Center for Data Science
NYU

DS-UA 111
Data Science for Everyone

Week 14: Lecture 2
Hypothesis Testing for Regression

How can we quantity our uncertainty about the slope and intercept determined in least squares regression?

Center for Data Science
NYU

DS-UA 111
Data Science for Everyone

Week 14: Lecture 2
Hypothesis Testing for Regression

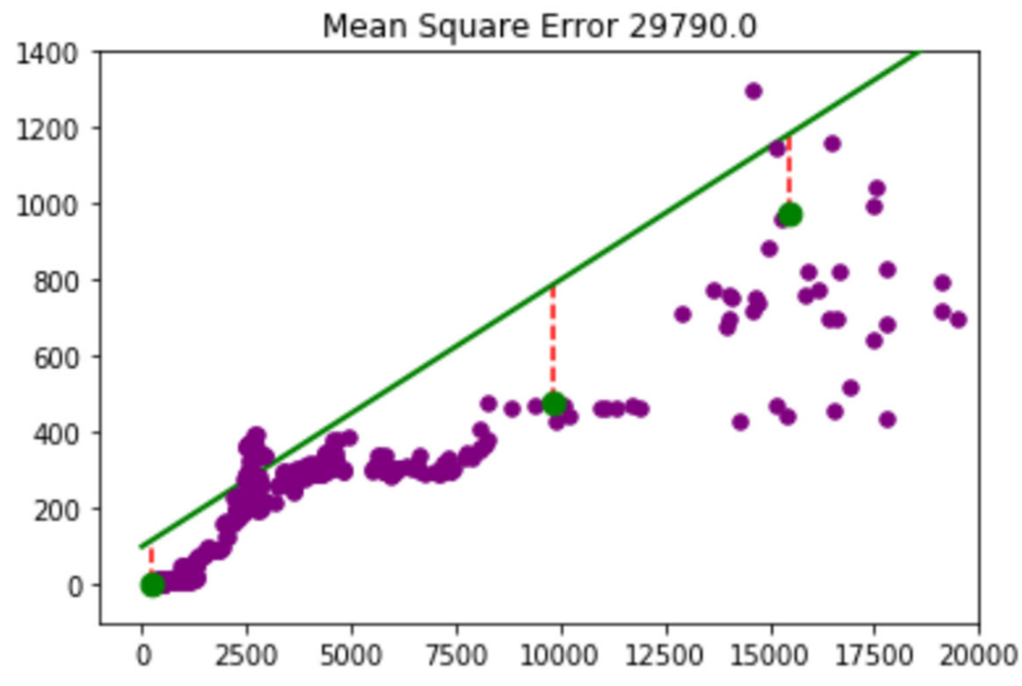
Adapted from Adhikari, DeNero, Wagner, Milner

Announcements

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

Review

- Remember that errors come from differences between predicted values and observed values. We call the errors residuals.
- For least squares regression, we compute the mean square error by
  1. squaring the residuals
  2. computing the mean

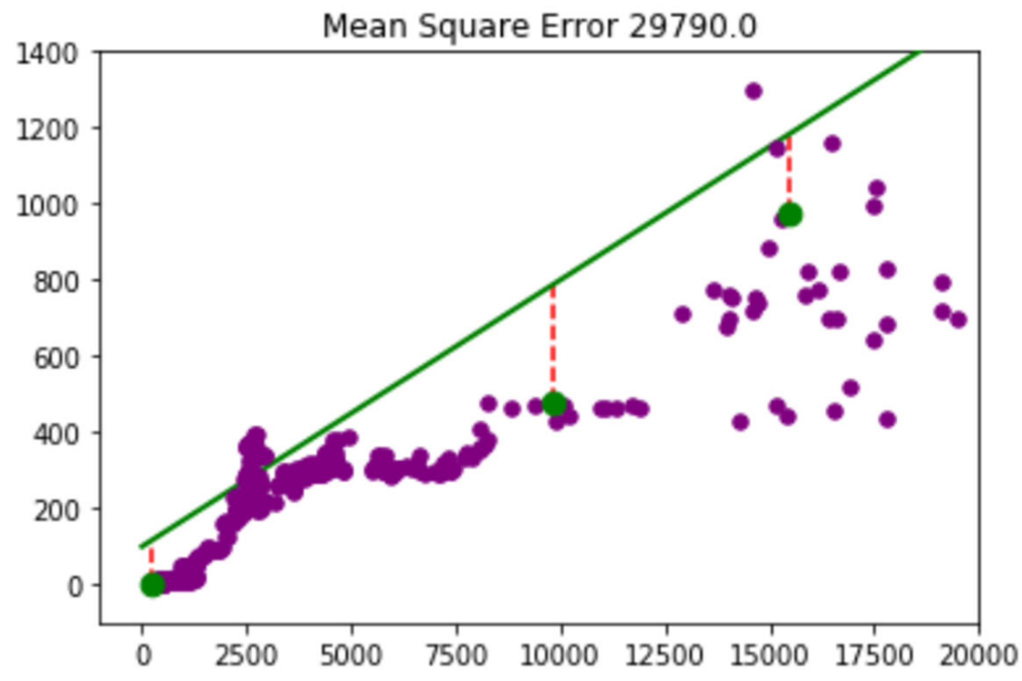


Review

- We can describe a line through a function of the form

  Output = Intercept + Slope * Input

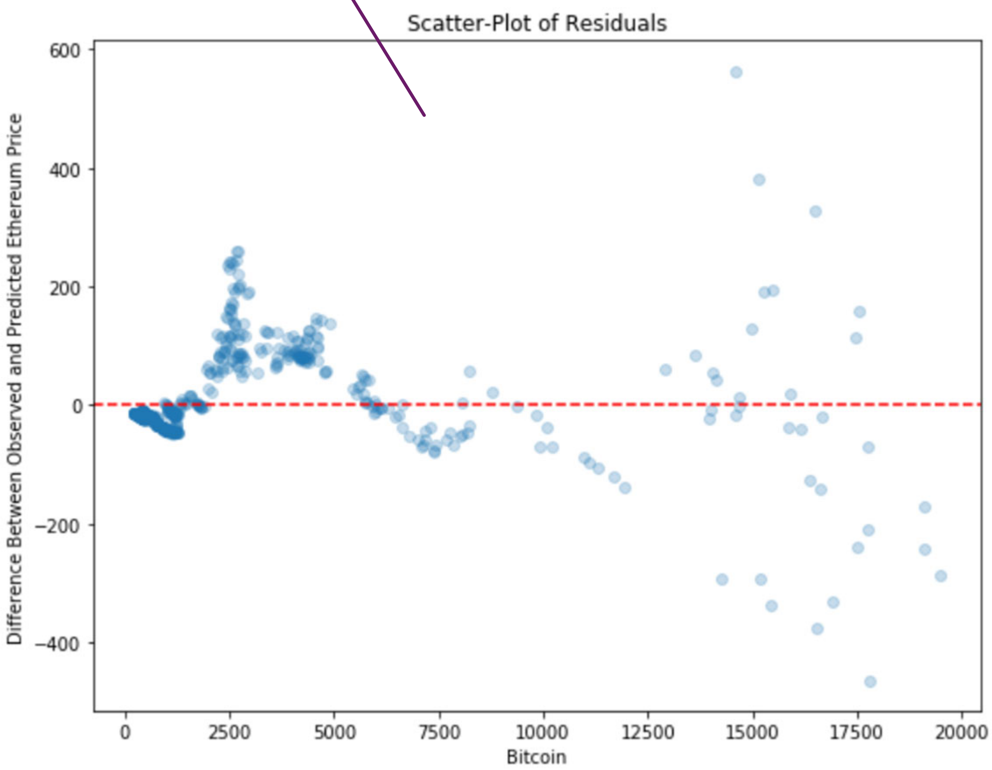
- The slope and intercept are the missing pieces in the model.
- We choose the slope and intercept that minimize the mean square error.



Review

Has a pattern like a funnel

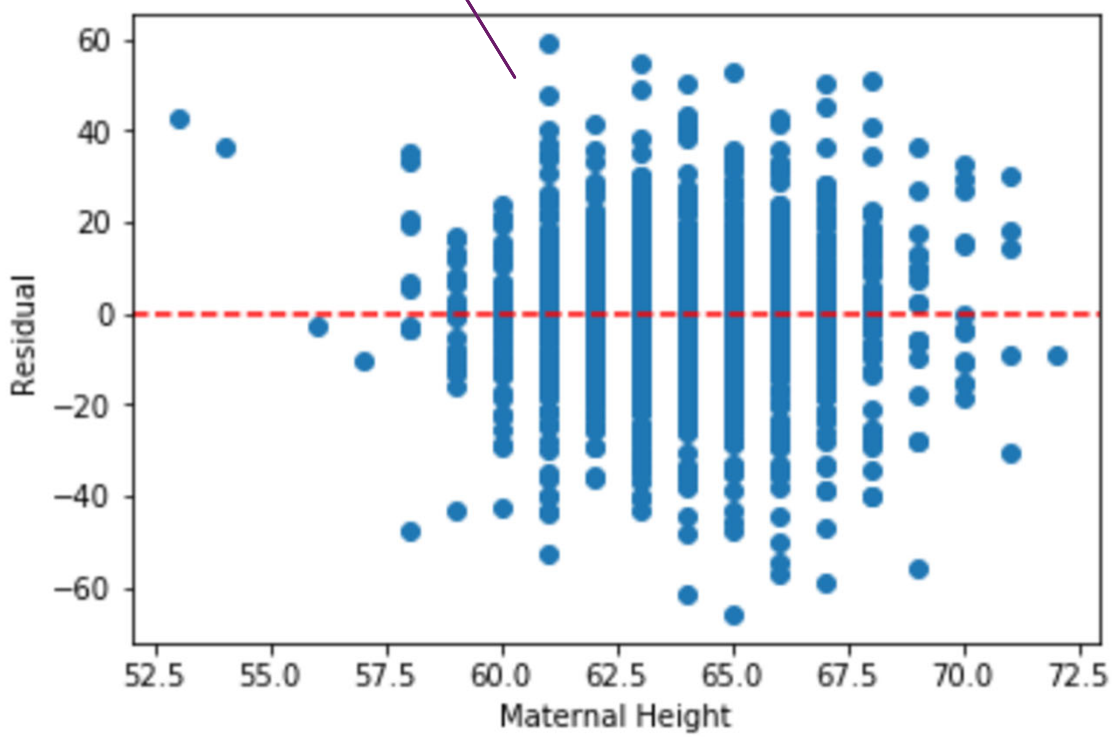
- We can generate a scatter-plot to visualize the residuals. We want
  - About half the points above 0 and about half the points below 0
  - Comparable differences from 0 throughout the points
  - No discernible trend or pattern
- Otherwise we should explore other explanatory variables



Review

- We can generate a scatter-plot to visualize the residuals. We want
  - About half the points above 0 and about half the points below 0
  - Comparable differences from 0 throughout the points
  - No discernible trend or pattern
- Otherwise we should explore other explanatory variables

Does not have a pattern



Agenda

- Multiple Explanatory Variables
- Confidence Intervals

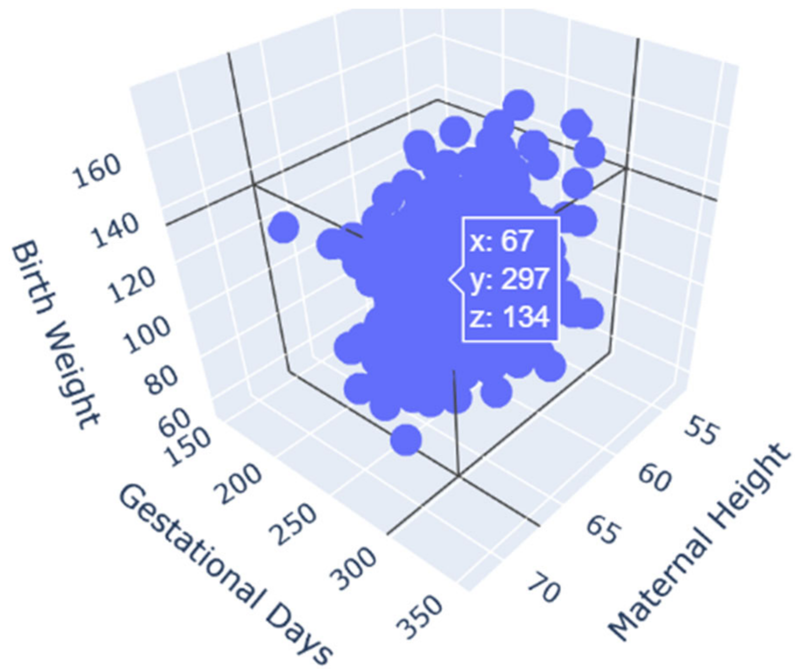
References

- Inference for Regression
  - Chapter 16.1-16.3

Explanatory Variables

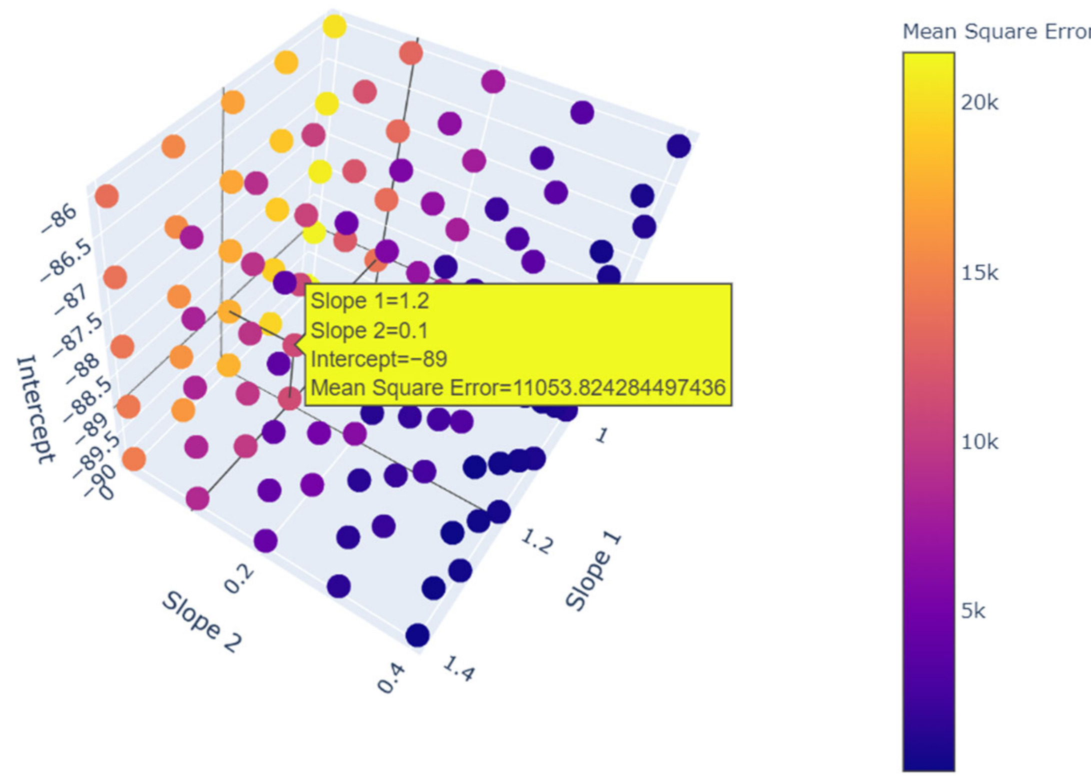
- We can have multiple explanatory variables in a least squares regression model.
- If we have two explanatory variables then the prediction determine a plane.
- We can describe a plane through a function of the form

  Output = Intercept + Slope 1 * Input 1 + Slope 2 * Input 2



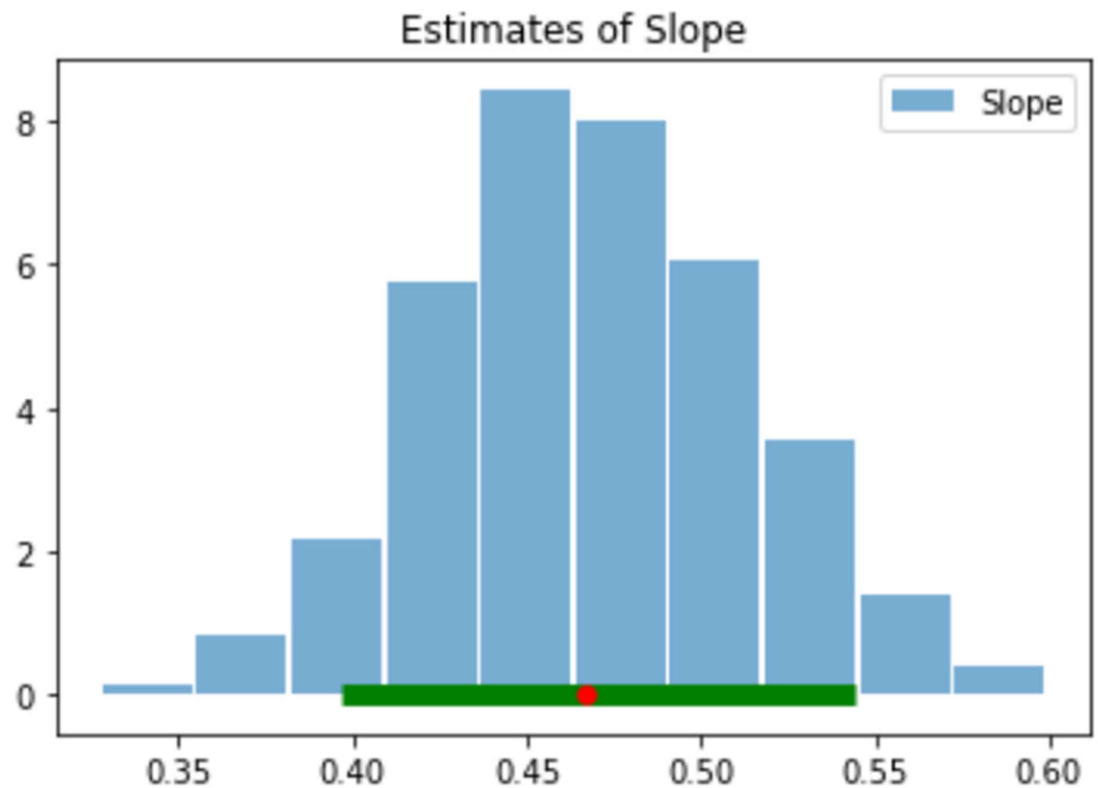
Finding a Minimum

- We choose the intercept and slopes that minimize the mean square error.
- We can use a package to find the slopes and intercept through guessing and check values



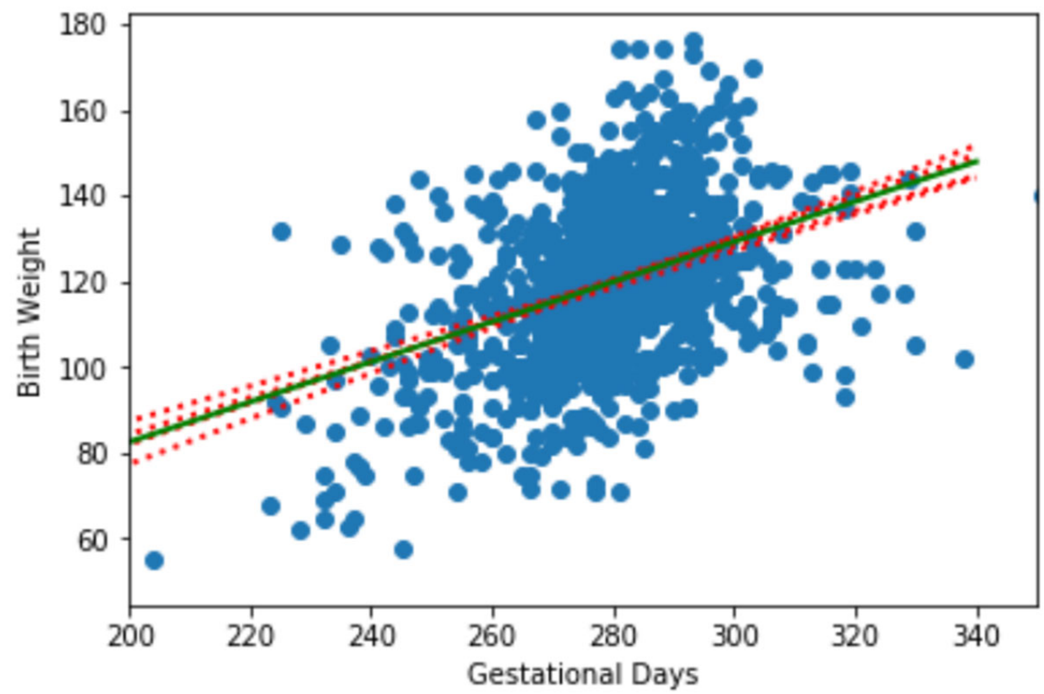
Confidence Intervals

- We determine the slope and intercept through fitting the line to the data. The data is a sample from the population.
- We can quantify the variation across samples in the slope and interval through resampling.
- Bootstrap resampling allows us to generate many slopes and intercepts across replications



Confidence Intervals

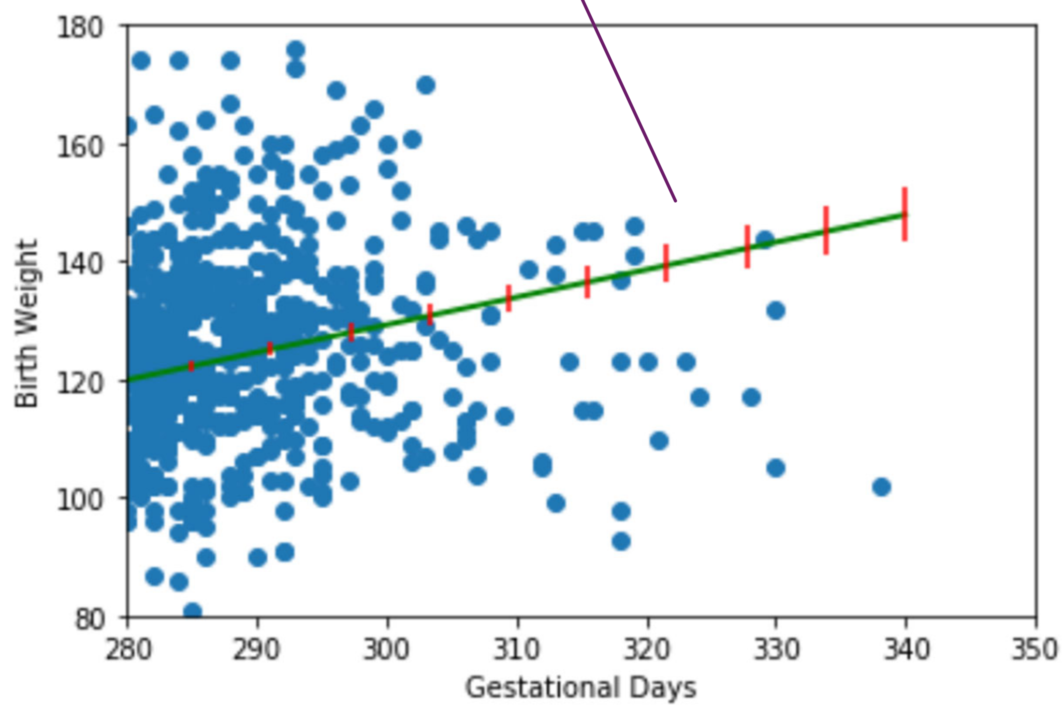
- For each replication, we have a resample. We can fit a line to the data in the resample to determine the slope and intercept.
- We can calculate confidence intervals from these numbers by determining percentiles like 5th and 95th
- Here we have bootstrap confidence intervals for slope and intercept



Confidence Intervals

Note that the confidence intervals become large for values far from the mean

- If we fix a value for the explanatory variable, then for each replication we have a slope and intercept to make a prediction.
- We can calculate confidence intervals from these numbers by determining percentiles like 5th and 95th
- Here we have bootstrap confidence intervals predictions



Summary

- Multiple Explanatory Variables
- Confidence Intervals

Goals

- Incorporate two explanatory variables into least squares regression
- Compute confidence intervals for slopes and intercepts