From Conditional Means to Linear Regression

Will Doyle

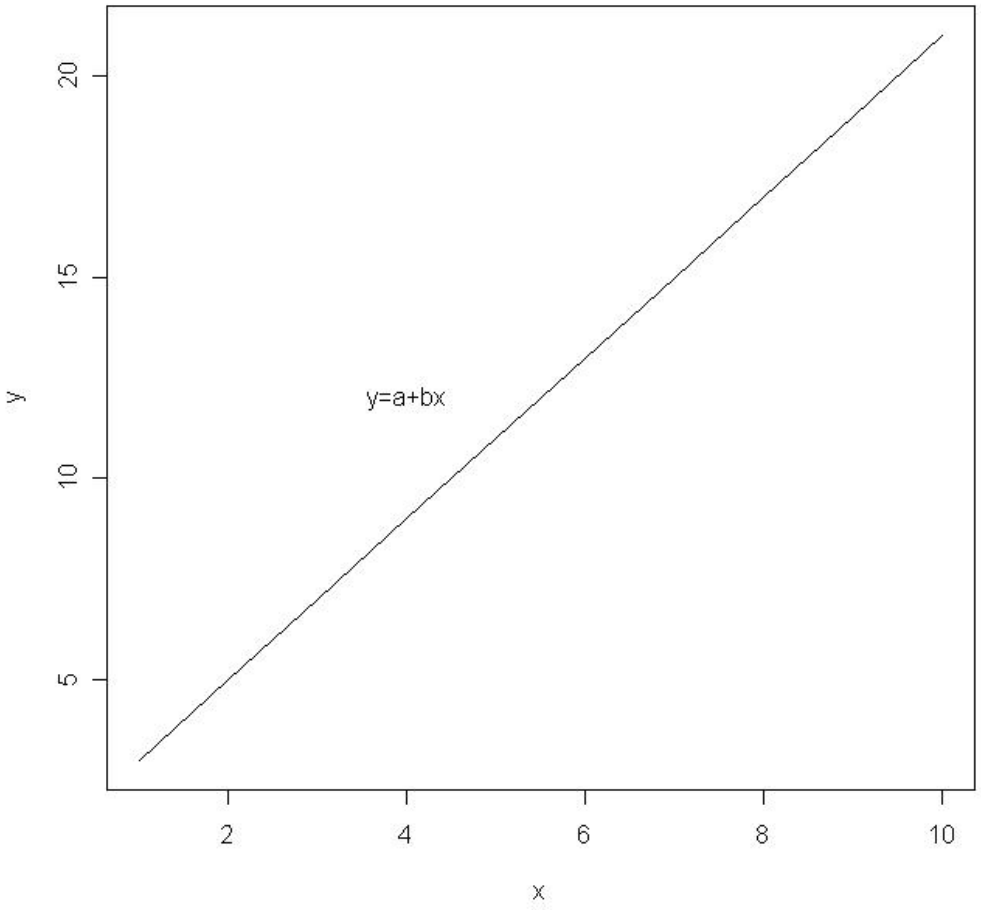
Expanding the Idea of Conditional Means

- We can generate a conditional mean for a single characteristic (e.g., average test scores for males and females).
- We can expand that to include more characteristics (e.g., average test scores for males and females from each of the four regions of the country).
- But it gets complicated fast. This is known as the curse of dimensionality.

Linear Regression

- Linear regression can be used to generate the expected value of the outcome for **every** level of multiple independent variables.
- It does this by making one simple assumption: the conditional expectation function—the expected value of the outcome given the predictor—is linear.

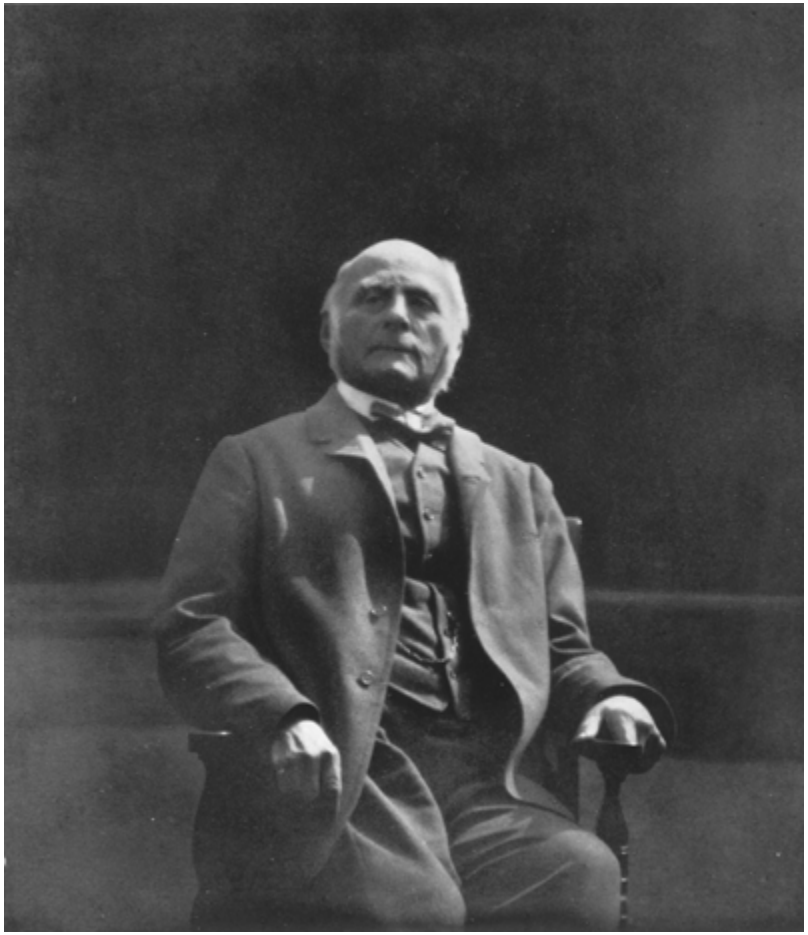
Flashback Time: Defining a Line



Regression: Line Fitting

- Regression is a means of fitting a line to an observed set of data.
- In two-dimensional space, the line is fit according to $y = a + bx$.
- In more dimensions, the same logic applies, but the line cannot be drawn simply.

Sir Francis Galton



Regression Lines

- In regression, we postulate a model of the world of the following linear (line) form:

$$y_i = \quad + \quad x_i + \quad _i$$

Where:

$y_i =$ the dependent variable for case i

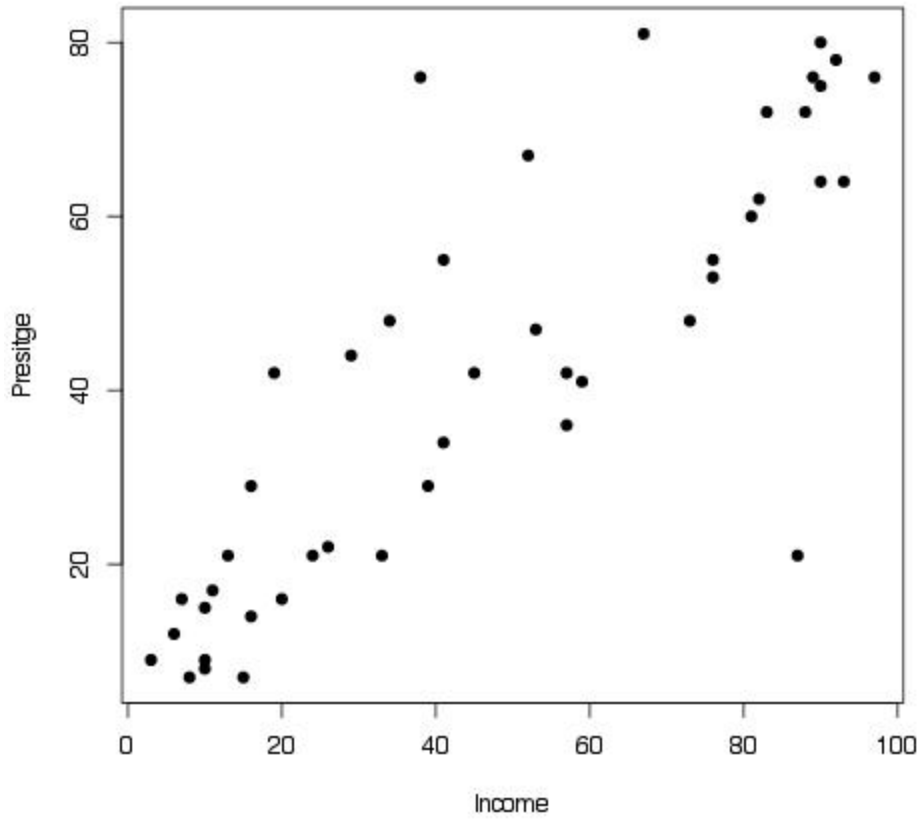
$x_i =$ the independent variable for case i

$=$ the intercept of the line describing the relationship between x and y

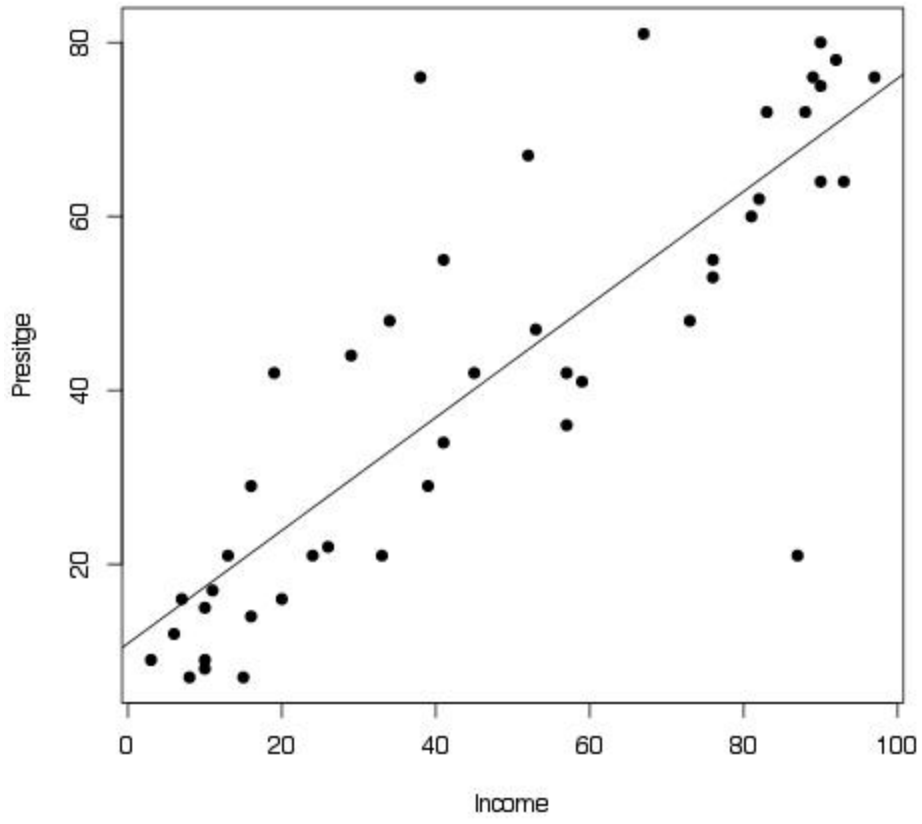
$=$ the slope of the line describing the relationship between x and y

$_i =$ an error term showing how far the fitted line is from x and y

Duncan Data on Income and Prestige



Duncan Data With Regression Line

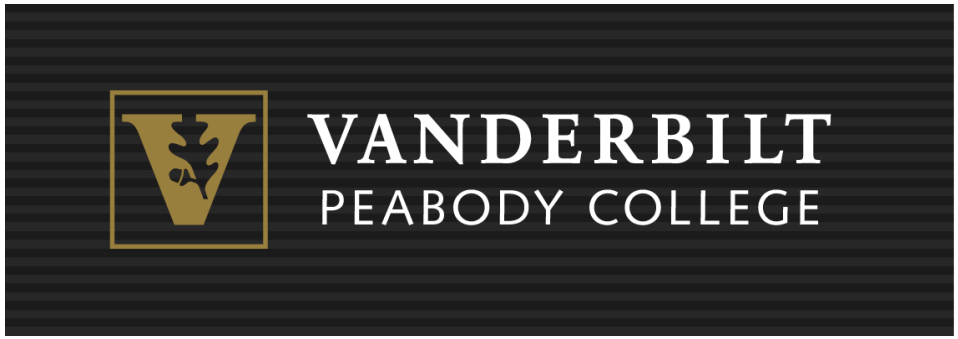


Ordinary Least Squares

- Ordinary least squares (OLS) is the most common (and usually best) form of regression when the dependent variable is continuous.
- OLS proceeds by finding a line that minimizes the sum of squared errors (SSE) between the line and the points in the data.
- OLS depends on a key set of assumptions.

Assumptions Under OLS

- *y* is continuous.
- There are fewer independent variables than there are cases in the data.
- The errors are distributed normally and have a mean of 0.
- Error terms are not correlated with one another, nor are they correlated with any of the regressors.


VANDERBILT
PEABODY COLLEGE

Training and Testing

Concepts

Will Doyle

The Problem of Prediction

- Our models can provide predictions given an observed set of characteristics.
- We can compare these to the actual data, but…
- We'll likely be overconfident, as the predictions will be based on the data at hand.

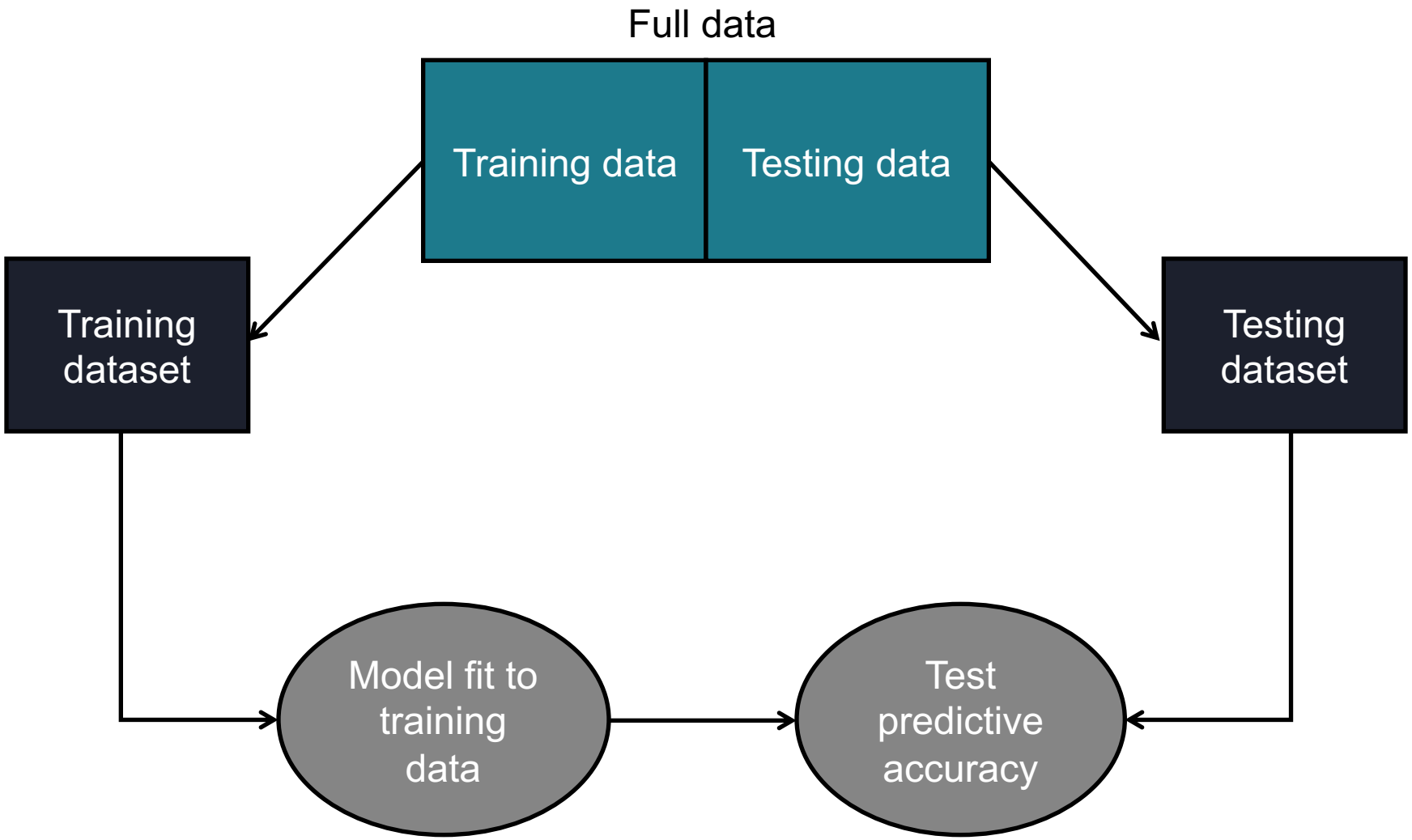
Solutions to the Problem of Prediction

- One solution (which we should do) is to make predictions prior to an event, then wait for the event to occur and see how we do.
- This can be costly and time consuming.
- The other solution is to use our predictions on out of sample data.

Training and Testing

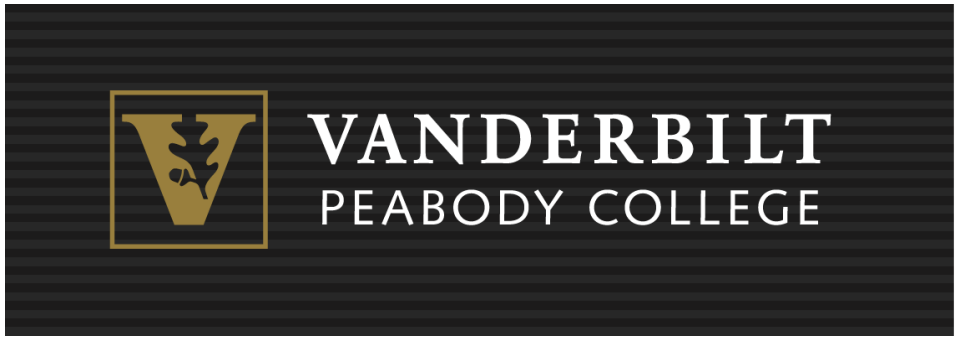
- A “training” dataset is used to create a model.
- A testing dataset is used to (you guessed it) test the predictions made by the training dataset.
- The testing dataset must **not** be used to fit the model.
- Instead, the predictions from the model are compared with the actual values of the outcome from the testing dataset.

Training and Testing



Training and Testing

- The testing dataset is also known as the validation dataset.
- We can do this once: split the data and test the model on the testing dataset. This is called validation.
- Or we can repeat the process. This is called cross validation.
- Right now, we're just going to do a single validation.
- But later on, we'll expand the scope.


VANDERBILT
PEABODY COLLEGE