Housing Prices

Mine Çetinkaya-Rundel

Table of contents

## 1 Introduction

In this analysis, we build a model predicting sale prices of houses based on data on houses that were sold in the Duke Forest neighborhood of Durham, NC around November 2020. Let’s start by loading the packages we’ll use for the analysis.

library(openintro) # for data  
library(tidyverse) # for data wrangling and visualization  
library(knitr) # for tables  
library(broom) # for model summary

We present the results of exploratory data analysis in [Section 2](#sec-eda) and the regression model in [Section 3](#sec-model).

We’re going to do this analysis using literate programming (Knuth 1984).

## 2 Exploratory data analysis

The data contains 98 houses. As part of the exploratory analysis let’s visualize and summarize the relationship between areas and prices of these houses.

### 2.1 Data visualization

[Figure 1](#fig-histogram) shows two histograms displaying the distributions of price and area individually.

ggplot(duke\_forest, aes(x = price)) +  
 geom\_histogram(binwidth = 50000) +  
 labs(title = "Histogram of prices")  
  
ggplot(duke\_forest, aes(x = area)) +  
 geom\_histogram(binwidth = 250) +  
 labs(title = "Histogram of areas")

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | (a) Histogram of prices | | |  | | --- | | (b) Histogram of areas | |

Figure 1: Histograms of individual variables

[Figure 2](#fig-scatterplot) displays the relationship between these two variables in a scatterplot.

ggplot(duke\_forest, aes(x = area, y = price)) +  
 geom\_point() +  
 labs(title = "Price and area of houses in Duke Forest")

|  |
| --- |
| Figure 2: Scatterplot of price vs. area of houses in Duke Forest |

### 2.2 Summary statistics

[Table 1](#tbl-stats) displays basic summary statistics for these two variables.

duke\_forest %>%  
 summarise(  
 `Median price` = median(price),  
 `IQR price` = IQR(price),  
 `Median area` = median(area),  
 `IQR area` = IQR(area),  
 `Correlation, r` = cor(price, area)  
 ) %>%  
 kable(digits = c(0, 0, 0, 0, 2))

Table 1: Summary statistics for price and area of houses in Duke Forest

| Median price | IQR price | Median area | IQR area | Correlation, r |
| --- | --- | --- | --- | --- |
| 540000 | 193125 | 2623 | 1121 | 0.67 |

## 3 Modeling

We can fit a simple linear regression model of the form shown in [Equation 1](#eq-slr).

[Table 2](#tbl-lm) shows the regression output for this model.

price\_fit <- lm(price ~ area, data = duke\_forest)  
   
price\_fit %>%  
 tidy() %>%  
 kable(digits = c(0, 0, 2, 2, 2))

Table 2: Linear regression model for predicting price from area

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 116652 | 53302.46 | 2.19 | 0.03 |
| area | 159 | 18.17 | 8.78 | 0.00 |

|  |
| --- |
| Note |
| This is a pretty incomplete analysis, but hopefully the document provides a good overview of some of the authoring features of Quarto! |

## References

Knuth, D. E. 1984. “Literate Programming.” *The Computer Journal* 27 (2): 97–111. <https://doi.org/10.1093/comjnl/27.2.97>.