When to Use Linear Regression

In this lesson:

Benefits of Linear Regression

How do I decide what kind of car to buy? Or what kind of house? Using machine learning tools such as linear regression can help make these decisions easier by using a data-driven method of evaluation.

However, not all datasets are a good fit for linear regression. In this lesson, you will examine two datasets: one about cars, and one about housing. This lesson will help you identify what kinds of datasets can be used for linear regression to ensure you have a good predictive model.

What is Linear Regression?

Linear regression is the most common type of machine learning algorithm. The algorithm will predict new values by determining the relationship between the data fed into the algorithm.

More information about linear regression can be found in the intermediate lesson.

Understanding the Data

The first dataset is on car performance and was extracted from the 1974 Motor Trend US magazine. It comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

> Can we predict the mpg based on other characteristics?

Here is the data recorded:

- mpg: Miles/(US) gallon
- disp: Displacement (cu.in.)
- hp: Gross horsepower
- drat: Rear axle ratio
- wt: Weight (1000 lbs)
- qsec: 1/4 mile time
- vs: Engine (0 = V-shaped, 1 = straight)
- am: Transmission (0 = automatic, 1 = manual)
- gear: Number of forward gears
- carb: Number of carburetors

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

The second dataset is on house prices and was extracted as a practice dataset from the following source: https://www.kaggle.com/egebozoglu/house-price-linear-regression (https://www.kaggle.com/egebozoglu/house-price-linear-regression). This dataset includes the cost of houses and 18 aspects of home design and location in Seattle, Washington, from 2014 to 2015.

> Can we predict the cost of a house based on other characteristics?

Here is the data recorded:

- id: House identifying number
- price: House price (USD)

- bedrooms: Number of bedrooms
- bathrooms: Number of bathrooms
- sqft_living: Area (ft^2) of the living room
- sqft_lot: Area (ft^2) of the property
- floors: Ranking of floor quality (1-3.5)
- waterfront: Is the house on the waterfront (1=Yes, 0=No)
- view: Does the house have a nice view, ranking 0-4
- condition: Ranking of house condition (1-5)
- grade: Ranking of house quality (1-13)
- sqft_above: Area (ft^2) upstairs
- sqft_basement: Area (ft^2) of the basement
- yr_built: Year the house was built
- yr_renovated: Year the house was renovated
- zipcode: Residential zipcode
- lat: Latitudinal position of the house
- long: Longitudinal position of the house
- sqft_living15: Area (ft^2) of the property in 2015
- sqft_lot15: Area (ft^2) of the property in 2015

id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_ren
7129300520	221900	3	1.00	1180	5650	1	0	0	3	7	1180	0	1955	
6414100192	538000	3	2.25	2570	7242	2	0	0	3	7	2170	400	1951	
5631500400	180000	2	1.00	770	10000	1	0	0	3	6	770	0	1933	
2487200875	604000	4	3.00	1960	5000	1	0	0	5	7	1050	910	1965	
1954400510	510000	3	2.00	1680	8080	1	0	0	3	8	1680	0	1987	
7237550310	1225000	4	4.50	5420	101930	1	0	0	3	11	3890	1530	2001	

Missing Data

The first step in cleaning a dataset is ensuring that there is not any data missing from the dataset. Missing data is when there are any rows in the table that are not filled with a value. Our datasets do not have any missing data, so nothing needs to be changed.

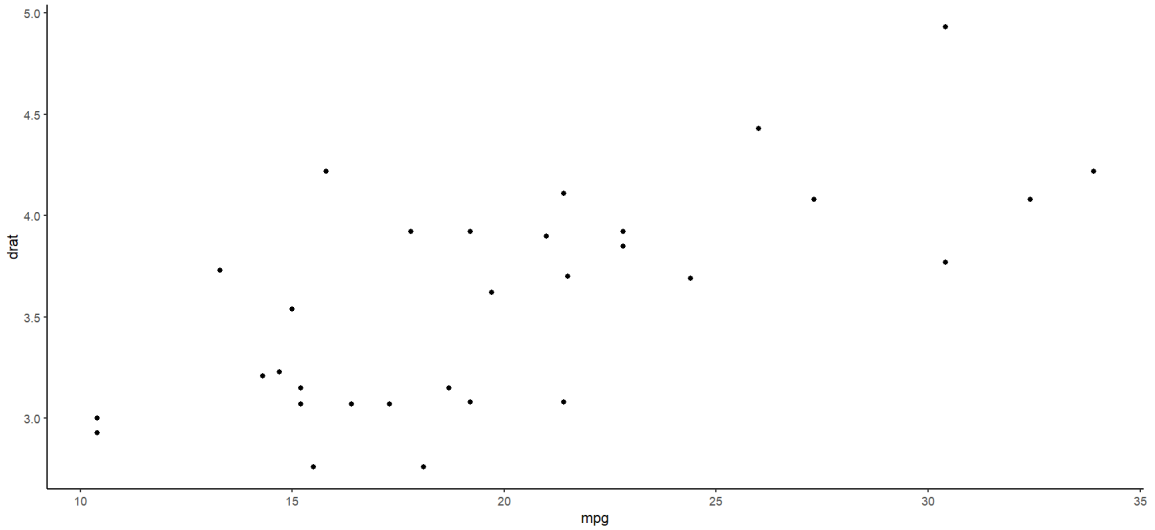
More information on how to address missing data can be found in the intermediate lesson.

Establishing Linear Relationships

The next step in deciding whether or not a dataset is good for linear regression is to examine whether there are linear relationships between the different parameters.

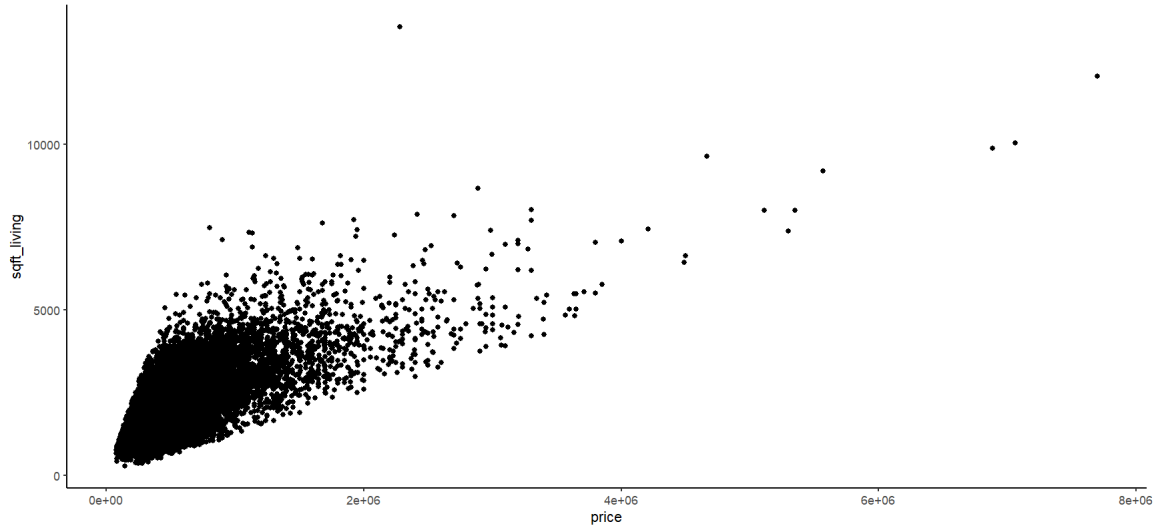
You can use the tool below to examine the relationship between any two variables in the datasets.

First, let's look at the cars dataset. Start by selecting mpg as variable 1. Then, change variable 2 and look for correlations.



Q1: What correlations do you see with mpg?

Next, let's look at the housing dataset. Start by selecting price as variable 1. Then, change variable 2 and look for correlations.



Q2: What correlations do you see with price?

Q3: Why is some data scattered, but others form lines?

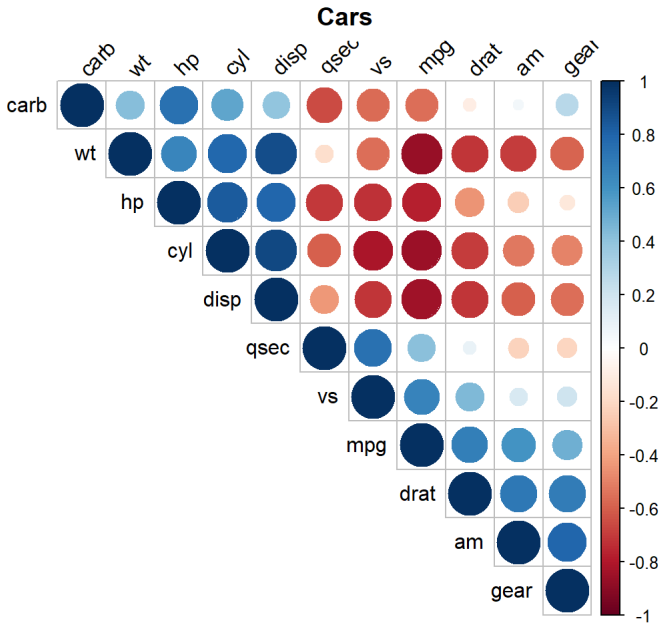
Now play around with this tool, changing variable 1 and variable 2 for both the cars and housing datasets.

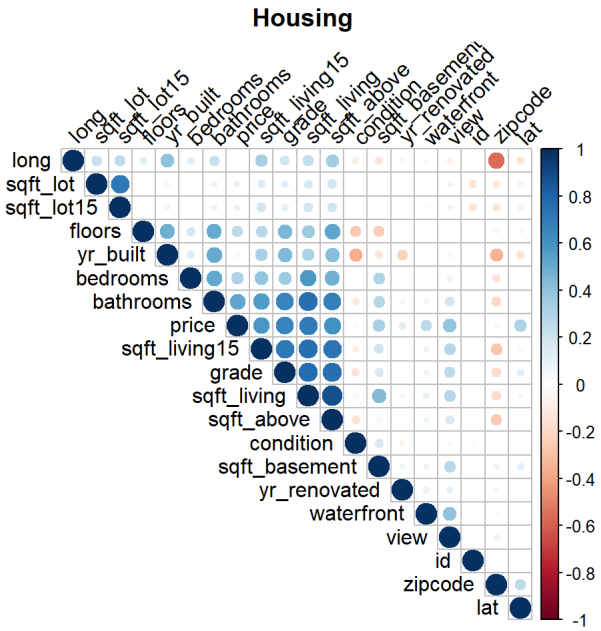
Q4: What other correlations can you find?

Next, you can examine the strength of the correlations between each of the variables.

The plots below shows all of the variables plotted against one another. The size and darkness of a circle indicate the strength of correlation between the two variables. Blue symbolizes a positive correlation, while red symbolizes a negative correlation.

To examine a specific correlation, choose the box you want to examine and follow the row to the left of the box and the column above the box. This will direct you to the two labels, and therefore the two variables that have a certain correlation.





First, look at the correlation plot for the cars dataset. Let's look specifically at mpg. Start at the top label 'mpg', and move down the column.

> Q5: Do you see any strong correlations with mpg in this column?

Now, start at the 'mpg' label on the left, and work your way across to the right.

> Q6: Do you see any strong correlations with mpg in this row?

There are a lot of strong correlations with mpg, so let's examine the relationship between mph and rear axle ratio. But first, let's examine the correlation plot for the housing dataset. Look specifically at price. Start at the top label 'price', and move down the column.

> Q7: Do you see any strong correlations with price in this column?

Now, start at the 'price' label on the left, and work your way across to the right.

> Q8: Do you see any strong correlations with price in this row?

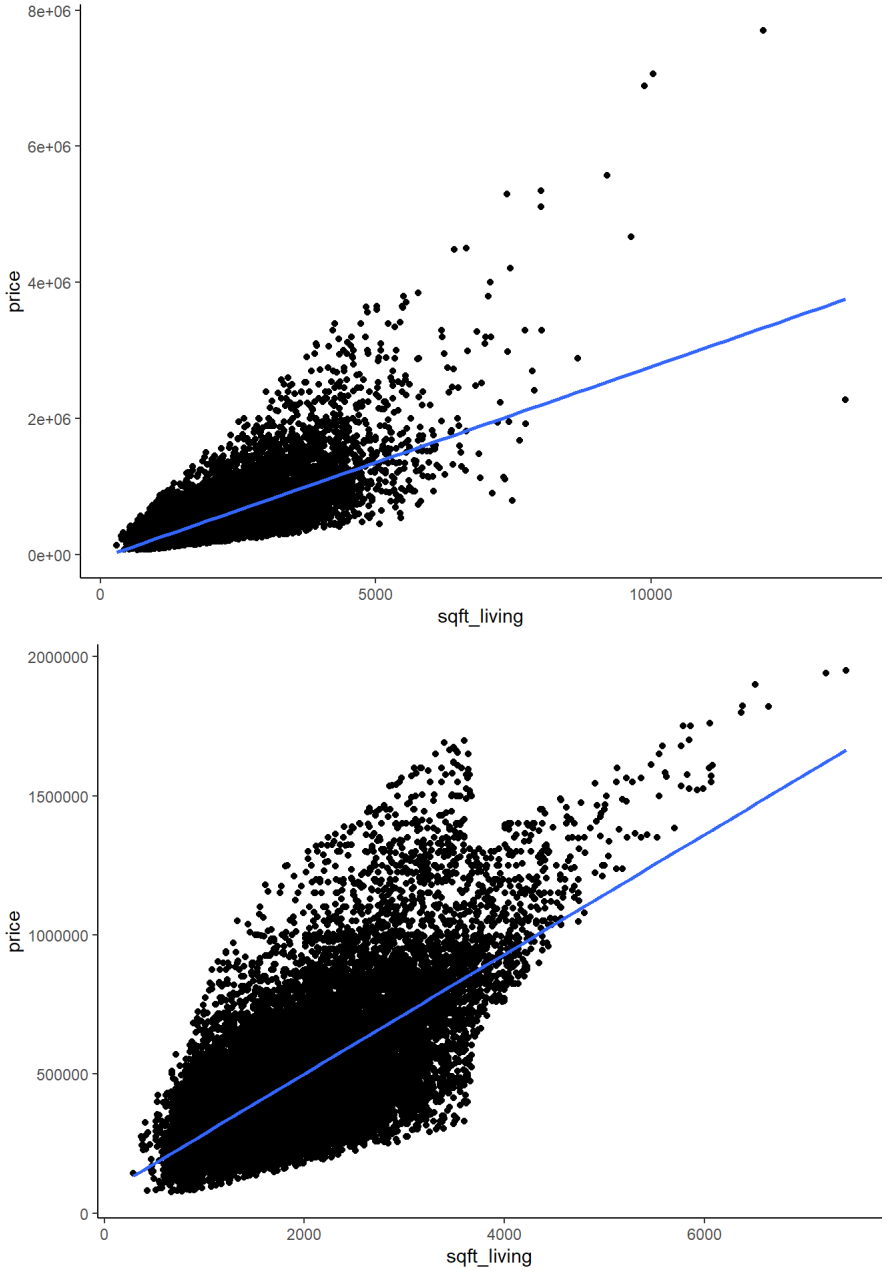
It appears the strongest correlation with price is sqft_living, so let's examine this relationship a little bit more.

Outliers and High-Leverage Points

The next step in cleaning our data is removing any outliers or high-leverage points.

> An outlier is a data point that has an extreme y-value, and a high leverage point is a data point that has an extreme x-value.

Examples of the dataset before and after removing significant outliers and high leverage points are shown below.



More information about outliers and high leverage points can be found in the intermediate lesson.

Four Assumptions of Linear Regression

There are obviously correlations within this dataset, but that doesn't necessarily mean that this data can be used for linear regression.

In order to use linear regression on a dataset, four assumptions about the data must be true: 1. Linearity of the data 2. Normality of the residuals 3. Homogeneity of residuals variance 4. Independence of residuals error terms

You can read more about these four assumptions at the following websites, or in the intermediate lesson:

http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/ (http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/)

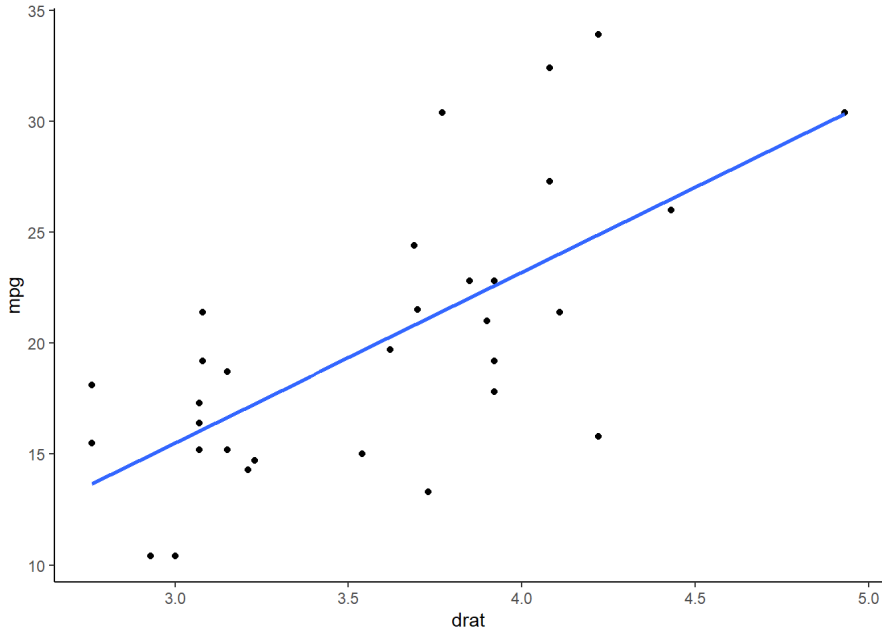
https://blog.uwgb.edu/bansalg/statistics-data-analytics/linear-regression/what-are-the-four-assumptions-of-linear-regression/ (https://blog.uwgb.edu/bansalg/statistics-data-analytics/linear-regression/what-are-the-four-assumptions-of-linear-regression/)

These assumptions were all checked for this data set, and it was found that the cars dataset meets all of the requirements while the housing dataset does not.

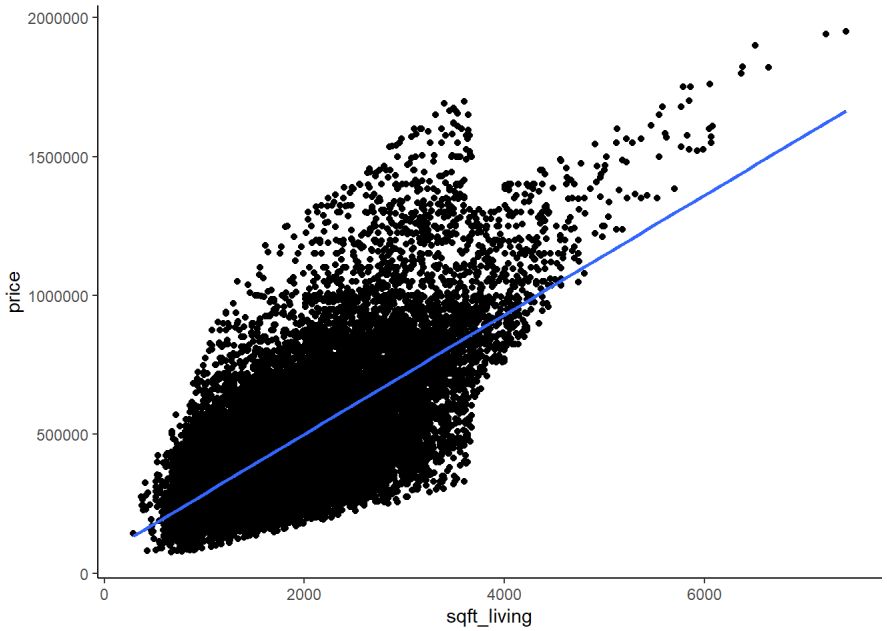
What Does This Mean?

We can draw the conclusion that even though there is a visible correlation in the housing data, the housing dataset is not a good dataset to model using linear regression. This is likely due to the increased number of datapoints clustered together towards the left side of the x-axis. We can demonstrate this by actually performing linear regression on each of these datasets.

Here are the results from a linear regression model for the cars and housing datasets:




```
##
## Call:
## lm(formula = mpg ~ drat, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0775 -2.6803 -0.2095  2.2976  9.0225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7.525      5.477  -1.374     0.18
## drat           7.678      1.507   5.096 1.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.485 on 30 degrees of freedom
## Multiple R-squared:  0.464,  Adjusted R-squared:  0.4461
## F-statistic: 25.97 on 1 and 30 DF,  p-value: 1.776e-05
```



```
##
## Call:
## lm(formula = price ~ sqft_living, data = x_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -513849 -129933  -17336  101228  888502
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70566.386   3499.694   20.16   <2e-16 ***
## sqft_living   214.980      1.635  131.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 184700 on 20817 degrees of freedom
## Multiple R-squared:  0.4538, Adjusted R-squared:  0.4538
## F-statistic: 1.729e+04 on 1 and 20817 DF,  p-value: < 2.2e-16
```

An explanation of each of the result parameters can be found here: http://www.learnbymarketing.com/tutorials/linear-regression-in-r/# (http://www.learnbymarketing.com/tutorials/linear-regression-in-r/#):~:text=lm()%20Function-,Linear%20Regression%20Example%20in%20R%20using%20lm()%20Function,variable%20from%20your%20new%20model.

Let's specifically look at the r-squared value. For linear regression involving only one independent variable, there is not much difference between the multiple and adjusted r-squared values. The r-squared values indicate the strength of the correlation between the x and y variables. The closer the r-squared value is to 1, the more accurate the model is. If the r-squared value is closer to 0, there is a weaker correlation. Notice that in this case, the r-squared values of both models are very similar.

> Just because a correlation is present, does not mean that you should use linear regression

Conclusion

Linear regression can be a useful tool in predicting the cost of your home or what kind of car you should buy. However, in order to develop an accurate model, a set of requirements must be met.

> Lesson materials prepared by Cortland Johns, MITRE