DSCI 100 - Introduction to Data Science

Lecture 9 - Introduction to linear regression

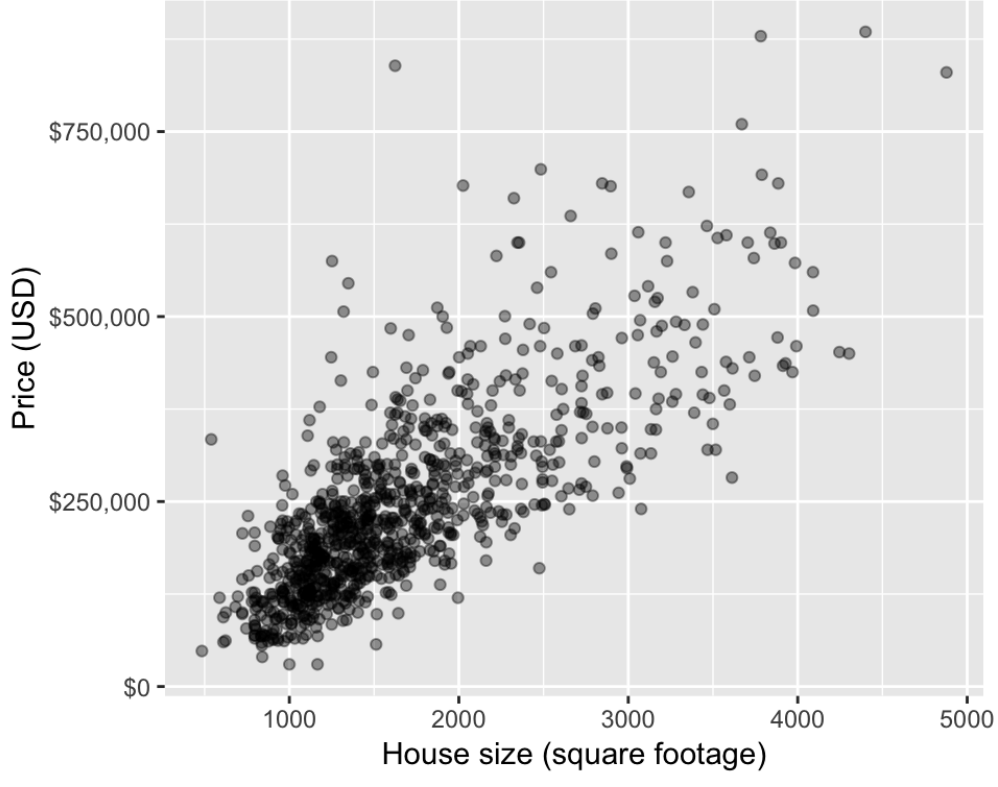
2019-03-13

News and reminders

- Tuesday, March 19th - in class peer review session
- Friday, April 26th at 19:00 - Final exam (format TBD)

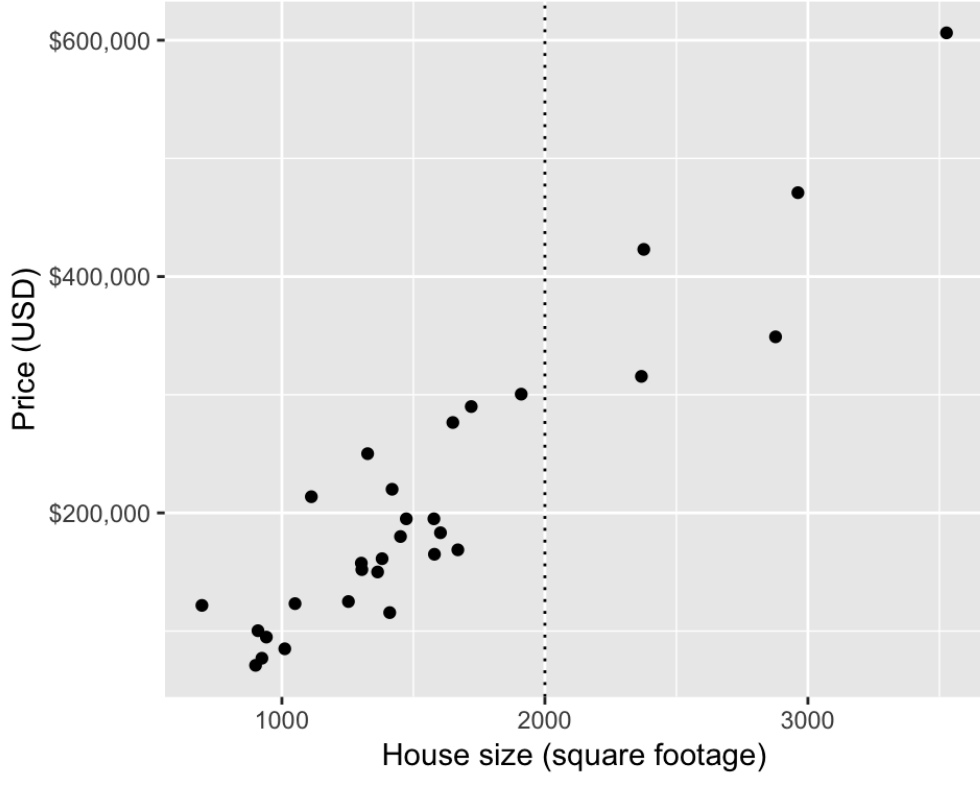
Regression prediction problem

What if we want to predict a quantitative value instead of a class label?



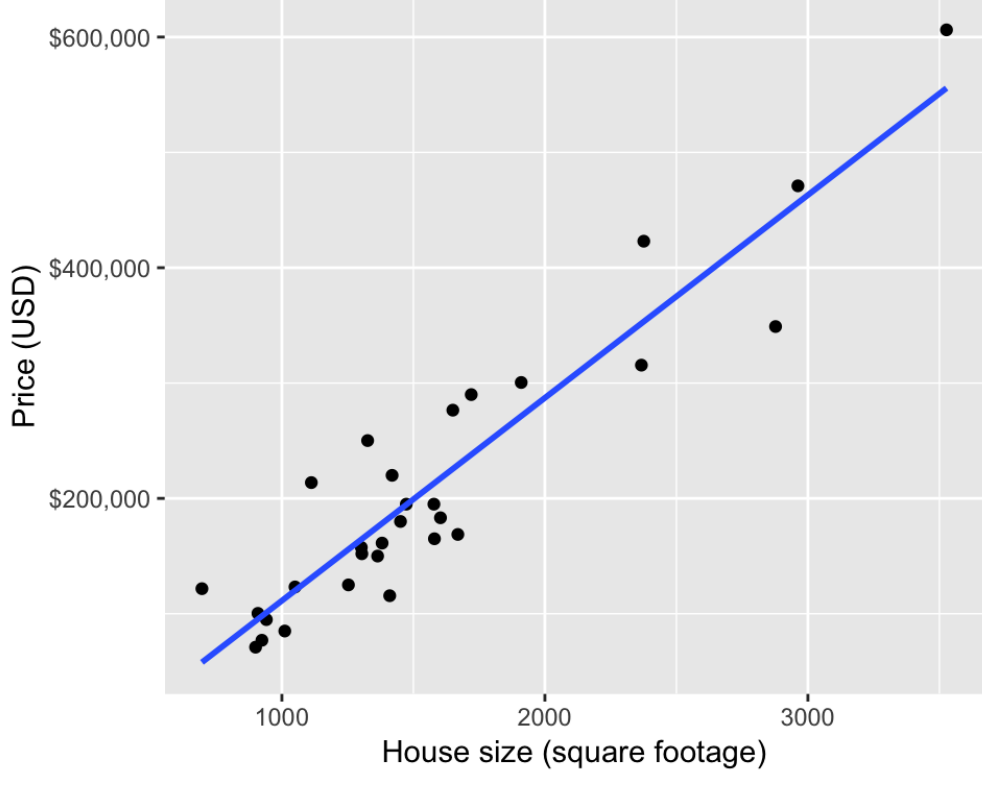
Today we will focus on another regression approach - linear regression.

For example, the price of a 2000 square foot home (from this reduced data set):



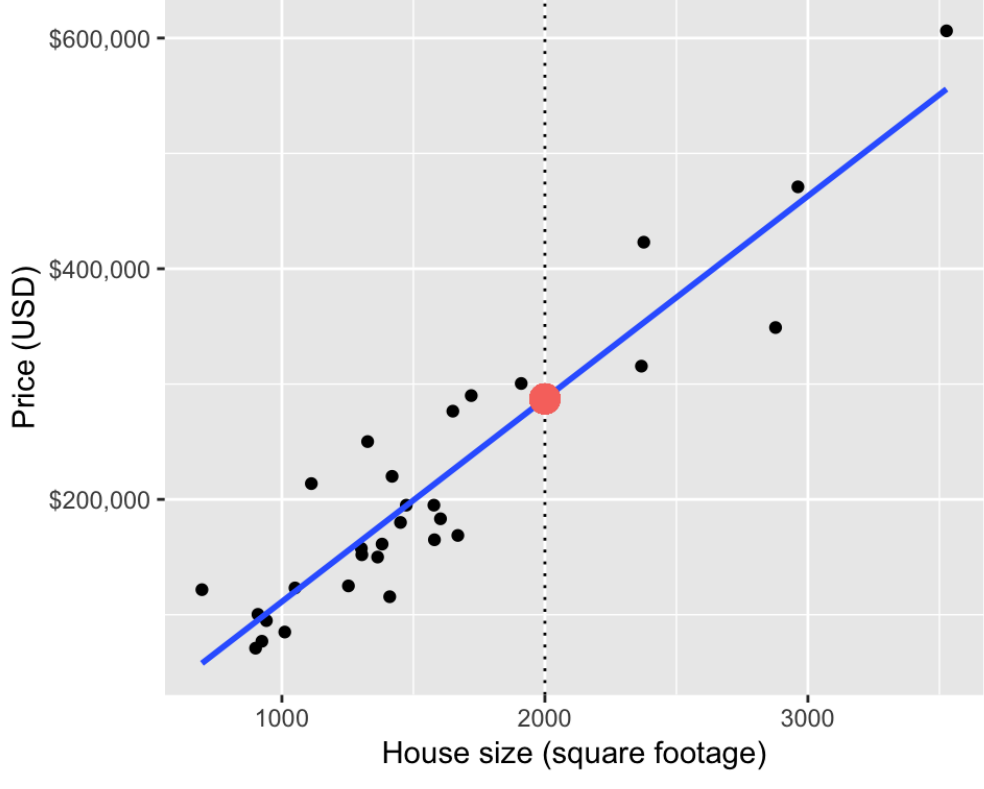
linear regression

First we find the line of "best-fit" through the data points:



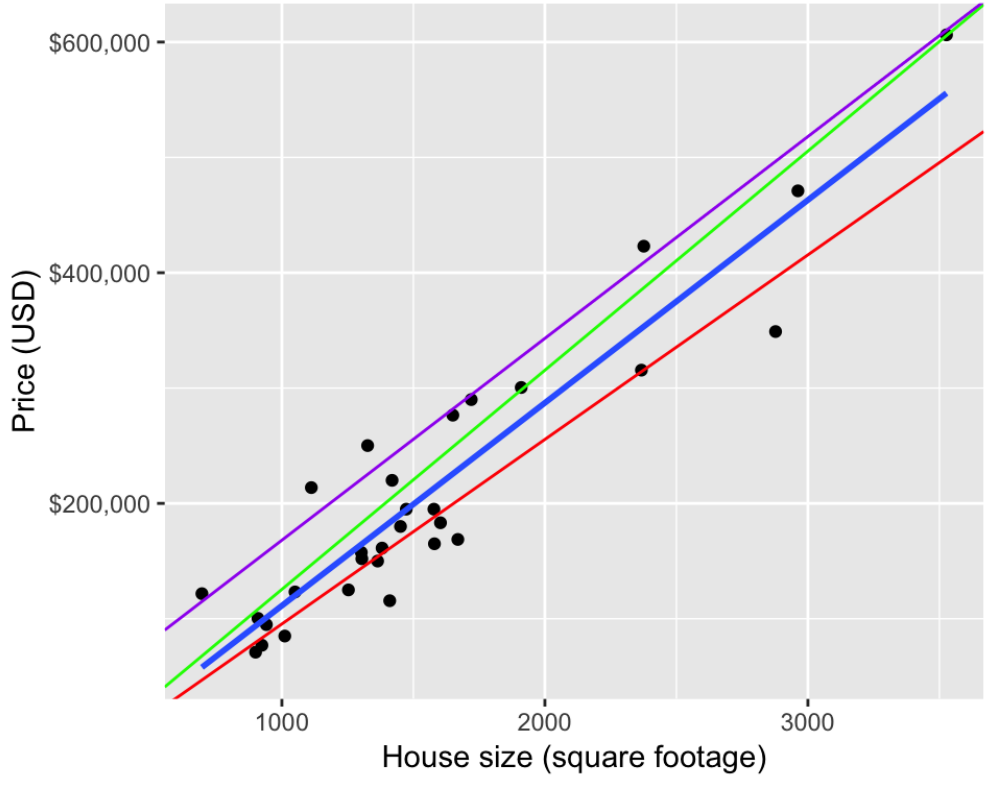
linear regression

And then we "look up" the value we want to predict of off of the line.



linear regression

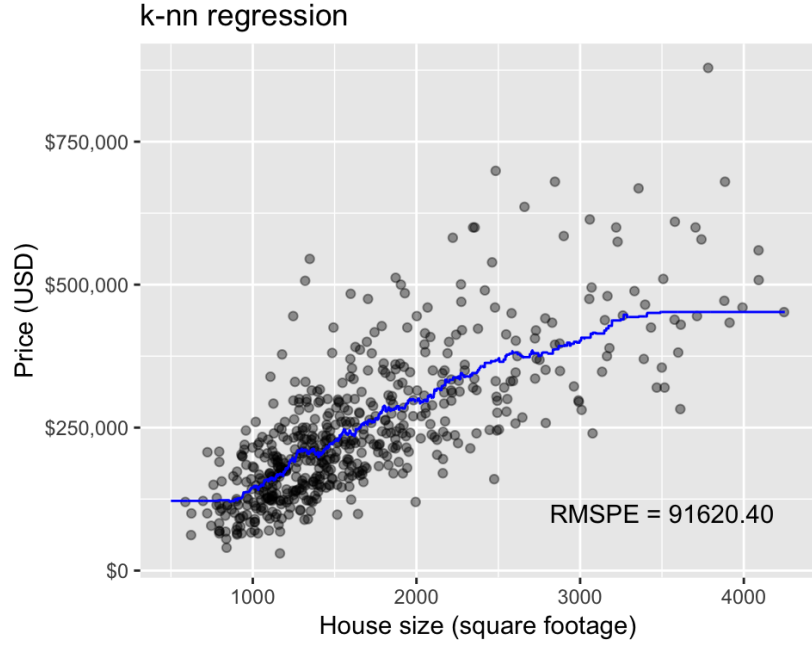
How do we choose the line of "best fit"? We can draw many lines through the data:



linear regression

We choose the line that minimzes the **average** vertical distance between itself and each of the observed data points

Linear vs k-nn regression



Why linear regression?

Advantages to restricting the model to straight line: interpretability!

Remembering that the equation for a straight line is: $Y = \beta_0 + \beta_1 X$

Where:

- β_0 is the y-intercept of the line (the value where the line cuts the y-axis)
- β_1 is the slope of the line

We can then write:

$$house\ price = \beta_0 + \beta_1 house\ size$$

And finally, fill in the values for β_0 and β_1:

$$house\ price = -64542.2 + 175.9 * house\ size$$

k-nn regression, as simple as it is to implement and understand, has no such interpretability from it's wiggly line.

Why not linear regression (sometimes?)

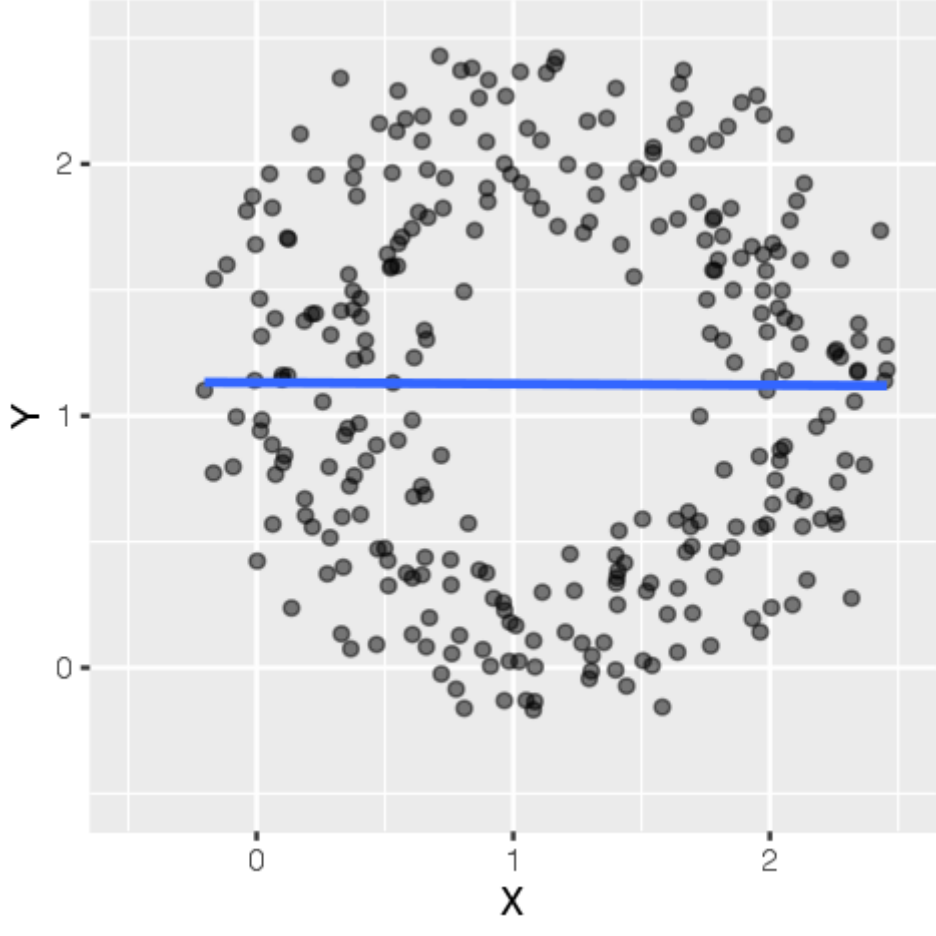
Models are not like kitten hugs

They are more like suits:

ONE SIZE DOES NOT FIT ALL!

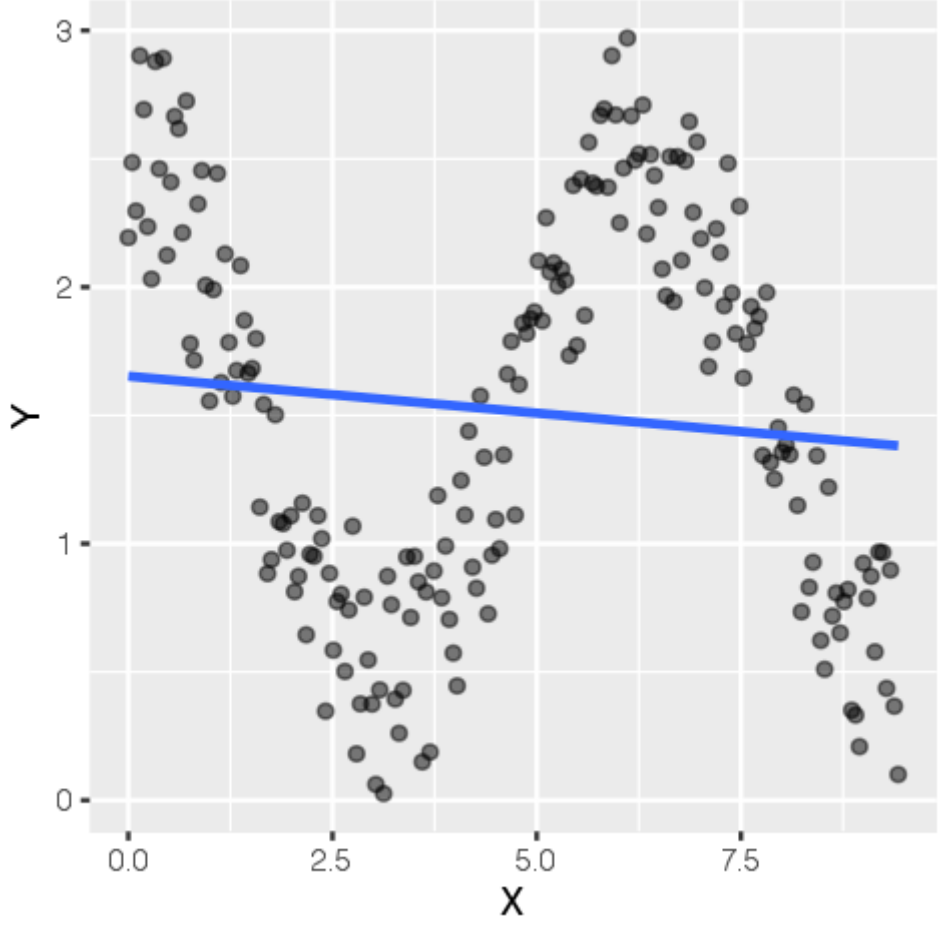
Be cautious with linear regression with data like this:

In [2]:
```
circle_plot
```



and this:

In [3]:
```
zigzag_plot
```



A cool app to explore more about linear regression

http://setosa.io/ev/ordinary-least-squares-regression/ (http://setosa.io/ev/ordinary-least-squares-regression/)

What did we learn

- linear regression
- has to be a straight line
- RMSE vs RMSPE
- `geom_smooth`
- don't need to use k or cross-validation to fit a linear regression