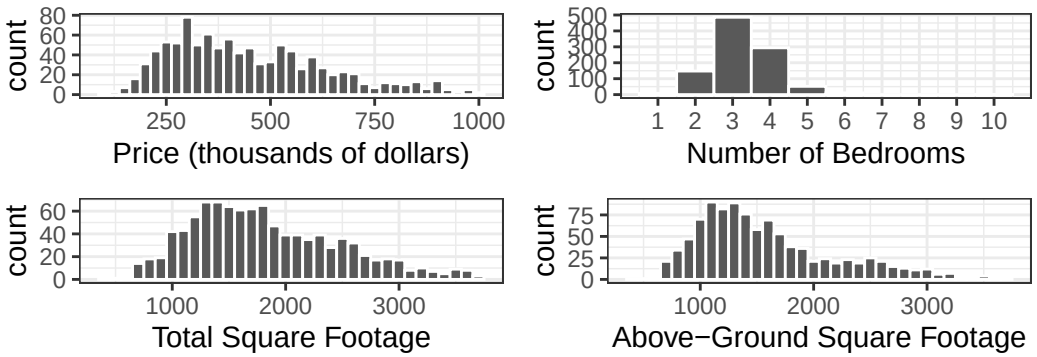
Multiple Linear Regression: Home Sales in King County, WA

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

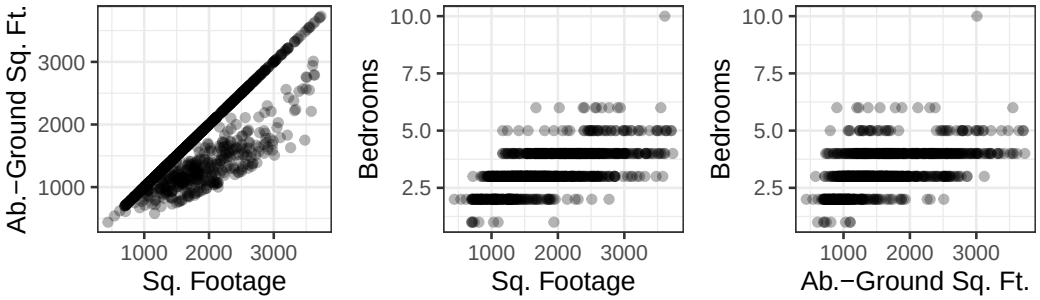
(a) Describe the distribution of 'price'. Consider shape (symmetric, left/right skewed), unimodality vs. multimodality, center, and spread.

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

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.

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

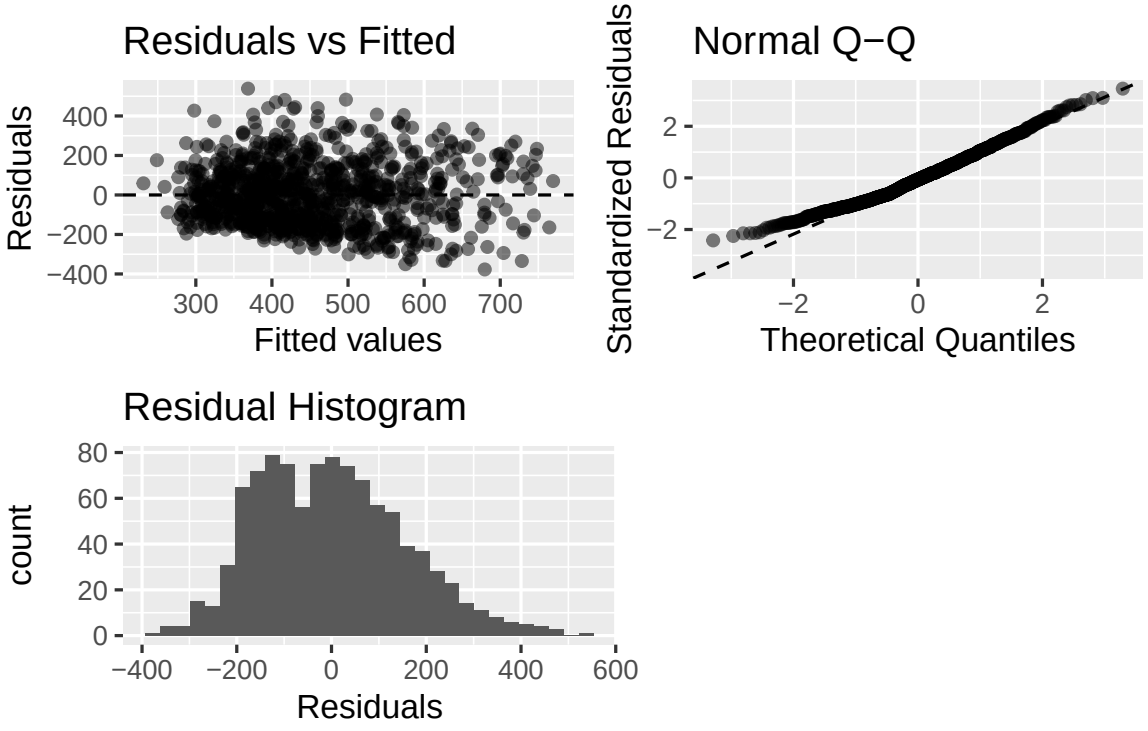
(a) Write down an equation, including coefficient values, for the estimated model:

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

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

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?

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

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

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

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.