Introduction to Linear Regression

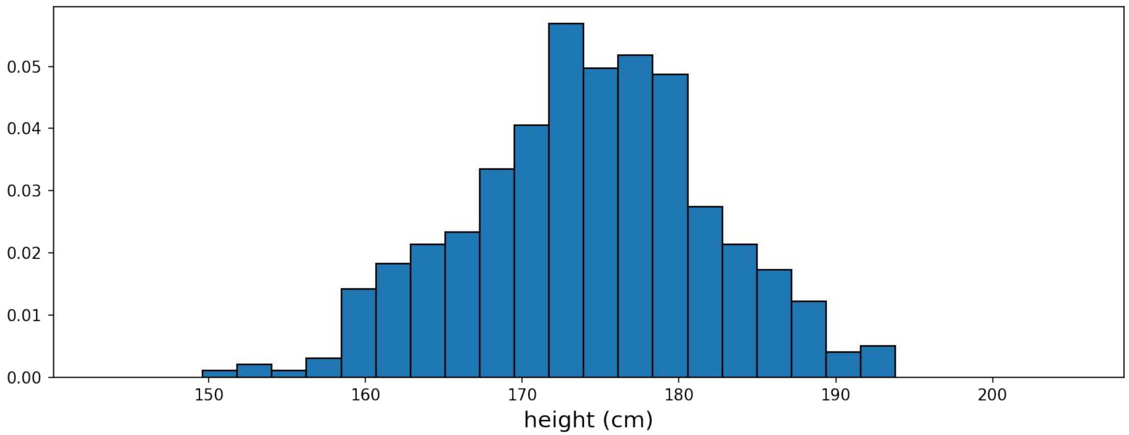
Statistical Models

A **statistical model** is a mathematical model that embodies a set of statistical assumptions concerning the generation of sample data (and similar data from a larger population).

A statistical model represents, often in considerably idealized form, the data-generating process.

Statistical Models

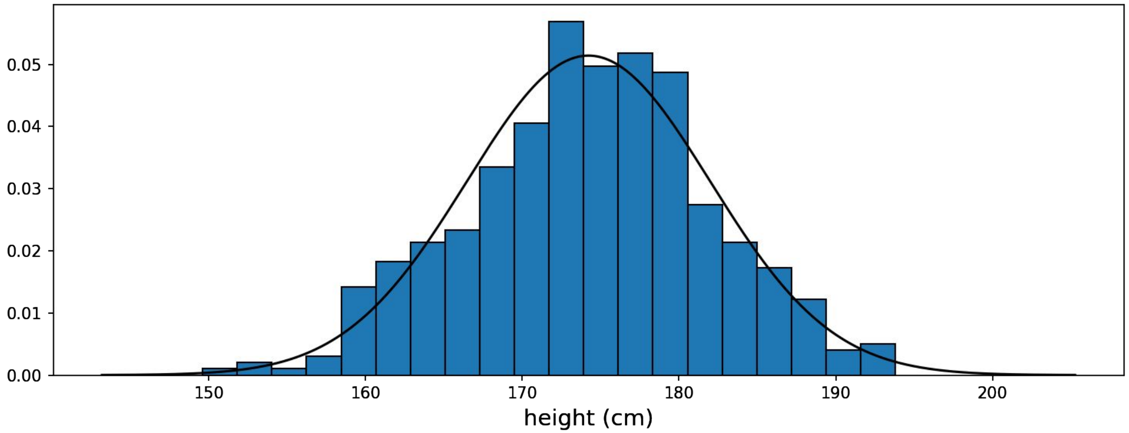
Example: In week 3, we looked at height data from the NHANES survey.



Statistical Models

Example: In week 3, we looked at height data from the NHANES survey.

We determined that we could describe this data quite well using a normal distribution with mean μ = 174.22 cm and standard deviation σ = 7.761.



Statistical Models

Purpose of a Statistical Model:

- **Summary of the Data**: Rather than retaining the dataset, just remember the model parameters.

From our example above, we only need to remember μ = 174.22 and σ = 7.761.

Statistical Models

Purpose of a Statistical Model:

- **Making Predictions:** Trying to say what other data will look like. We don’t try to be exactly right, but at least be able to say that our errors will show probabilistic patterns.

Statistical Models

Purpose of a Statistical Model:

- **Simulating the Data Generation Process:** Trying to say what new data would look like. This is closer to a *scientific* model rather than just a statistical one.

Statistical Models - Linear Regression

So far, we have seen a model for a single variable, but what if we want to study the relationship between two or more variables?

Linear regression is a method that allows us to study relationships between two (or more) quantitative variables.

- One variable, often denoted by *y*, is called the **response**, **outcome**, **dependent**, or **target variable**.
- The other variable, often denoted by *x*, is called the **predictor**, **explanatory**, or **independent variable**.

Example - Home Prices

Let's say we're interested in houses in 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?

Example - Home Prices

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

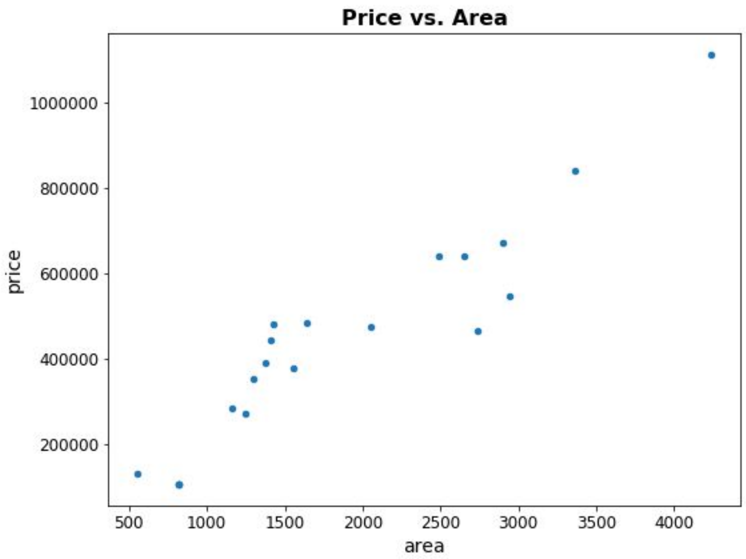
In our sample, we find that the average sales price was $482,000.

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?

Example - Home Prices

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?

Example - Home Prices

Big Idea: When performing regression, we are investigating whether the expected value of the target variable (here, sales price) depends on the predictor variable (here, sqft).

Example - Home Prices

Big Idea: When performing regression, we are investigating whether the expected value of the target variable (here, sales price) depends on the predictor variable (here, sqft).

In general, the relationship between two variables can be quite complex and difficult to estimate, so what we can make a simplifying assumption: the expected value of the target variable is a linear function of the value of the predictor variable.

Example - Home Prices

$$E[\text{price}|\text{sqft}] = \beta_0 + \beta_1 \cdot (\text{sqft})$$

Example - Home Prices

$$E[\text{price}|\text{sqft}] = \beta_0 + \beta_1 \cdot (\text{sqft})$$

average home price
for a given home
area (in sqft)

Example - Home Prices

$$E[\text{price}|\text{sqft}] = \beta_0 + \beta_1 \cdot (\text{sqft})$$

average home price for a given home area (in sqft)

slope: the change in average price for corresponding to a one-unit increase in sqft

Example - Home Prices

$$E[\text{price}|\text{sqft}] = \beta_0 + \beta_1 \cdot (\text{sqft})$$

This is often written in a different form:

$$\text{price} = \beta_0 + \beta_1 \cdot (\text{sqft}) + \epsilon$$

random error term

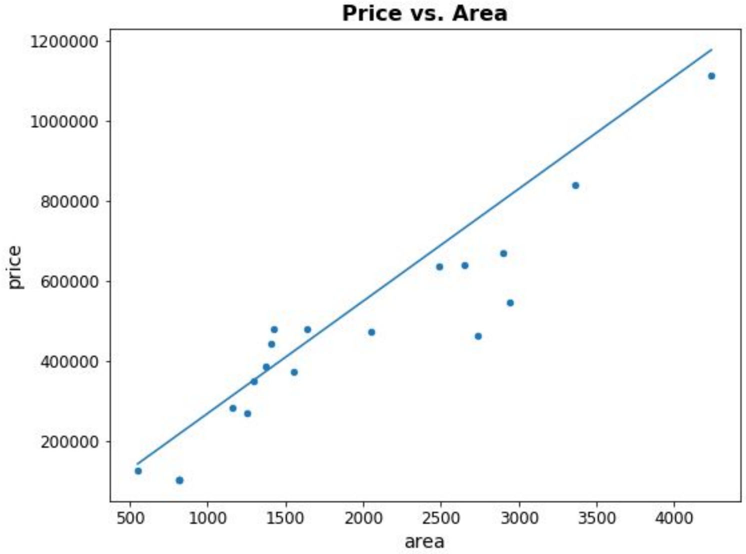
Example - Home Prices

$$\text{price} = \beta_0 + \beta_1 \cdot (\text{sqft}) + \epsilon$$

Finding the values of the β coefficients corresponds to finding a line that describes the relationship between the square footage of a home and its sales price.

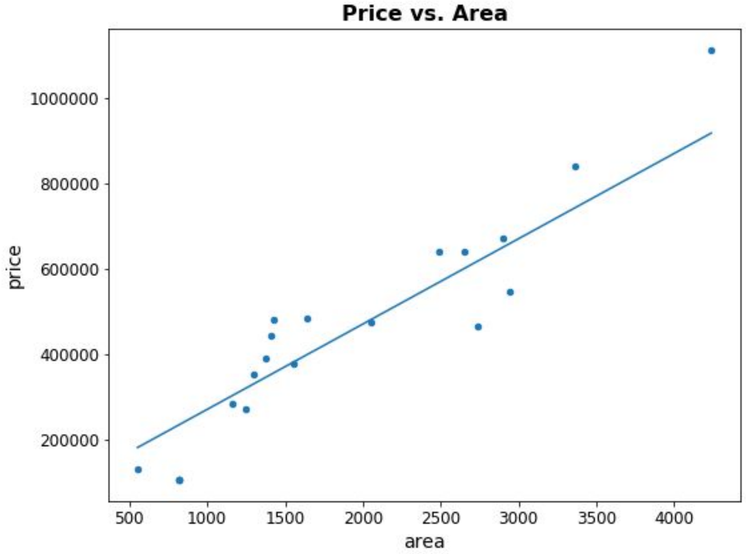
Example - Home Prices

There are many possible lines that we could use to try and describe this relationship.



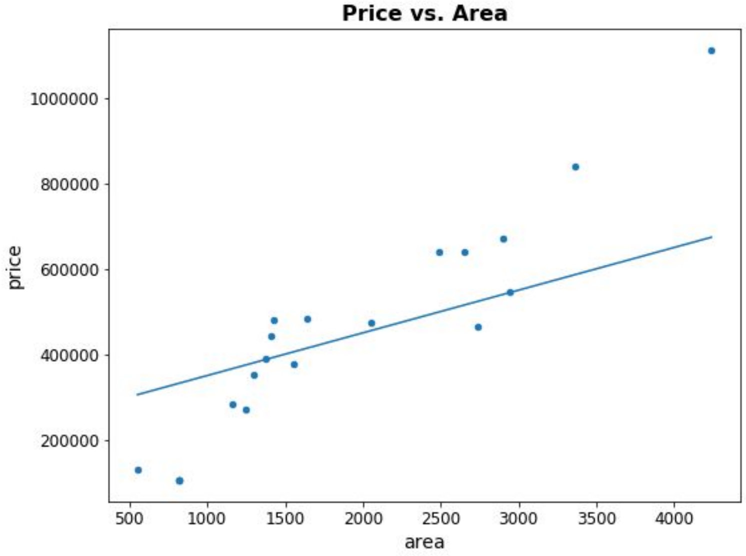
Example - Home Prices

There are many possible lines that we could use to try and describe this relationship.



Example - Home Prices

There are many possible lines that we could use to try and describe this relationship.



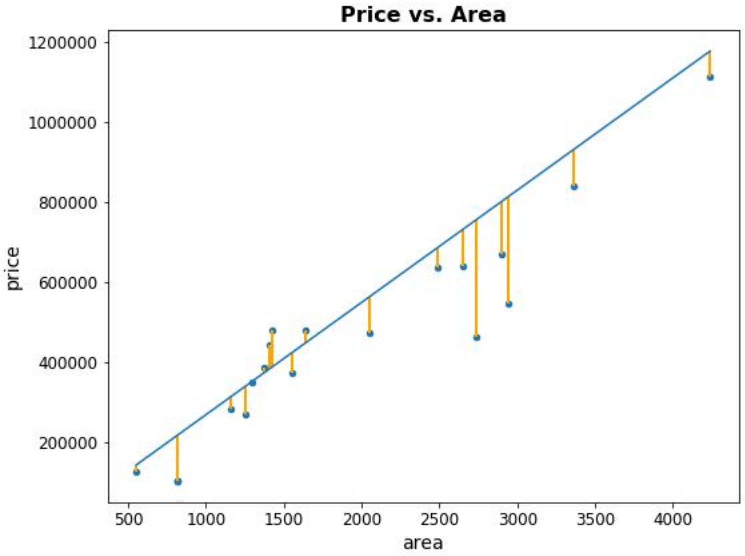
Example - Home Prices

How do we choose the “best” line?

Example - Home Prices

How do we choose the “best” line?

When performing **least squares regression**, what we do is to look at the **residuals** - the vertical distance between the line and the observation point.

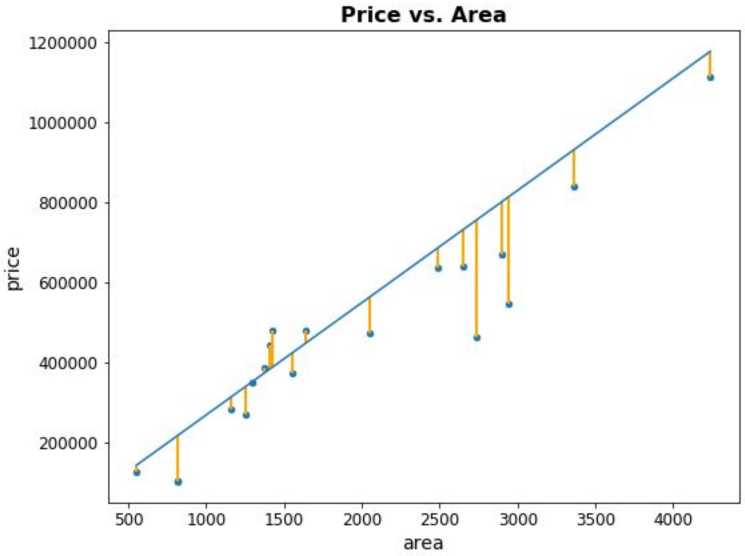


Example - Home Prices

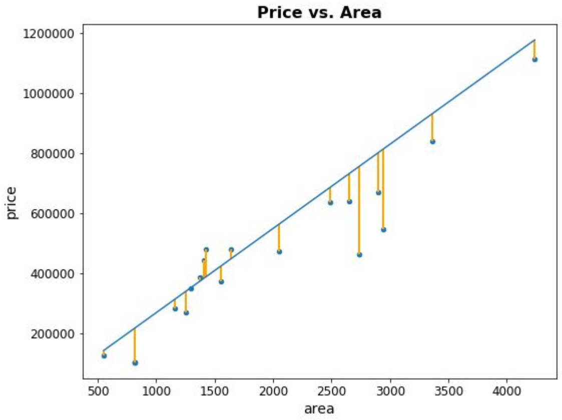
How do we choose the “best” line?

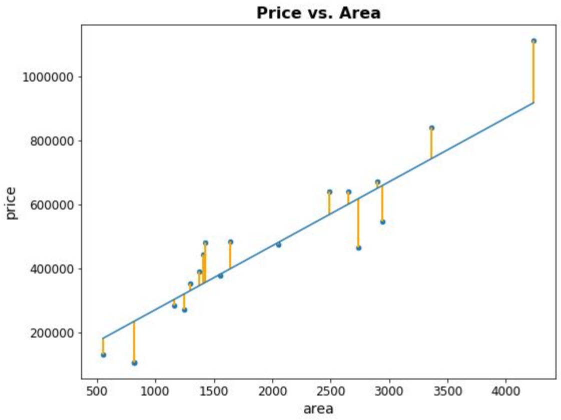
When performing **least squares regression**, what we do is to look at the **residuals** - the vertical distance between the line and the observation point.

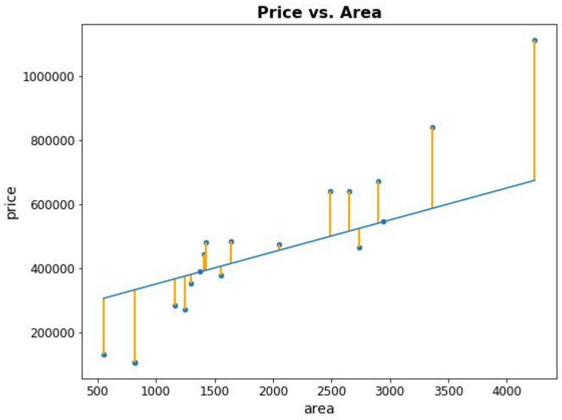
We actually look at the squared residuals since it makes the math work out better.



Example - Home Prices



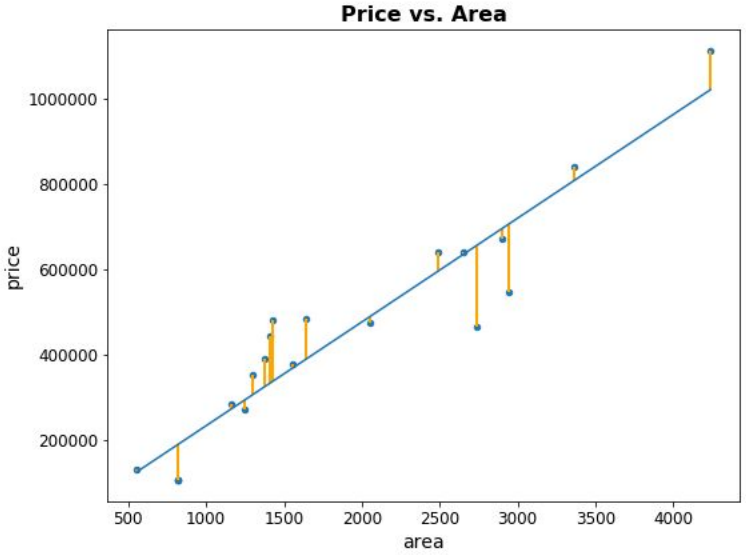




Example - Home Prices

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

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

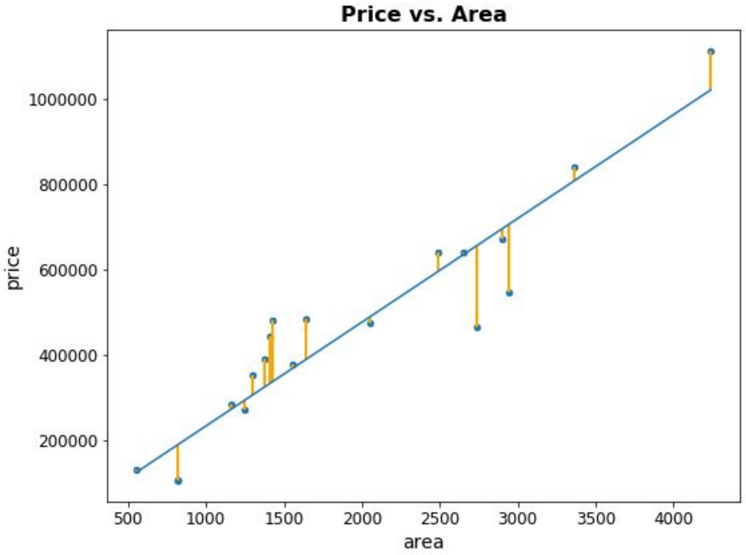


Example - Home Prices

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

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

For every additional one square foot of area, the average price increases by $243.49.

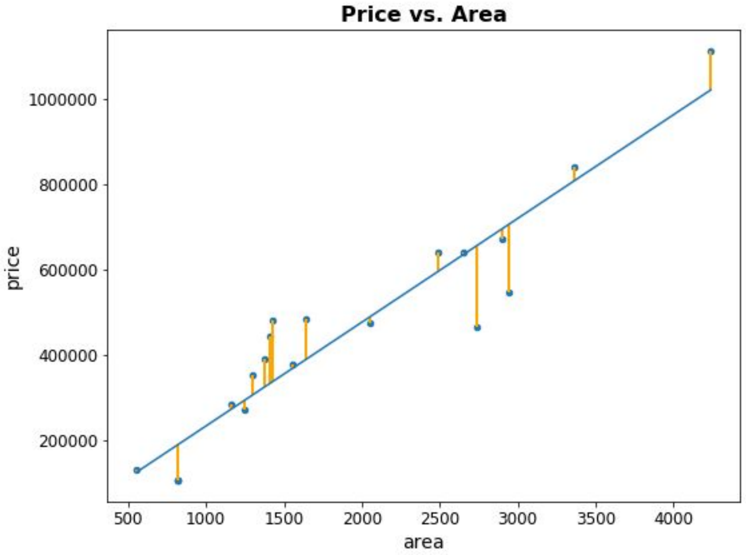


Example - Home Prices

We can ask how “good” this model is

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

A measure of this is the **coefficient of determination**, R^2.



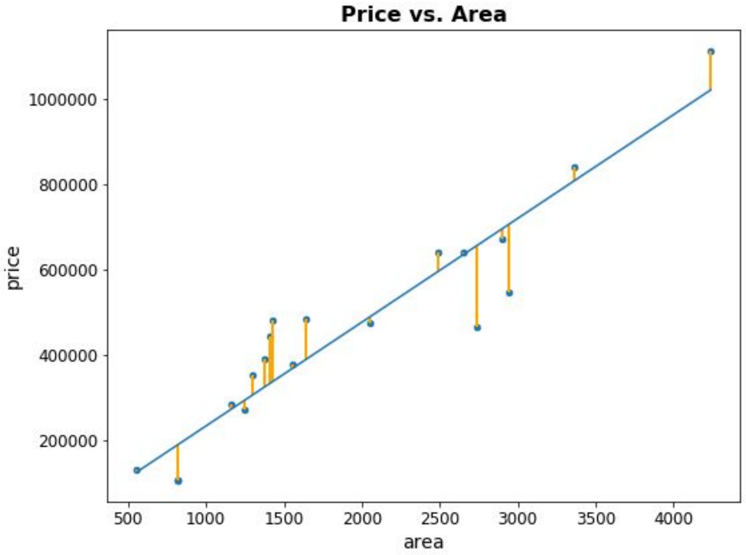
Example - Home Prices

We can ask how “good” this model is

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

A measure of this is the **coefficient of determination**, R^2.

One way of understanding R^2 is that it compares the residuals from the model to what they would be if we only used the average price to make predictions.



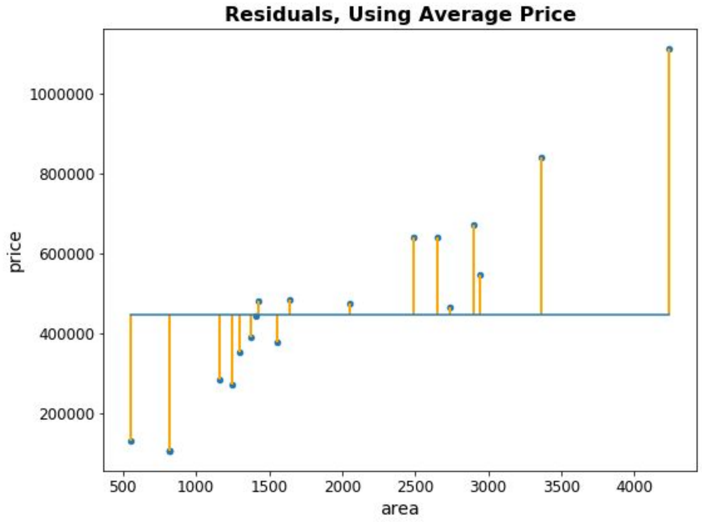
Coefficient of Determination

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

Coefficient of Determination

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

$$TSS = \sum_i (y_i - \bar{y})^2$$

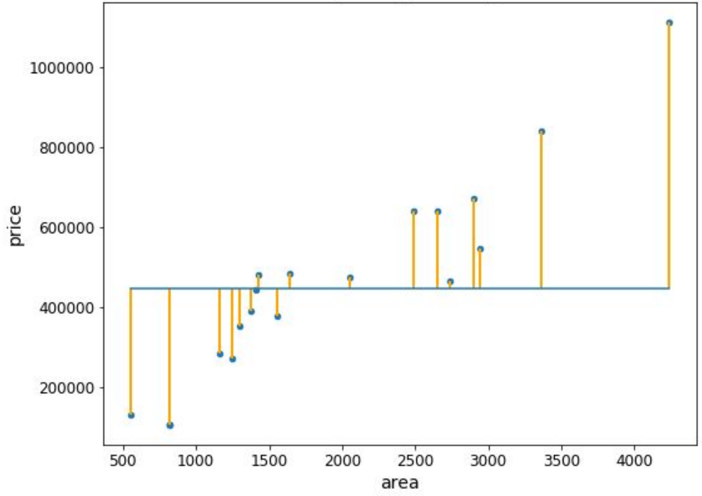


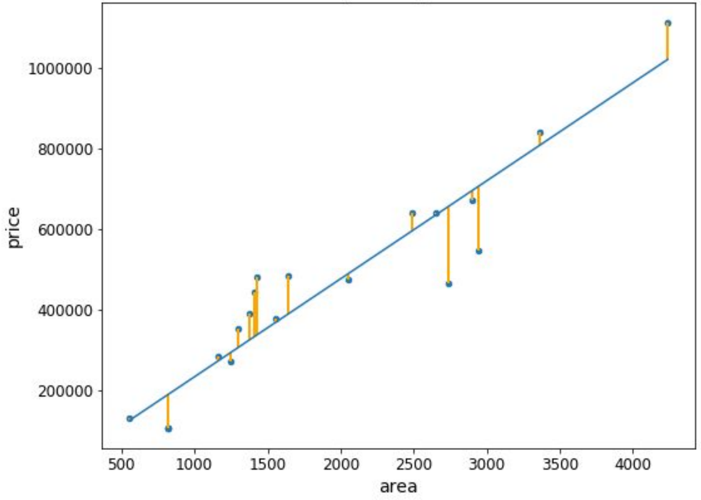
Coefficient of Determination

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

$$TSS = \sum_i (y_i - \bar{y})^2$$

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$


Residuals, Using Average Price


Residuals, Using Fitted Line

Coefficient of Determination

This measures the reduction in residuals (as a percentage) that you get by using the model compared to using just the average.

A perfect model (one which no residuals) would have an R^2 value of 1. In this case, we have a R^2 value of 0.886.

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

Example - Home Prices

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

Based on this, what do we estimate the average price of a 3200 square foot home will be?

Example - Home Prices

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

Based on this, what do we estimate the average price of a 3200 square foot home will be?

$$\text{price} = -10970 + 243.49 \cdot (3200)$$

$$= \$768{,}198$$

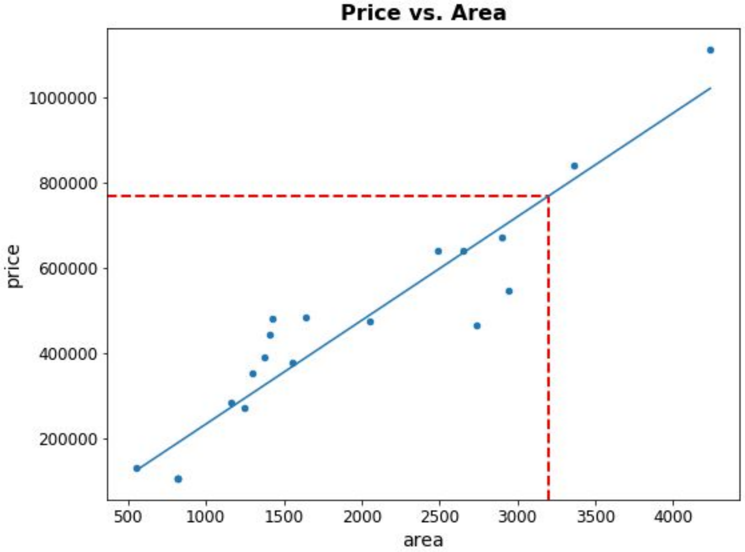
Example - Home Prices

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

Based on this, what do we estimate the average price of a 3200 square foot home will be?

$$\text{price} = -10970 + 243.49 \cdot (3200)$$

$$= \$768{,}198$$



Example - Home Prices

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

Based on our dataset, we have our estimated line, but there are a number of questions that we can ask from here:

- Is the relationship statistically significant? (Is it possible that the slope is actually 0?)
- How precise is our estimate of the slope?
- How precise is our estimate of the average price of a 3200 square foot home?
- What price can we predict for a new observation with area 3200?

The Linear Regression Model

So far, all we've done is to assume that the average response variable can be described as a linear function of the predictor variable(s). If we want to be able to do additional inference, we need to make some additional assumptions. The four assumptions of the linear regression model are as follows:

The Linear Regression Model

So far, all we've done is to assume that the average response variable can be described as a linear function of the predictor variable(s). If we want to be able to do additional inference, we need to make some additional assumptions. The four assumptions of the linear regression model are as follows:

- **L**inear Function: The mean of the response at each value of the predictor is a linear function of the predictor variable.

The Linear Regression Model

So far, all we've done is to assume that the average response variable can be described as a linear function of the predictor variable(s). If we want to be able to do additional inference, we need to make some additional assumptions. The four assumptions of the linear regression model are as follows:

- **L**inear Function: The mean of the response at each value of the predictor is a linear function of the predictor variable.
- **I**ndependent: The errors, ε are independent.

The Linear Regression Model

So far, all we've done is to assume that the average response variable can be described as a linear function of the predictor variable(s). If we want to be able to do additional inference, we need to make some additional assumptions. The four assumptions of the linear regression model are as follows:

- **L**inear Function: The mean of the response at each value of the predictor is a linear function of the predictor variable.
- **I**ndependent: The errors, ε are independent.
- **N**ormally Distributed: The errors at each value of the predictor are normally distributed.

The Linear Regression Model

So far, all we've done is to assume that the average response variable can be described as a linear function of the predictor variable(s). If we want to be able to do additional inference, we need to make some additional assumptions. The four assumptions of the linear regression model are as follows:

- **L**inear Function: The mean of the response at each value of the predictor is a linear function of the predictor variable.
- **I**ndependent: The errors, ε are independent.
- **N**ormally Distributed: The errors at each value of the predictor are normally distributed.
- **E**qual variances: The errors at each value of the predictor have equal variances.

The Linear Regression Model

$$\text{price} = \beta_0 + \beta_1 \cdot (\text{sqft}) + \epsilon$$

When we make the assumptions of the linear regression model, we can think of it as saying that the error terms, ε, are normally distributed with a mean of 0 and a constant (across all values of the predictor variable) variance.

Example - Home Prices

Questions about the slope:

- Is the relationship statistically significant? (Is it possible that the slope is actually 0?)
- How precise is our estimate of the slope?

These can be addressed by performing inference on β_1. One can show that the sampling distribution of the estimated slopes follows a normal distribution, which lets us either perform a hypothesis test or build a confidence interval.

Example - Home Prices

Questions about predictions:

- How precise is our estimate of the average price of a 3200 square foot home?
- What price can we predict for a new observation with area 3200?

Note that these questions are asking different things - the first is about the *average* price, but we are assuming that there will some variability in the prices for a given area, so when predicting a new price, we should consider this variability.

Example - Home Prices

Questions about predictions:

- How precise is our estimate of the average price of a 3200 square foot home?
- What price can we predict for a new observation with area 3200?

In the first case, we can build a confidence interval for the average response.

In the second, we can build what is called a **prediction interval**.

Linear Regression in Python

We'll be using the *statsmodels* library to perform linear regression.

Let's see this in action in the notebook.