Introduction to Linear Regression

NASHVILLE SOFTWARE SCHOOL

Least-Squares Regression

Let's say we're interested in houses near downtown Nashville. We gather a sample of 20 homes in the area and look at sales prices for those homes.

In our sample, we find that the average sales price was $482,000.

If we are trying to predict the price of another house in this area, what would our best guess be?

Least-Squares Regression

Let's say we're interested in houses near downtown Nashville. We gather a sample of 20 homes in the area and look at sales prices for those homes.

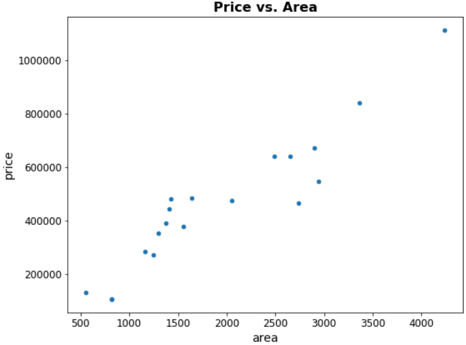
In our sample, we find that the average sales price was $482,000.

If we are trying to predict the price of another house in this area, what would our best guess be?

In the absence of other information, we could go with the average price from our sample, $482,000, but could we do better if we had more information?

Least-Squares Regression

What if we also looked at the square footage of the homes in our sample. Here's a scatterplot of our sample:



If we know that a house we're interested in is 3200 sqft, could we now make a better guess versus just guessing the average price for the area?

Least-Squares Regression

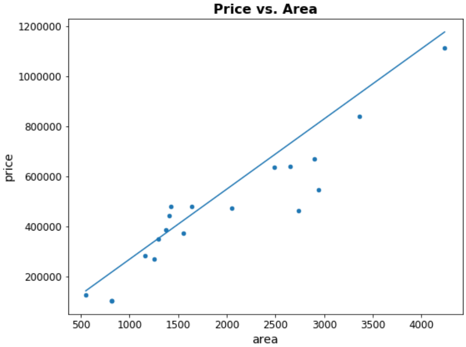
The big idea of least-squares regression is to find a line which describes the relationship between two variables.

Such a line will rarely completely describe the relationship since there will be some uncertainty due either to randomness or due to variables we did not measure.

For example, the price of a home is never completely determined by the square footage, but we can probably make a decent guess about the price by knowing how large the house is.

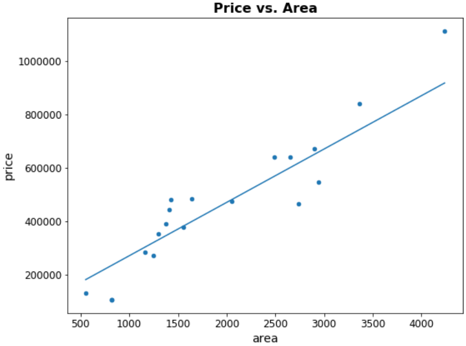
Least-Squares Regression

There are many possible lines that we could use to try and describe the relationship between square footage and price.



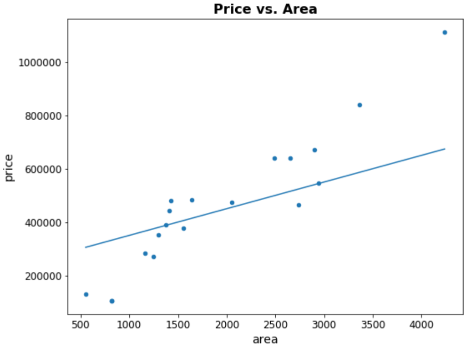
Least-Squares Regression

There are many possible lines that we could use to try and describe the relationship between square footage and price.



Least-Squares Regression

There are many possible lines that we could use to try and describe the relationship between square footage and price.



Least-Squares Regression

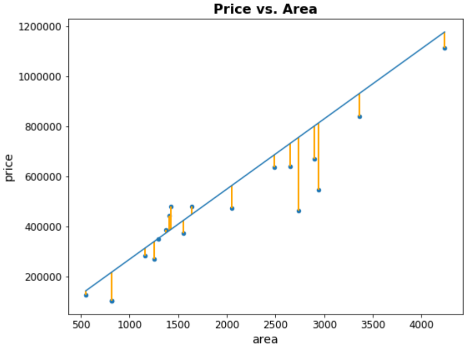
How do we choose the line to use out of all of the options?

We'll choose by looking at the **residuals** - the vertical distance from the line to each point. In this case, the residuals represent the difference between the true price and what we would guess for the price is we used the line to predict price based on square footage.

Because it makes the math work out nicer, we'll really be looking at the squared residuals, but we can get a pretty good idea by looking at the regular residuals.

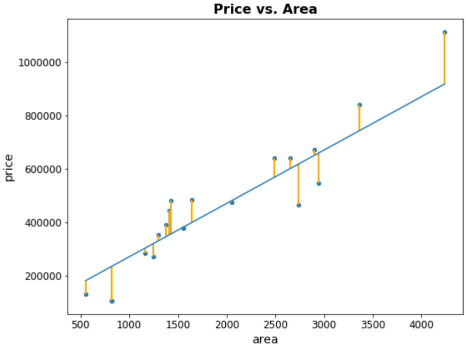
Least-Squares Regression

There are many possible lines that we could use to try and describe the relationship between square footage and price.



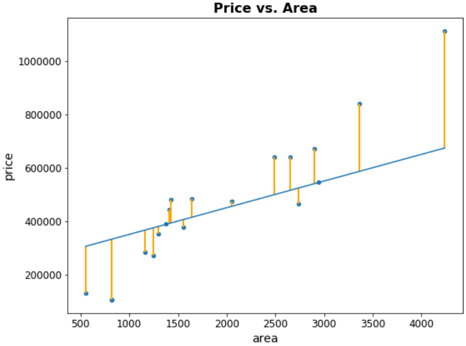
Least-Squares Regression

There are many possible lines that we could use to try and describe the relationship between square footage and price.



Least-Squares Regression

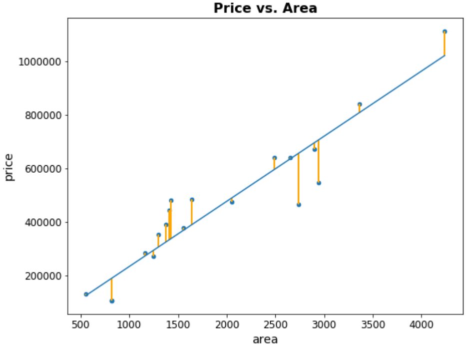
There are many possible lines that we could use to try and describe the relationship between square footage and price.



Least-Squares Regression

It turns out that the line with the smallest squared residuals is this one, with equation

$$\text{price} = 243.49 \cdot (\text{sqft}) - 10970$$



NASHVILLE SOFTWARE SCHOOL

Least-Squares Regression

Least-Squares Regression Line:

$$\text{price} = 243.49 \cdot (\text{sqft}) - 10970$$

This equation tells us that for every additional one square foot of area, the price tends to increase by about \$243.49.

Least-Squares Regression

Least-Squares Regression Line:

$$\text{price} = 243.49 \cdot (\text{sqft}) - 10970$$

This equation tells us that for every additional one square foot of area, the price tends to increase by about \$243.49.

This means that increasing the square footage of a house by 50 square feet will tend to increase the price by about

Least-Squares Regression

Least-Squares Regression Line:

$$\text{price} = 243.49 \cdot (\text{sqft}) - 10970$$

This equation tells us that for every additional one square foot of area, the price tends to increase by about \$243.49.

This means that increasing the square footage of a house by 50 square feet will tend to increase the price by about

$$50 \cdot \$243.49$$

Least-Squares Regression

Least-Squares Regression Line:

$$\text{price} = 243.49 \cdot (\text{sqft}) - 10970$$

This equation tells us that for every additional one square foot of area, the price tends to increase by about \$243.49.

This means that increasing the square footage of a house by 50 square feet will tend to increase the price by about

$$50 \cdot \$243.49 = \$12,174.50$$

Least-Squares Regression

Least-Squares Regression Line:

$$\text{price} = 243.49 \cdot (\text{sqft}) - 10970$$

Also, for a house that is 3200 sqft, we can expect the price to be around

Least-Squares Regression

Least-Squares Regression Line:

$$\text{price} = 243.49 \cdot (\text{sqft}) - 10970$$

Also, for a house that is 3200 sqft, we can expect the price to be around

$$\$243.49 \cdot 3200 - \$10,970$$

Least-Squares Regression

Least-Squares Regression Line:

$$\text{price} = 243.49 \cdot (\text{sqft}) - 10970$$

Also, for a house that is 3200 sqft, we can expect the price to be around

$$\$243.49 \cdot 3200 - \$10,970 = \$768,198$$

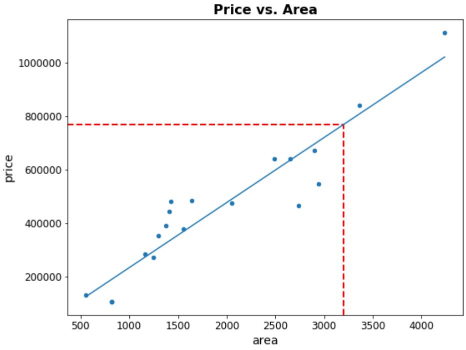
Least-Squares Regression

Least-Squares Regression Line:

$$\text{price} = 243.49 \cdot (\text{sqft}) - 10970$$

Also, for a house that is 3200 sqft, we can expect the price to be around

$$\$243.49 \cdot 3200 - \$10,970 = \$768,198$$

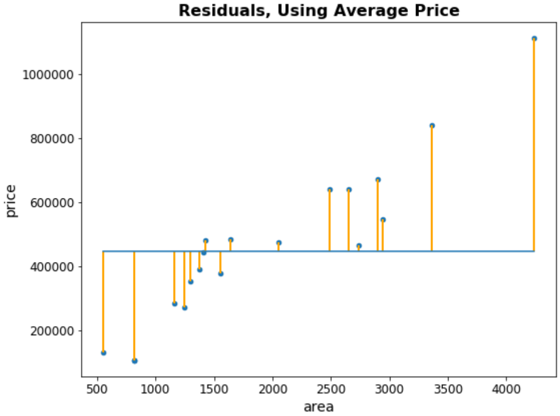


Least-Squares Regression

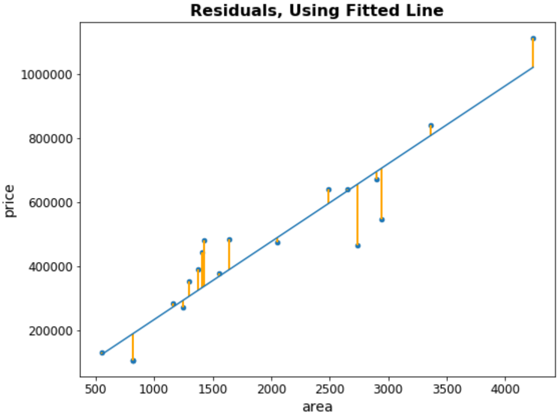
We can quantify how well our line fits the data by using the coefficient of determination, or R^2.

We can understand R^2 by comparing the resiudals for our fitted line versus the residuals if we only used the average price to make our predictions.

Least-Squares Regression



Least-Squares Regression



Least-Squares Regression

The coefficient of determination, R^2 roughly measures the proportion by which we reduce the residuals by using the fitted line instead of just using the average price to predict.

Least-Squares Regression

The coefficient of determination, R^2 roughly measures the proportion by which we reduce the residuals by using the fitted line instead of just using the average price to predict.

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

Here TSS is the total squared residuals from using the average home price and RSS is the squared residuals from using the regression line.

The coefficient of determination will take values between 0 and 1, with 1 representing a perfect fit.

Least-Squares Regression

The coefficient of determination, R^2 roughly measures the proportion by which we reduce the residuals by using the fitted line instead of just using the average price to predict.

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

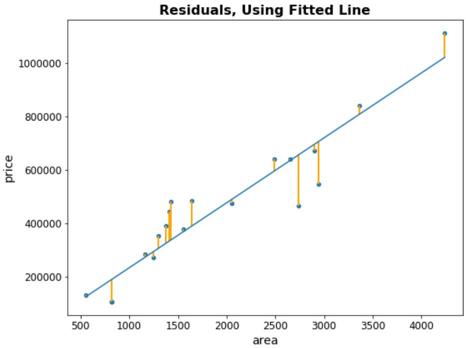
Here TSS is the total squared residuals from using the average home price and RSS is the squared residuals from using the regression line.

The coefficient of determination will take values between 0 and 1, with 1 representing a perfect fit.

In this case, the residuals were reduced by a significant amount. Consequently, the R^2 value is equal to 0.886.

Least-Squares Regression

There are a number of other metrics to evaluate a linear model in addition to R^2. For example, the **mean absolute error** measures the average magnitude of the residuals.



Here, the mean absolute error is equal to \$66,043.55. That means that, on average, the predictions from the line are off by \$66,043.55.

Least-Squares Regression

Now that you have seen the basic concepts of linear regression, let's see how we can create linear regression lines using Python, including how to include multiple predictor variables.