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, using the formula y = mx + b. More information about linear regression can be found at these links:

https://www.geeksforgeeks.org/ml-linear-regression/# (https://www.geeksforgeeks.org/ml-linear-regression/#):~:text=Linear%20Regression%20is%20a%20machine,value%20based%20on%20independent%20variables.&text=Linear%20regression%20perform

https://machinelearningmastery.com/linear-regression-for-machine-learning/ (https://machinelearningmastery.com/linear-regression-for-machine-learning/)

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. This would appear in the tables above as 'NA'.

```
## There are 0 rows with missing data from the cars dataset.
```

```
## There are 0 rows with missing data from the housing dataset.
```

As you can see above, there are no rows with missing data, so we do not have to address missing values.

If these datasets did have missing data, there are several ways to address it.

One good rule of thumb, according to the links below, is that if there is less than 5% of the data in a dataset missing, those data points can simply be dropped without affecting the results of the dataset. Missing data in these small amounts is called Missing Completely At Random, or MCAR data.

For more information about MCAR, check out this link: http://www.hubresearch.ca/bridging-the-data-gap-how-to-deal-with-missing-data-in-observational-studies/# (http://www.hubresearch.ca/bridging-the-data-gap-how-to-deal-with-missing-data-in-observational-studies/#):~:text=As%20a%20rule%20of%20thumb,any%20significant%20ramifications%20(3).&text=In%20this%20case%2C%20it%20is,fill%20in%20the%20miss

Other types of missing data include data that is Missing At Random (MAR) and Missing Not At Random (MNAR). For these larger amounts of missing, it is necessary to perform some kind of imputation - filling in the missing data with substituting values. There are many different methods of imputation, ranging from simply filling in the missing values with the average to interpolating the most likely value of each missing chunk of data based on the values surrounding them.

To read more about MCAR, MAR, and MNAR, check out this link: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0442-1 (https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0442-1)

To read more about the different imputation methods available to address missing data, check out this link: http://www.stat.columbia.edu/~gelman/arm/missing.pdf (http://www.stat.columbia.edu/~gelman/arm/missing.pdf)

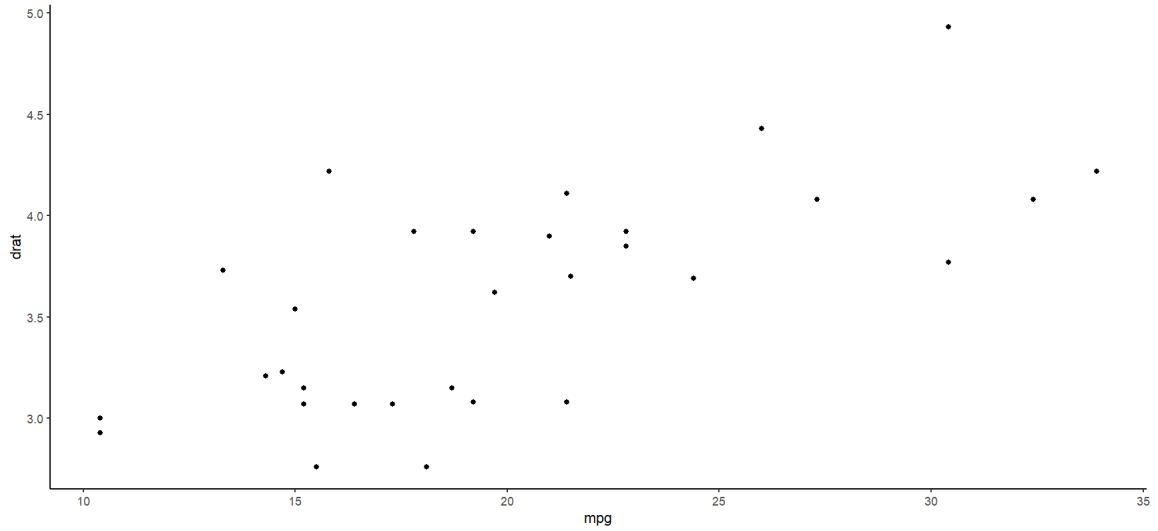
> Q1: Why is it important to ensure that no data is missing when you're doing some kind of data analysis?

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.

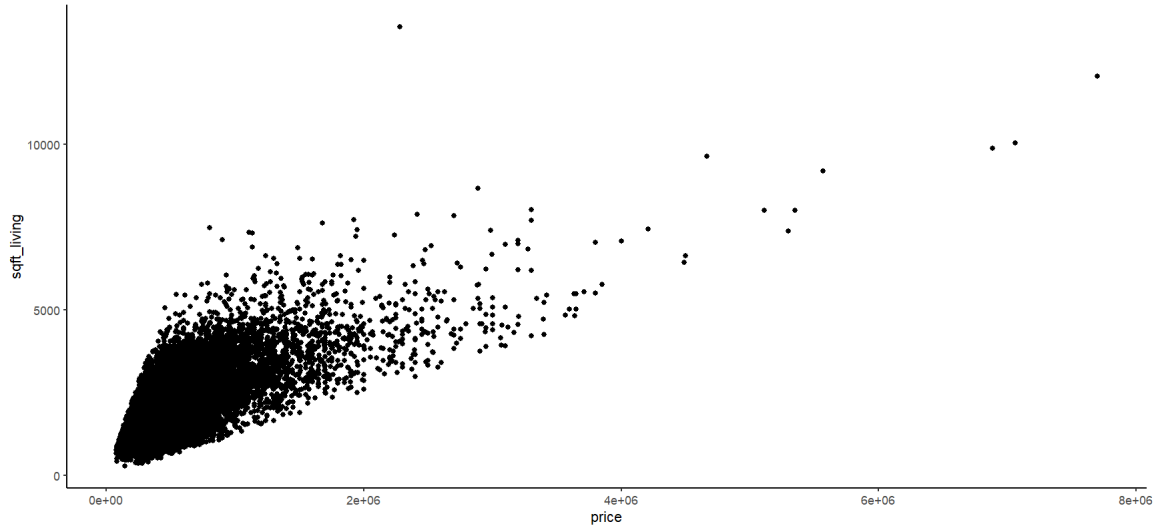
You can use the tool below to examine the relationship between any two variables in the datasets.

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



Q2: 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.



Q3: What correlations do you see with price?

Q4: Why is some data scattered, but others form lines?

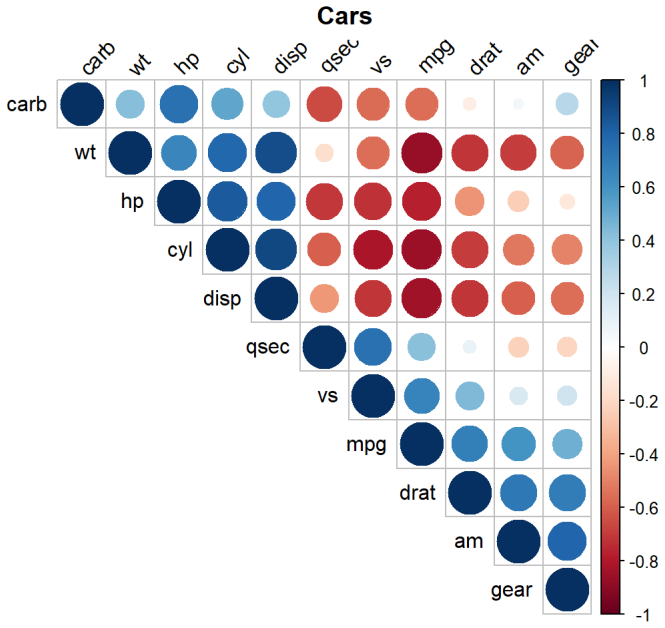
Now play around with this tool, changing variable 1 and variable 2 for both the cars and housing datasets.

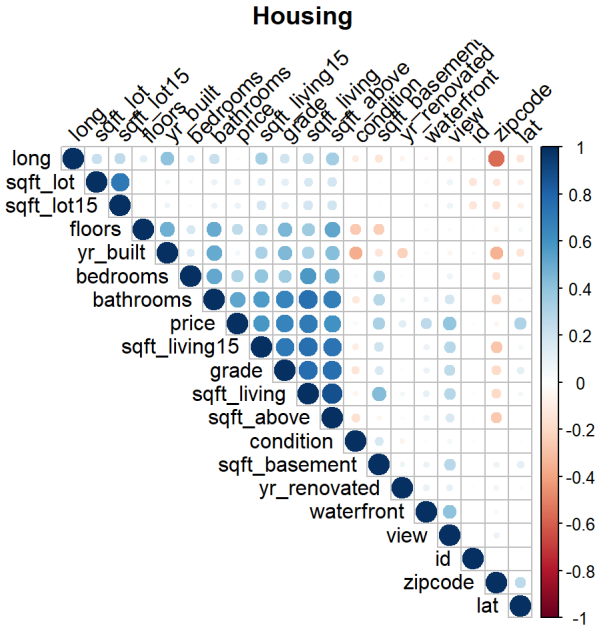
Q5: What other correlations can you find?

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.





Q6: Why is there a perfect positive correlation in the leftmost box of every row?

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.

Q7: 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.

Q8: 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.

Q9: 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.

Q10: 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 data preparation for linear regression is to get rid of any data points that might skew the results of the model. Here's an example of removing these data points using the housing dataset.

The table below shows all of the data points in this correlation that are outliers or high leverage points that significantly affect the data, nearly 800 data points out of over 21,000.

price	sqft_living	.hat	.cooksd	.std.resid
1225000	5420	0.0006582	0.0003071	-0.9656887
2000000	3050	0.0000979	0.0010102	4.5428337
775000	4220	0.0002975	0.0002911	-1.3987419
1040000	4770	0.0004432	0.0002110	-0.9755111
740500	4380	0.0003365	0.0004878	-1.7025070
2250000	5180	0.0005734	0.0029626	3.2135487

> An outlier is a data point that has an extreme y-value.

An example of this is a house with a much higher or lower price than you would expect for a certain square footage based on the linear model. Outliers are identified as data points whose standardized residuals (.std.resid) are greater than 3 or less than -3. The standardized residual represents the number of standard errors away from the linear regression line.

> A high leverage point is a data point that has an extreme x-value.

High leverage points are quantified by the leverage statistic or hat-value (.hat). A hat-value is considered high when the value is greater than the equation 2(p+1)/n, where p equals the number of predictors (in this case, we are predicting price based on one variable, so p = 1) and n equals the number of observations (in this case, 21613). For this dataset, a hat-value greater than 0.0002 indicates a high leverage point.

Just because a value is an outlier or a high leverage point does not mean that value has a significant effect on the data. The significance of a value's influence on the linear model is measured using Cook's distance (.cooksd).

> Cook's distance is a combination of a data point's leverage and residual size.

A value is considered significantly influential with a Cook's distance greater than 4/(n - p - 1), where n equals the number of observations and p equals the number of predictors. For this dataset, a Cook's distance greater than 0.0002 indicates a significantly influential value.

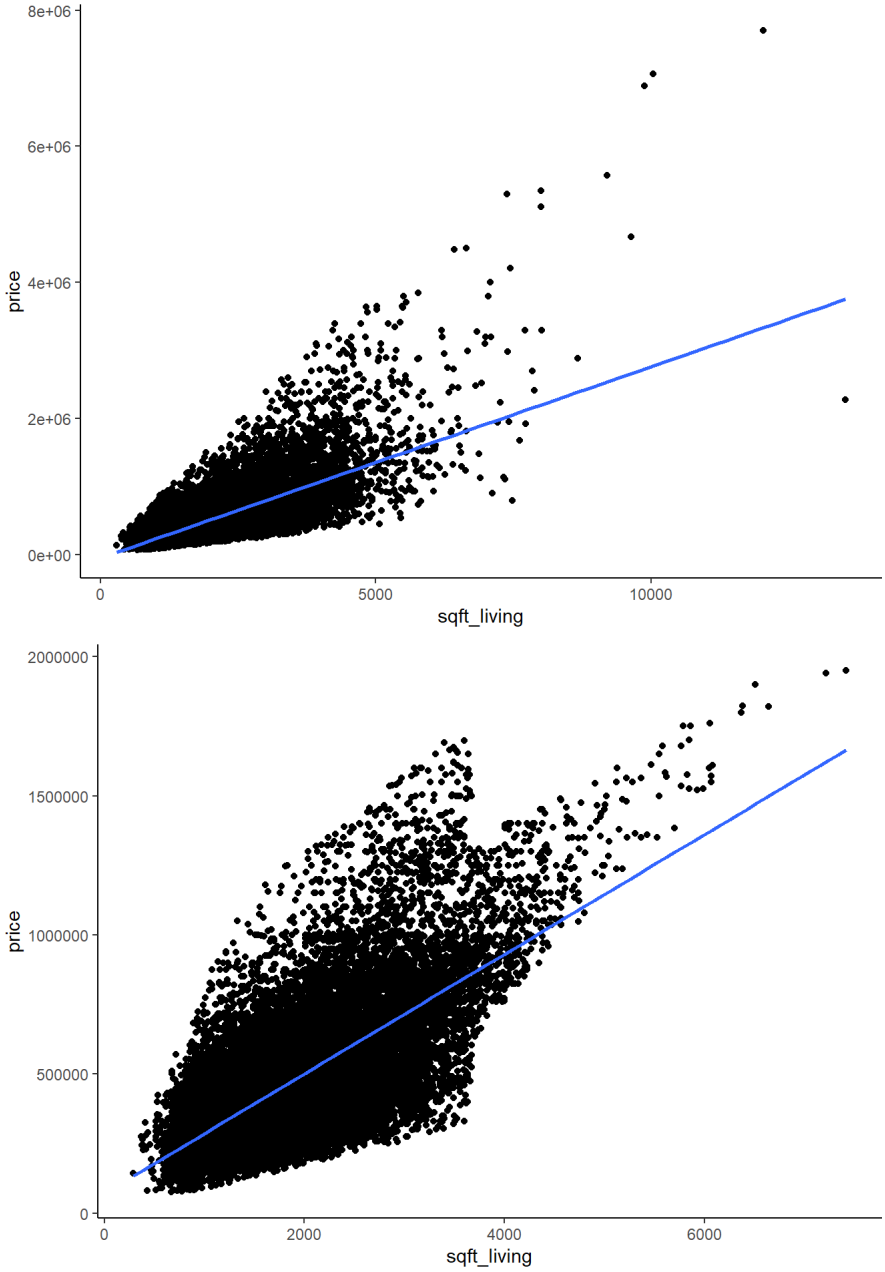
To learn more about how Cook's distance is calculated and utilized, check out these links:

https://www.statisticshowto.com/cooks-distance/ (https://www.statisticshowto.com/cooks-distance/)

https://online.stat.psu.edu/stat462/node/173/ (https://online.stat.psu.edu/stat462/node/173/)

The table above shows all of the outliers and high leverage points that have a significant impact on the outcome of the linear model. These values will be removed from the dataset.

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



Q11: How can outliers affect a dataset? What are the advantages and disadvantages to removing outliers?

Four Assumptions of Linear Regression

There are obviously correlations within the cars and housing datasets, but that doesn't necessarily mean that these datasets 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

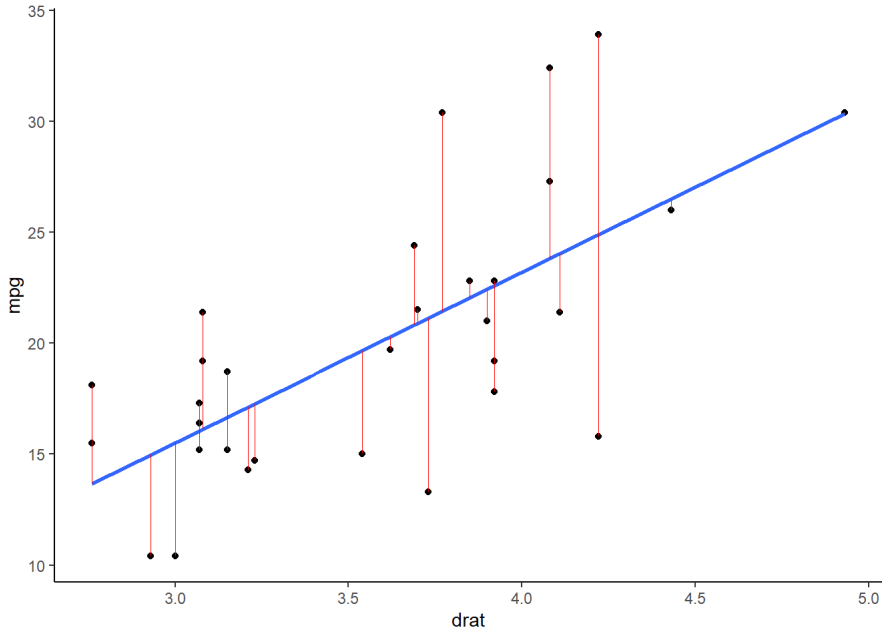
You can read more about these four assumptions at the following websites:

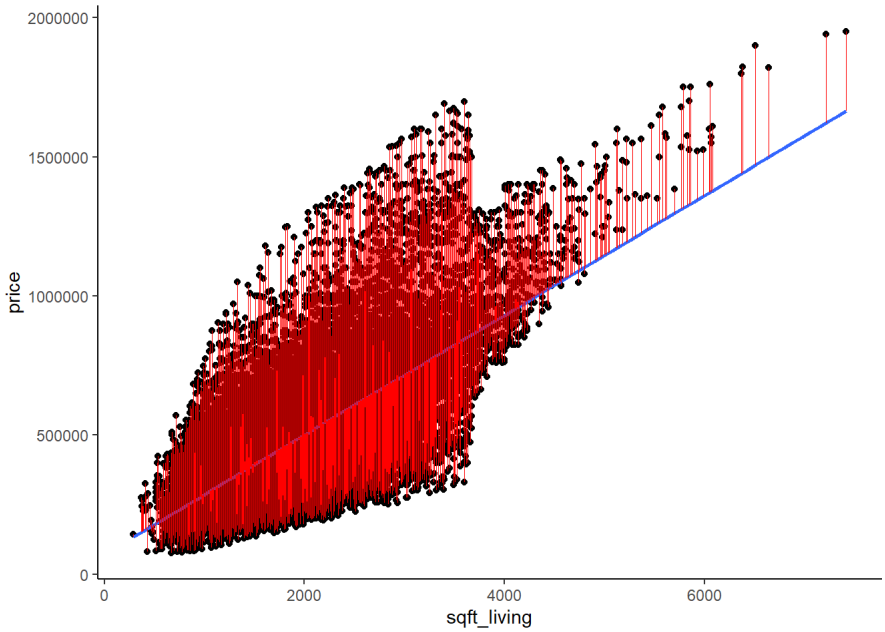
http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/ (http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/)

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/)

In a linear regression model, the residual is the distance from the actual y-value to the predicted value corresponding to the same x-value from the linear model.

The plots below show the residuals of the cars and housing datasets with respect to the regression line.





For the housing dataset, this table shows the actual y-value (price), the corresponding x value (sqft_living), the y-value predicted using linear regression (fitted), and the difference between the actual and predicted y-value (residual). You can fact check this by summing the fitted and residual values, and seeing that they add up to the actual price.

price	sqft_living	.fitted	.resid
221900	1180	324242.5	-102342.49
538000	2570	623064.3	-85064.34
180000	770	236100.8	-56100.79
604000	1960	491926.7	112073.31
510000	1680	431732.4	78267.64
257500	1715	439256.7	-181756.66

For more information about residuals, check out this link:

https://www.statisticshowto.com/residual/#
(https://www.statisticshowto.com/residual/#):~:text=A%20residual%20is%20the%20vertical,at%20that%20point%20is%20zero.

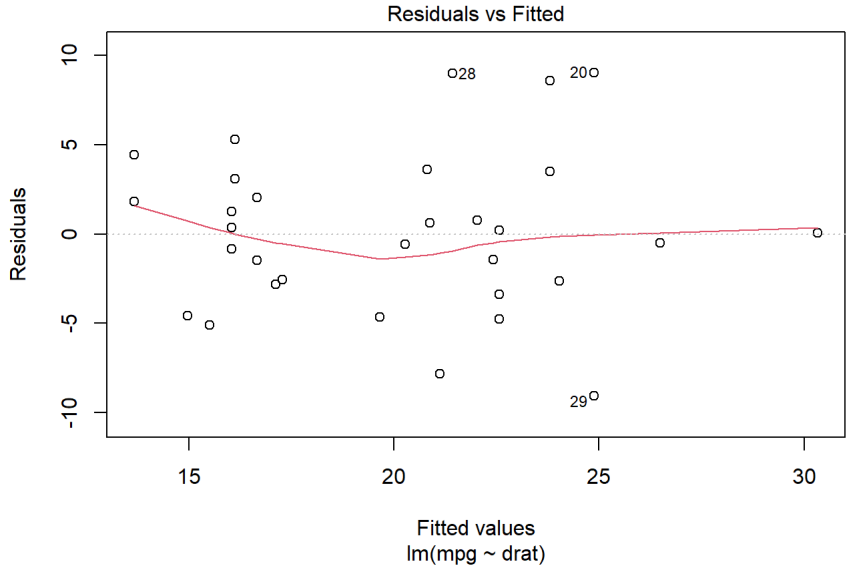
First Assumption

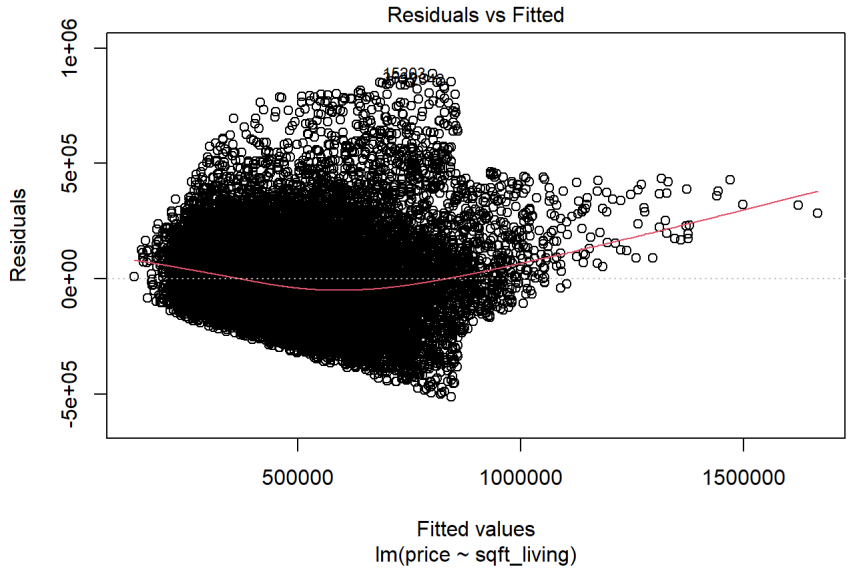
> The first assumption for linear regression is that the data is linear.

This can be tested by making a plot of the residual values versus the fitted values.

If the data is linear, we should see a plot with a horizontal line near zero.

The cars dataset shows a good example of a residuals versus fitted plot. The housing dataset, however, does not show a very good example.



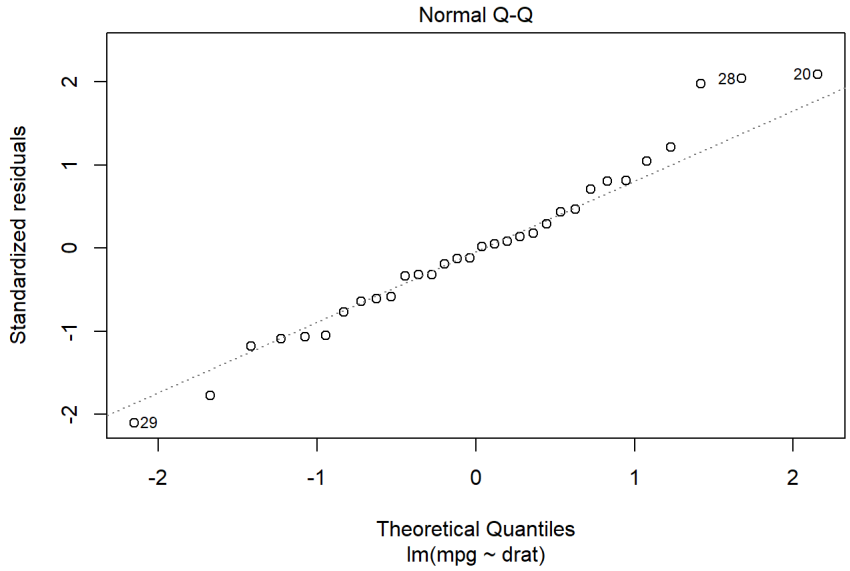


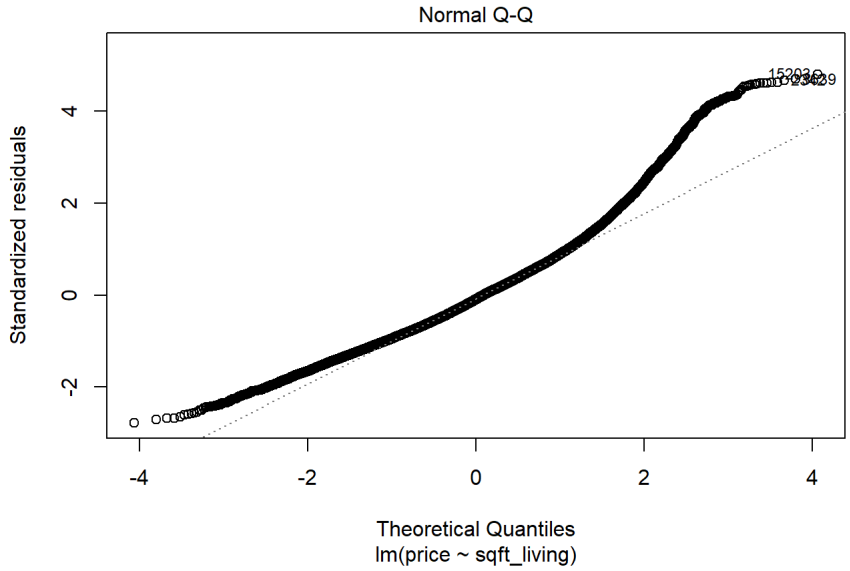
Second Assumption

> The second assumption for linear regression is that the residual values are normally distributed.

This can be tested by making a quantile-quantile (qq) plot, where the standardized residuals are plotted against the theoretical quantiles.

If the residuals are normally distributed, we should see a plot where the circles primarily line up with a diagonal line. Again, the qq plot for the cars dataset shows a good example of a dataset with normally distributed residuals, whereas the housing dataset's residuals are not nearly as normally distributed.



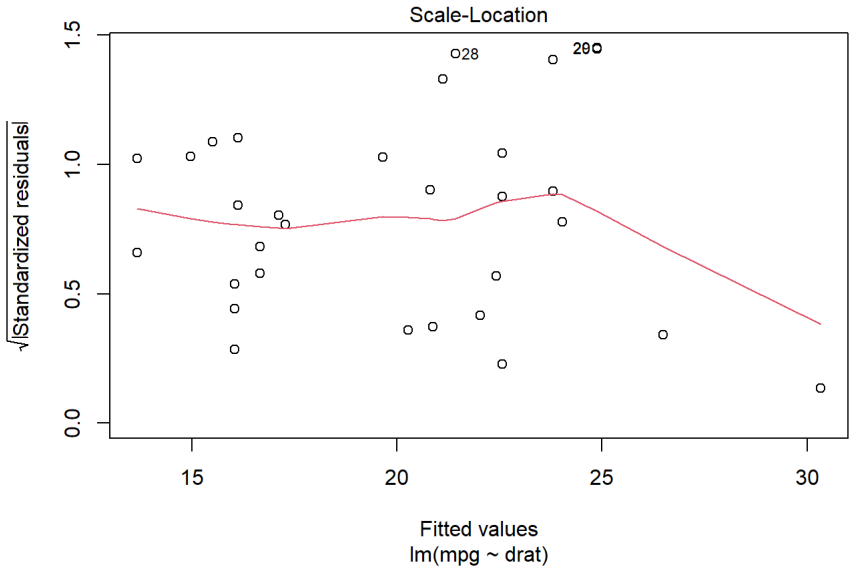


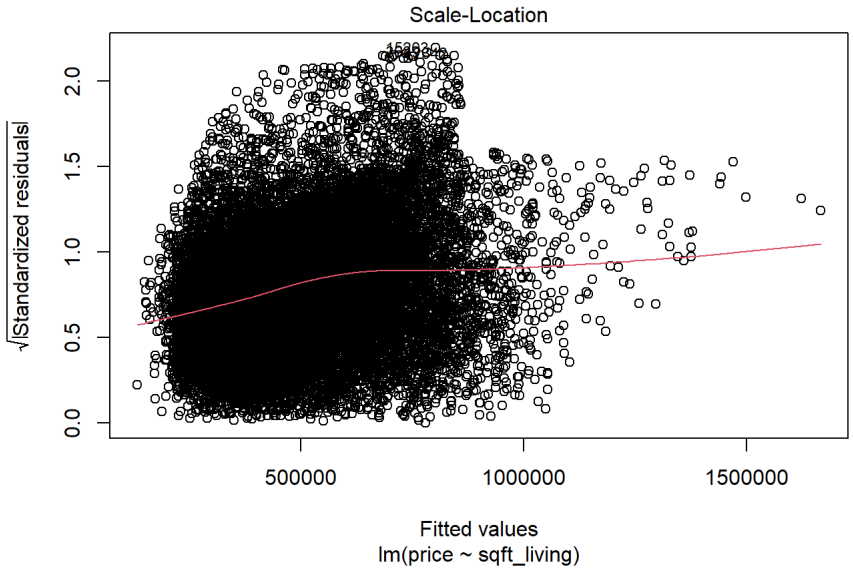
Third Assumption

> The third assumption for linear regression is that the residual values have a constant variance.

This characteristic is called 'homoscedasticity'. This can be tested by making a scale-location or spread-location plot.

If the residuals have a constant variance, we should see a plot with a horizontal line roughly in the center of the individual data points. It is not surprising that the scale-location plot for the cars dataset shows a good example of a dataset with constant residual variance, while the residual variance in the housing dataset is not as constant.





Fourth Assumption

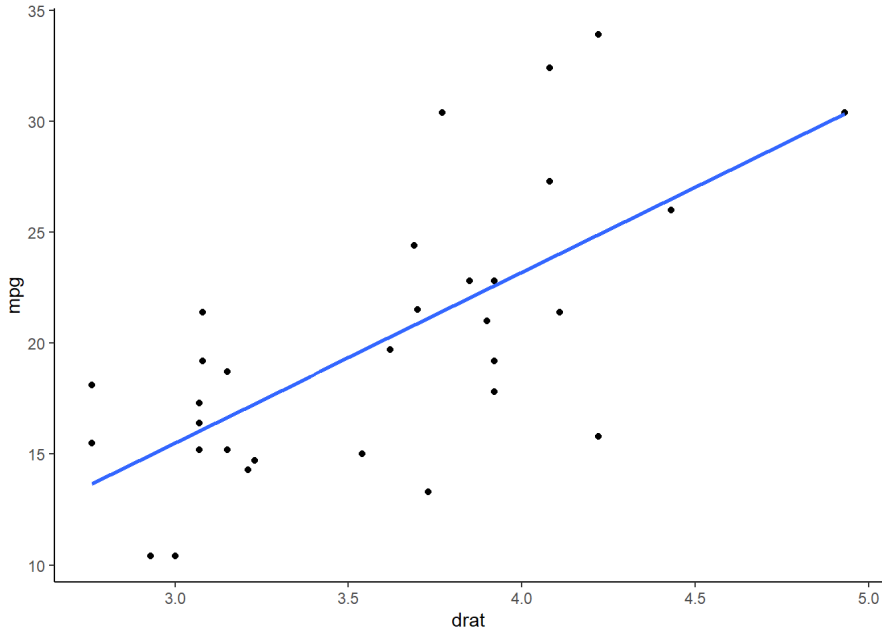
> The fourth assumption for linear regression is that the residual error terms are independent of one another.

This assumption only has to be evaluated for longitudinal data. Longitudinal data is data collected over a long period of time. Because our data was collected for each car or house at one time, like a snapshot of data, we can assume that the residual error terms are independent. An example of a longitudinal dataset would be if someone collected pricing information for one house once a month for 20 years to see how the price changed over time.

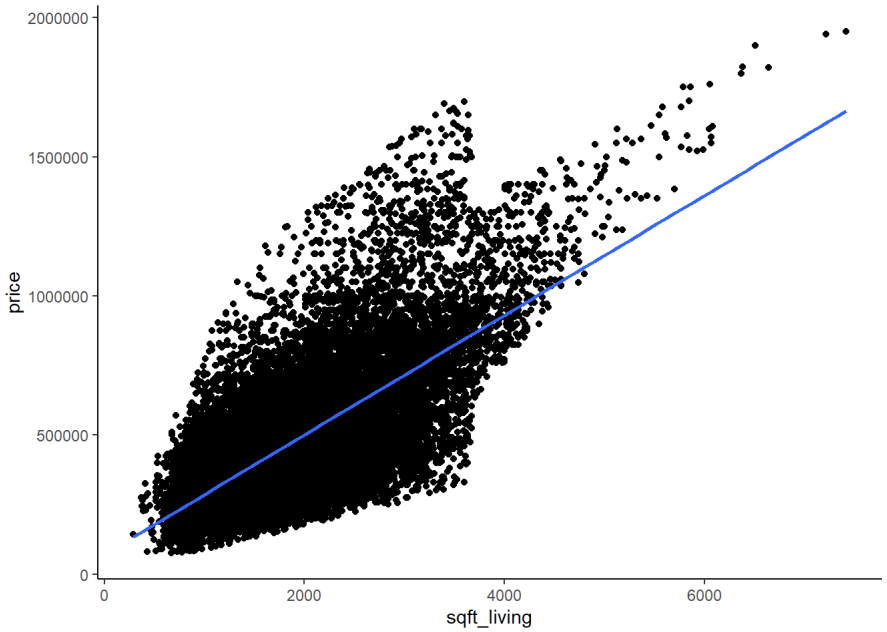
What Does This Mean?

From examining the plots above, 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_clean)
##
## 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 linear regression should be used

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 created by Cortland Johns, MITRE