Multiple Linear Regression: Home Sales in King County, WA

KEY

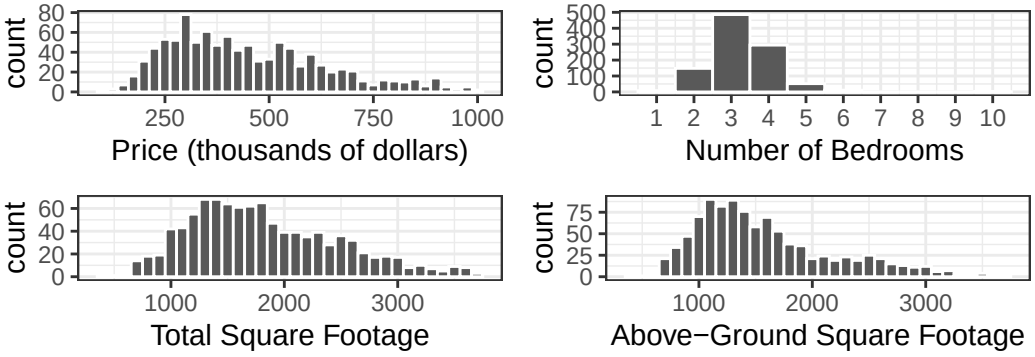
141-01-01

Learning Objective: Practice the process of checking assumptions, assessing, and interpreting multiple linear regression model to answer a research question.

Introduction

Today we will analyze data on 1000 homes sold in King County, WA between May 2014 and May 2015. We would like to answer the following research question: **What's the relationship between a home's square footage and number of bedrooms on the home's sale price?**

We'll consider **four variables**: Sale price (response); and square footage, above-ground square footage, and number of bedrooms (potential explanatory variables). Check out the exploratory plots below:



Question 1: Exploration

(a) Describe the distribution of 'price'. Consider shape (symmetric, left/right skewed), unimodality vs. multimodality, center, and spread.

Price has a right-skewed distribution, is unimodal, has an average around 300-400k, and has a range of approximately 200k to 1 million dollars.

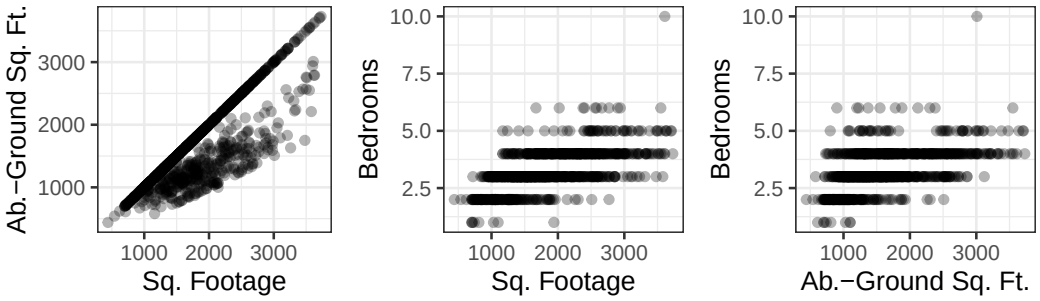
(b) Hypothesize the (i) direction and (ii) strength of relationship between 'price' and the number of 'bedrooms'.

I hypothesize that 'price' and 'bedrooms' have a moderately strong, positive relationship, because as the number of bedrooms increases, I would expect that price would be very likely to increase as well

Multicollinearity refers to situations when explanatory variables are highly correlated with one another. Multicollinearity violates the independence assumption of MLR, and often results in coefficients that are distorted in erroneous ways! We want to avoid multicollinearity (*some* correlation is okay!)

Question 2: Multicollinearity

To check for multicollinearity, examine the following scatterplots displaying each pair of **explanatory** variables and then answer the following question.



Which pair of explanatory variables suffers *most* from multicollinearity? Hypothesize an explanation in context for the observed pattern.

Square footage and above-ground square footage are the most multicollinear. We notice that points on the diagonal in their pairwise scatterplot represent homes with no basement; these variables are identical for such homes. Points off the diagonal are homes with a basement. In short, these variables essentially represent the same thing, and thus, we should include at most one of them in our regression model.

Model Estimation

Based on our research question and exploration so far, we will regress Sale Price (`price`) against Square Footage (`sqft_living`) and Number of Bedrooms (`bedrooms`). Notice that one variable was excluded! A model summary is below:

```
# A tibble: 3 x 7
  term        estimate std_error statistic p_value lower_ci upper_ci
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept    197.        19.4      10.1    0      159.     235.
2 sqft_living    0.177      0.01     18.4    0        0.158    0.196
3 bedrooms     -21.6        7.27     -2.97   0.003  -35.9     -7.34
```

Question 3: Regression Model

(a) Write down an equation, including coefficient values, for the estimated model:

$$\hat{\text{Price}} = 197 + 0.177\text{SquareFootage} - 21.6\text{Bedrooms}$$

(b) Carefully interpret the **intercept** coefficient in our model.

For a home with 0 bedrooms and 0 square footage, we predict a price of $197k, on average.

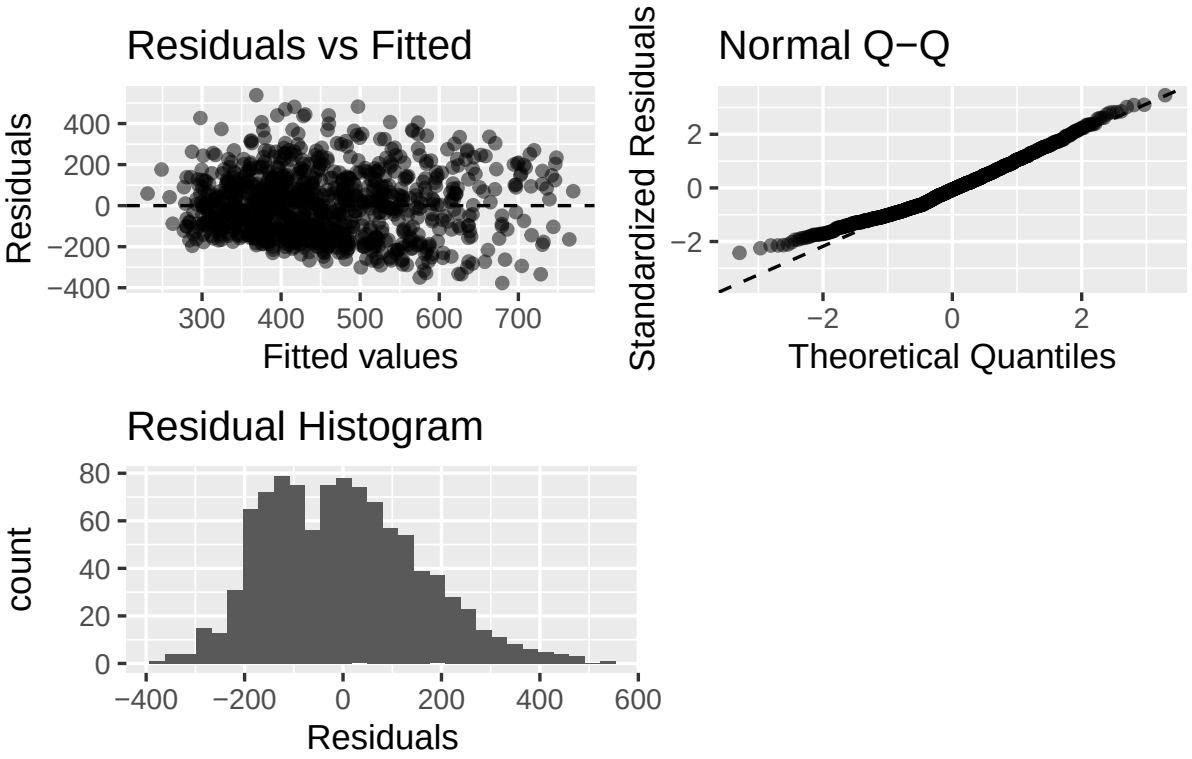
(c) Carefully interpret the coefficient for **Bedrooms** in our model. Did this match your expectation?

For every 1-bedroom increase, we expect the price of a home to decrease by $21.6k, on average, holding square footage constant.

This did not match my expectations! I expected prices to generally increase as we have more bedrooms. We may think of this unintuitive result as being a result of some multicollinearity between our remaining explanatory variables.

Question 4: Assumptions

It is important to assess our assumptions for linear regression whenever conducting an analysis. When our assumptions are not met, we should feel more skeptical about our coefficient estimates and their interpretations (or ignore them entirely). Examine the following plots below:



(a) Which plot above can be used to assess **Linearity**? Do you believe the assumption is satisfied?

The plot on the top left can be used to assess linearity. I believe the assumption is generally satisfied, because the points fall roughly equally above and below the line.

(b) We will consider the **Independence** assumption without a plot! Based on our data and chosen model, do you believe the assumption is satisfied?

I believe the assumption is roughly satisfied. I would assume that this random sample of 1000 homes in King County, WA are reasonably-separated geographically, and I don't think that there's any other evidence to suggest non-independence between these observations. Still, I have some concerns about multicollinearity which casts some doubt on independence.

(c) Which plot above can be used to assess **Normality of Residuals**? Do you believe the assumption is satisfied?

The plots on the top right and bottom left can be used to assess normality of residuals. I believe the assumption is questionable, specifically because the

residual histogram appears to be somewhat bimodal and we have a bit of a heavy tail from the qq-plot.

(d) Which plot above can be used to assess **Equal Variance**? Do you believe the assumption is satisfied?

The plot on the top left can be used to assess equal variance. I believe the assumption is arguably satisfied, although I have some concern about potential fanning (there appears to be less variability when predicted price is smaller) This could go either way.

Question 5: Conclusions

Based on our data exploration, model estimation, and assessment of assumptions, answer the research question. Be sure to state any doubts or concerns you have about the may have about our analysis.

We have learned that, in general, higher home prices are associated with more square footage and/or fewer bedrooms. However, I have some concerns about this model result, specifically due to potential independence/multicollinearity issues (yielding an unintuitive result about bedrooms) as well as concerns regarding normality of results and equal variance.