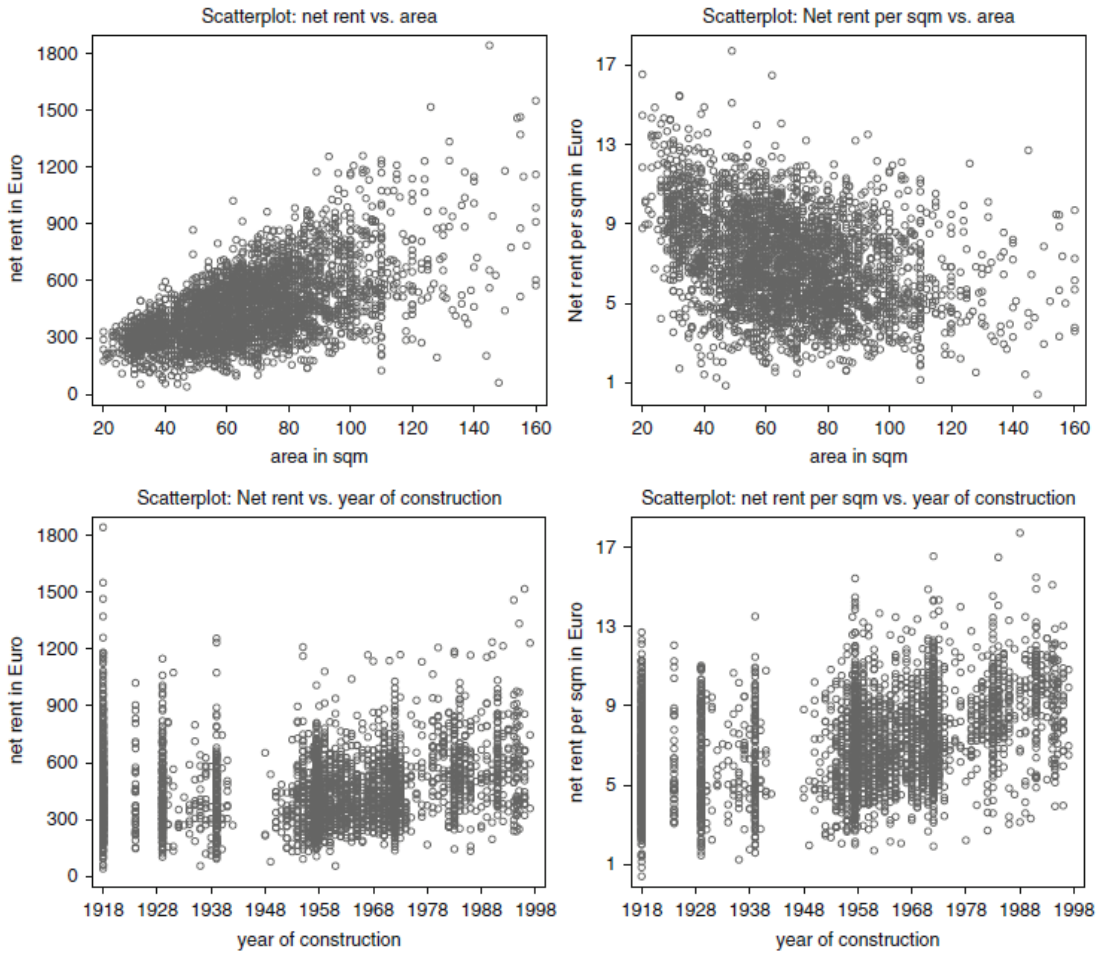


Fig. 1.5 Munich rent index: scatter plots between net rent (*left*) / net rent per sqm (*right*) and the covariates area and year of construction

- Scatter plots are not very informative due to the large sample size (more than 3000)
- There is an approximately linear relationship between net rent and living area
- Variability of the net rent increases with an increased living area
- The relationship between net rent per square meter and living area is harder to determine
- The net rent per square meter for larger apartments seems to decrease but hard to define the relationship (linear or nonlinear)
- The relationship of either of the two response variables and the year of construction is hardly visible (it exists at all), but there is a monotonic increase of rents (and rents per square meter) for flats built after 1948

Cluster Scatter Plots - Continuous Variables

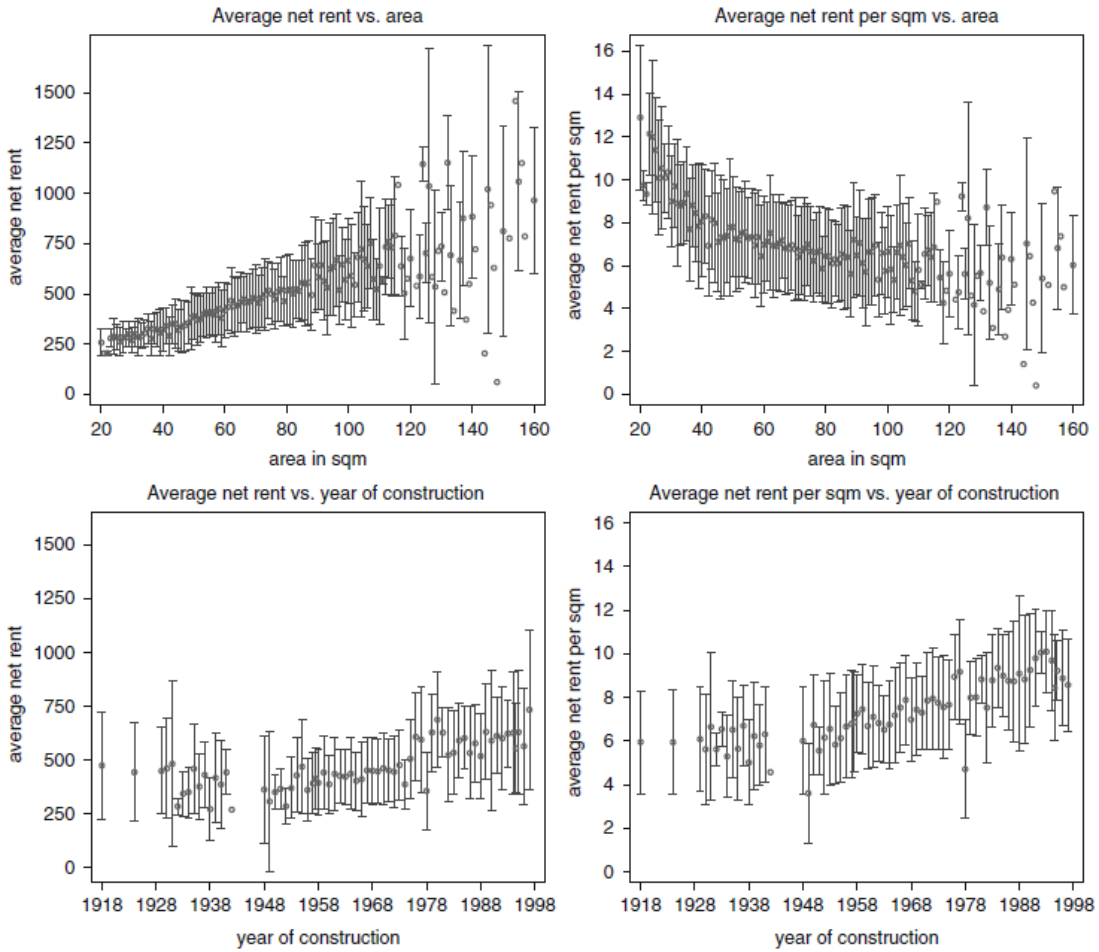


Fig. 1.6 Munich rent index: average net rent (*left*) and net rent per sqm (*right*) plus/minus one standard deviation versus area and year of construction

- It is clearer to see the relationship between the variables
- The net rent per square meter has a nonlinear relationship monotonically decreasing with living area
- For larger apartments (120 square meters or larger) has wider variability of average rents
- There exists a weak nonlinear relationship between the year of construction and the net rent per square meter

Box plots - Categorical Variables

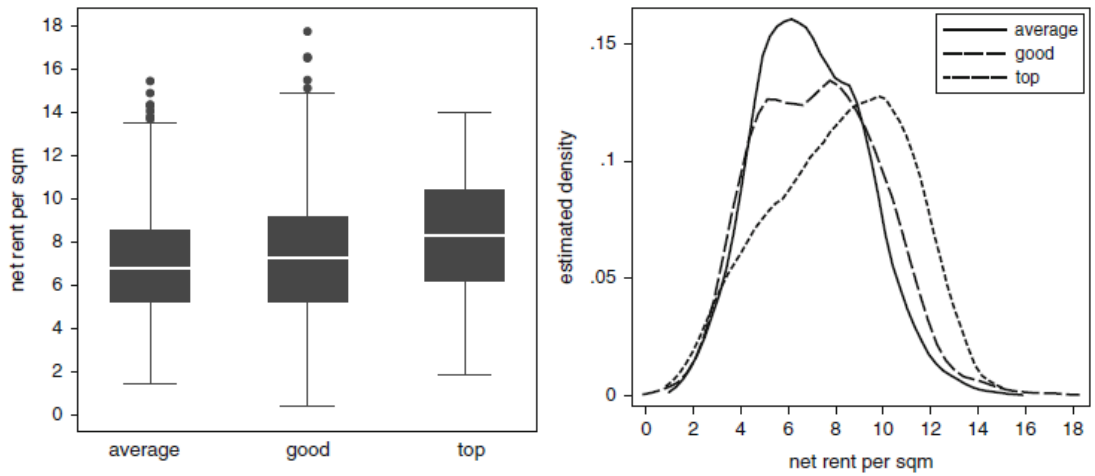


Fig. 1.7 Munich rent index: distribution of net rent per sqm clustered according to location

- The median rent (as well as the variation) increases as the location of the apartment improves
- The smooth density estimators offer similar information but not as obvious as box plots