**Microsoft:** DAT210x Programming with Python for Data Science

Bookmarks

- Start Here
- 1. The Big Picture
- 2. Data And Features
- 3. Exploring Data
- 4. Transforming Data
- **5. Data Modeling**

**Lecture: Clustering**
Quiz

**Lab: Clustering**
Lab

**Lecture: Splitting Data**
Quiz

**Lecture: K-Nearest Neighbors**
Quiz

**Lab: K-Nearest Neighbors**



# Linear Regression Gotchas!

MOD37

Lab

**Lecture: Regression**
Quiz

**Lab: Regression**
Lab

**Dive Deeper**

6. Data Modeling II

0:00 / 2:03 1.0x

Download video Download transcript .srt

Linear regression is a very powerful technique if used correctly. With just a few instances of well correlated samples, linear regression can capture the underlying pattern in your dataset, making its use of your data very efficient. This applies even more so to smaller datasets.

All of the math and theory associated with linear regression is well understood, which means by modeling with it, even the multi-output version of it supported by SciKit-Learn, the results remain easily interpretable.

Linear regression isn't all roses and cherries. For one, as the name hints, it only works with linear data, by identifying linear relationships between your continuous output and your independent input features. Sometimes there simply isn't a linear correlation, and in these cases the regression completely fails. Luckily, it's pretty easy to detect this by looking at the resulting $R^2$ coefficient, or by visually plotting your data.

Under the hood, linear regression examines the relationship between the mean value of your output variable and your input variables. So if you're trying to model stock market security price as a function of the date, linear regression will only factor in the average stock price taken at different date intervals. If you wanted to know what the highest and lowest values were for any date interval, linear regression wouldn't be able to provide that.

Of all this, the major thing to watch out for while using linear regression (even more so than its sensitivity to outliers) is that linear regression assumes your variables are linearly independent. So in your dataset, if you have multiple observations, it would assume that the feature values of one sample have nothing to do with the values of another subject. This is often not the case. In our above, stock market example, it is often observed that one company's stock price fluctuations have a ripple effect to other companies in the same markets.

The last thing to watch out for with linear regression is that the further you extrapolate from the range of your training data, the less reliable the results of the regression become. Keep these thoughts in mind while using linear regression!