NYU Center for Data Science

DS-UA 111
Data Science for Everyone

Week 14: Lecture 1
Regression

How can we measure the difference between observed values and predicted values?

DS-UA 111
Data Science for Everyone

Week 14: Lecture 1
Regression

Adapted from Adhikari, DeNero, Wagner, Milner

Announcements

- Please check Week 14 agenda on NYU Classes
  - Homework 3/4
  - Lab 9
  - Project Milestone
- Refer to the Calendar linked to NYU Classes

Review

- Suppose we want to draw a line-chart through the scatter-plot to fit the pattern.
- The equation for a line is

Output = Intercept + Slope * Input

- The correlation helps us determine the slope of the line in standard units

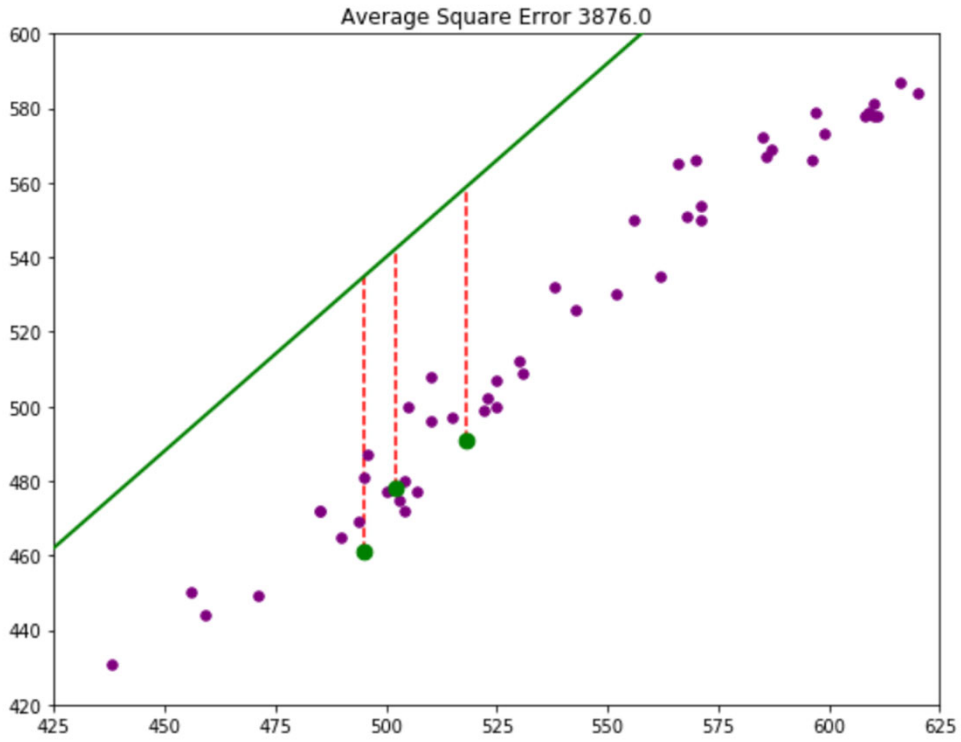
$$\frac{\text{estimate of } y \ - \ \text{average of } y}{\text{SD of } y} = r \times \frac{\text{the given } x \ - \ \text{average of } x}{\text{SD of } x}$$

estimate of *y* in standard units

x in standard units

Review

- We call the difference between observation and prediction a residual.
- In least squares regression we fit a line to the scatter-plot by minimizing the mean square error
- The mean square error is the average of the squared residuals.



Agenda

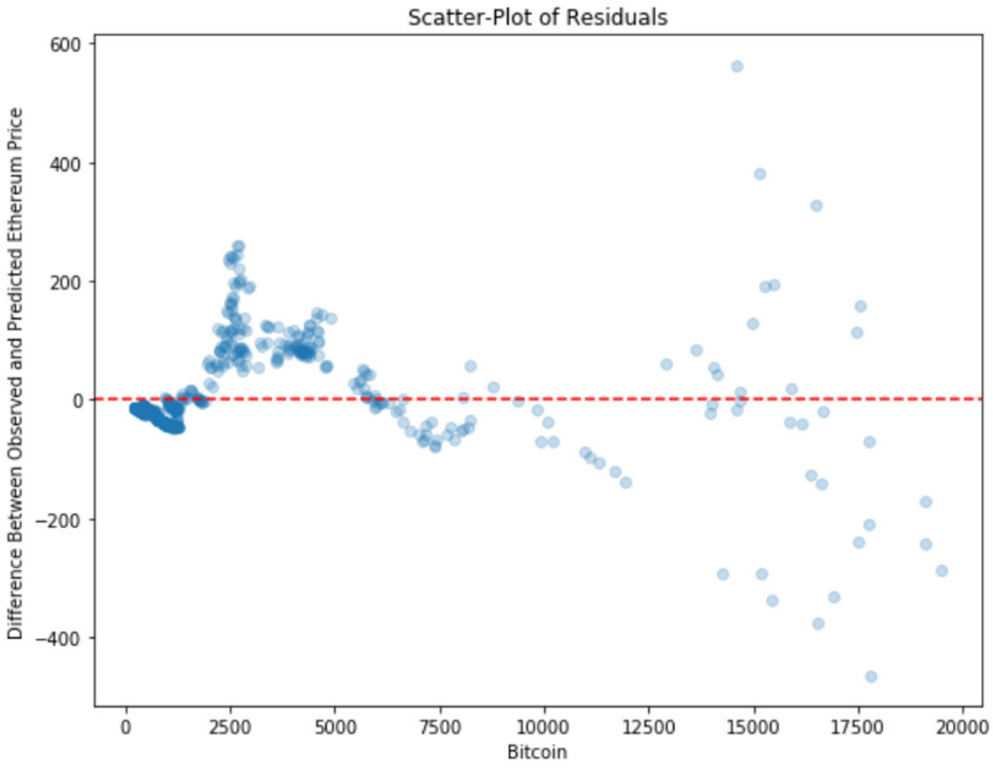
- Residuals
  - Understanding the difference between Predicted Values and Observed Values

References

- Prediction
  - Chapter 15.3-15.5

Residuals

- We should not find a association in scatter-plot of residuals.
  - If we find a positive or negative trend, then we might need to find a different explanatory variable
  - If we find a pattern like a funnel, then we might need to add an explanatory variable
- How can we have a model with two explanatory variables?



Summary

- Residuals
  - Understanding the difference between Predicted Values and Observed Values

Goals

- Understand impact of outliers on regression
- Relate mean and standard deviation of observations and predictions