Regression

APAM E4990
Modeling Social Data

Jake Hofman

Columbia University

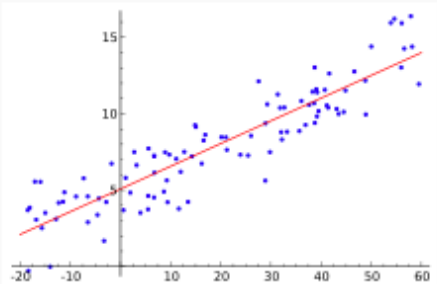
March 8, 2019

Definition

?

Definition

Regression analysis



Models

Linear regression · Simple regression · Ordinary least squares · Polynomial regression · General linear model

Generalized linear model · Discrete choice · Logistic regression · Multinomial logit · Mixed logit · Probit · Multinomial probit · Ordered logit · Ordered probit · Poisson

Multilevel model · Fixed effects · Random effects · Mixed model

Nonlinear regression · Nonparametric · Semiparametric · Robust · Quantile · Isotonic · Principal components · Least angle · Local · Segmented

Errors-in-variables

Estimation

Least squares · Ordinary least squares · Linear (math) · Partial · Total · Generalized · Weighted · Non-linear · Iteratively reweighted · Ridge regression · LASSO

Least absolute deviations · Bayesian · Bayesian multivariate

Background

Regression model validation · Mean and predicted response · Errors and residuals · Goodness of fit · Studentized residual · Gauss–Markov theorem

Definition

"The primary goal in a regression analysis is to **understand**, *as far as possible* **with the available data**, *how the conditional distribution of the* **response varies across subpopulations** *determined by the possible values of the predictor or* **predictors**."

-"Applied Regression Including Computing and Graphics"
Cook & Weisberg (1999)

Goals

Describe

Provide a compact summary of outcomes under different conditions

Predict

Make forecasts for future outcomes or unobserved conditions

Explain

Account for associations between predictors and outcomes

Goals

Describe

Provide a compact summary of outcomes under different conditions

Never "false", but may be wasteful or misleading

Predict

Make forecasts for future outcomes or unobserved conditions

Varying degrees of success, often room for improvement

Explain

Account for associations between predictors and outcomes

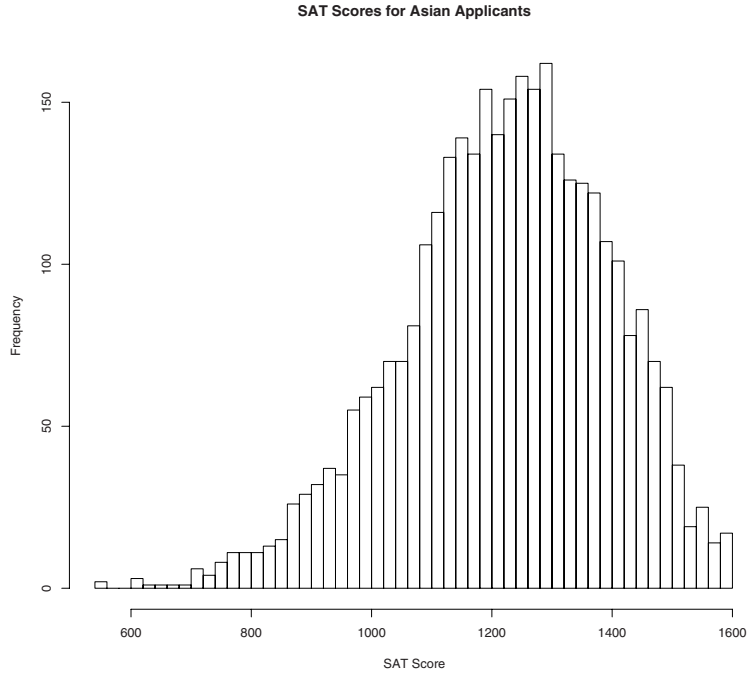
Difficult to establish causality in observational studies

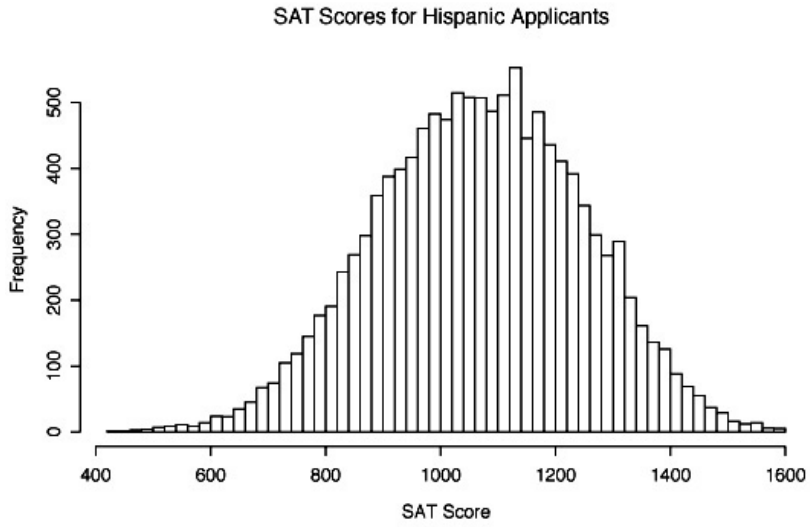
See "Regression Analysis: A Constructive Critique", Berk (2004)

Goals

Models should be flexible enough to describe observed phenomena but simple enough to generalize to future observations

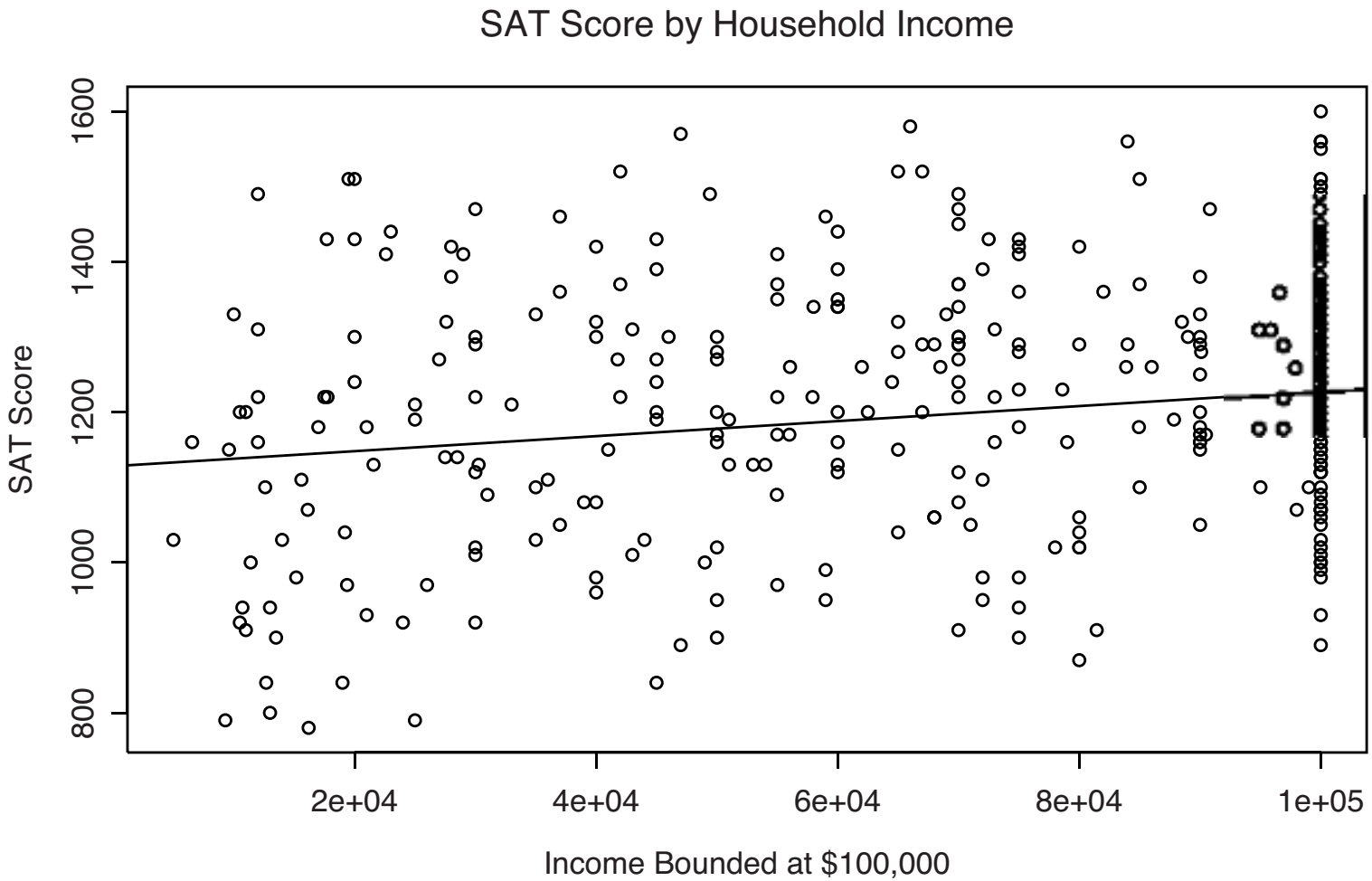
Examples[1]





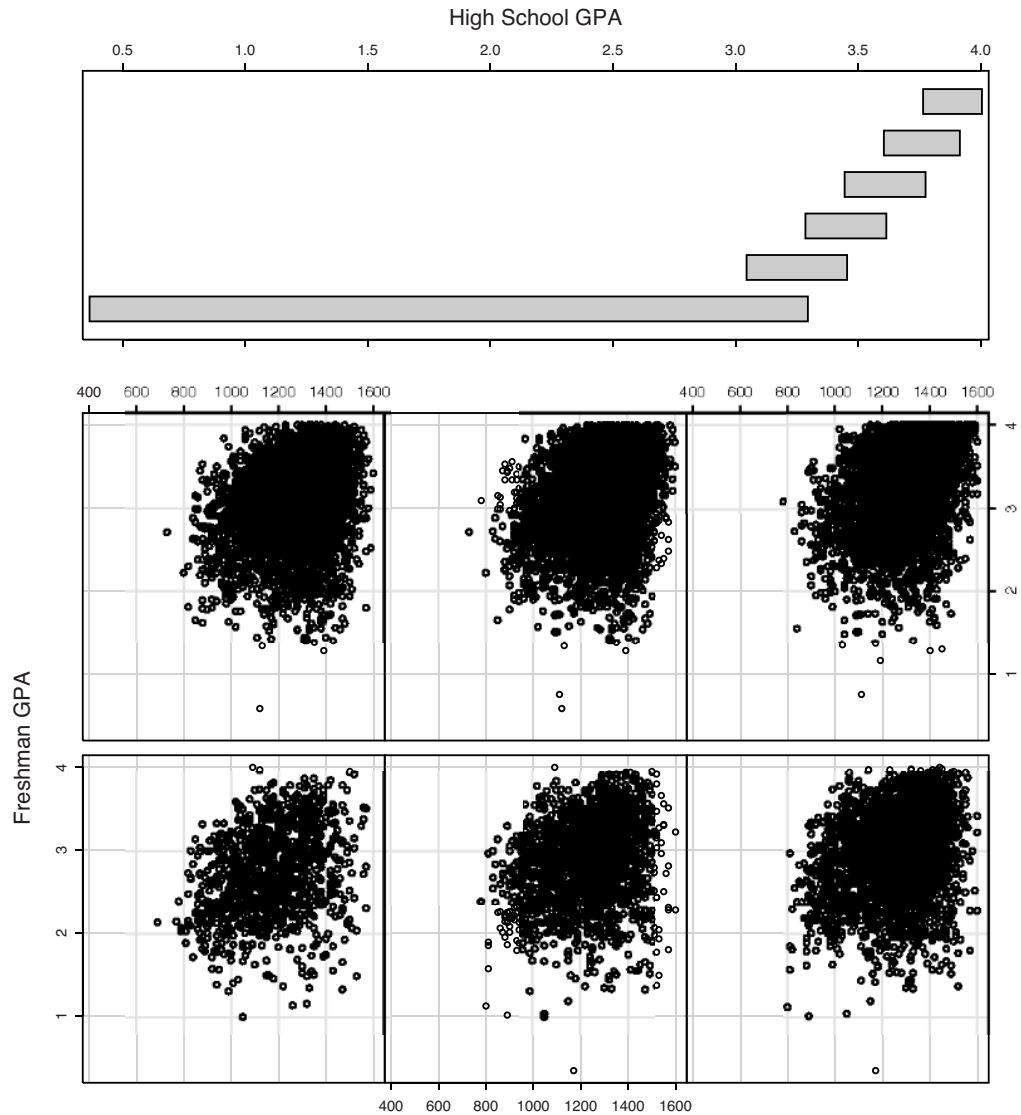
[1] "Statistical Learning from a Regression Perspective", Berk (2008)

Examples[1]



[1] "Statistical Learning from a Regression Perspective", Berk (2008)

Examples[1]



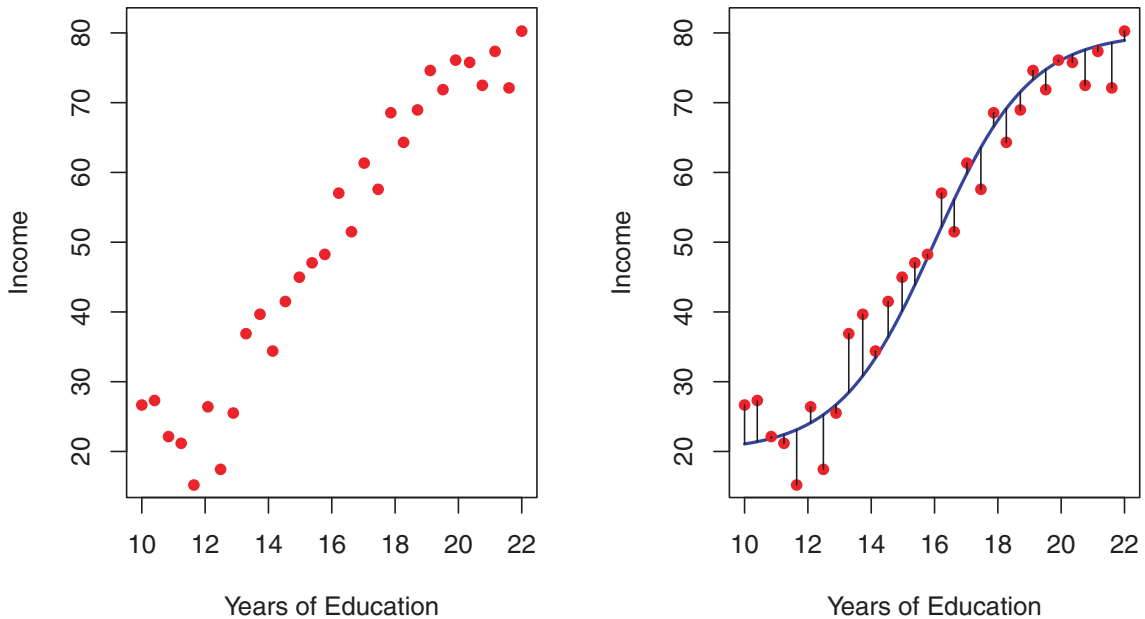
[1] "Statistical Learning from a Regression Perspective", Berk (2008)

Framework

- Specify the outcome and predictors, along with the form of the model relating them
- Define a loss function that quantifies how close a model's predictions are to observed outcomes
- Develop an algorithm to fit the model to the observations by minimizing this loss
- Assess model performance and interpret results.

Regression

Regression is an *supervised* learning task by which we aim to *predict a real-valued outcome* for an example given its features

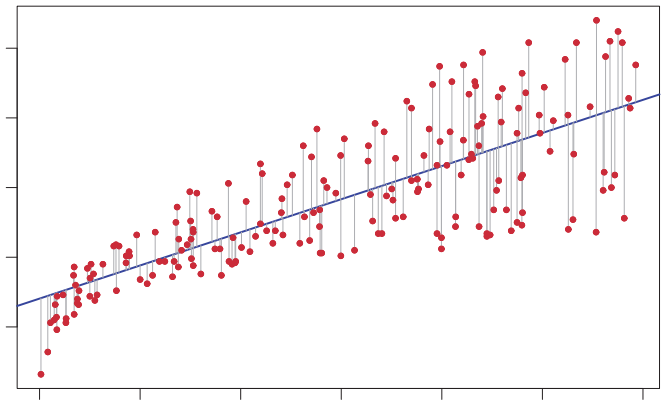


e.g., predict someone's income given their education

Linear regression

We'll use a **linear model** to make predictions $\hat{y}$ given features x:

$$\hat{y} = mx + b$$



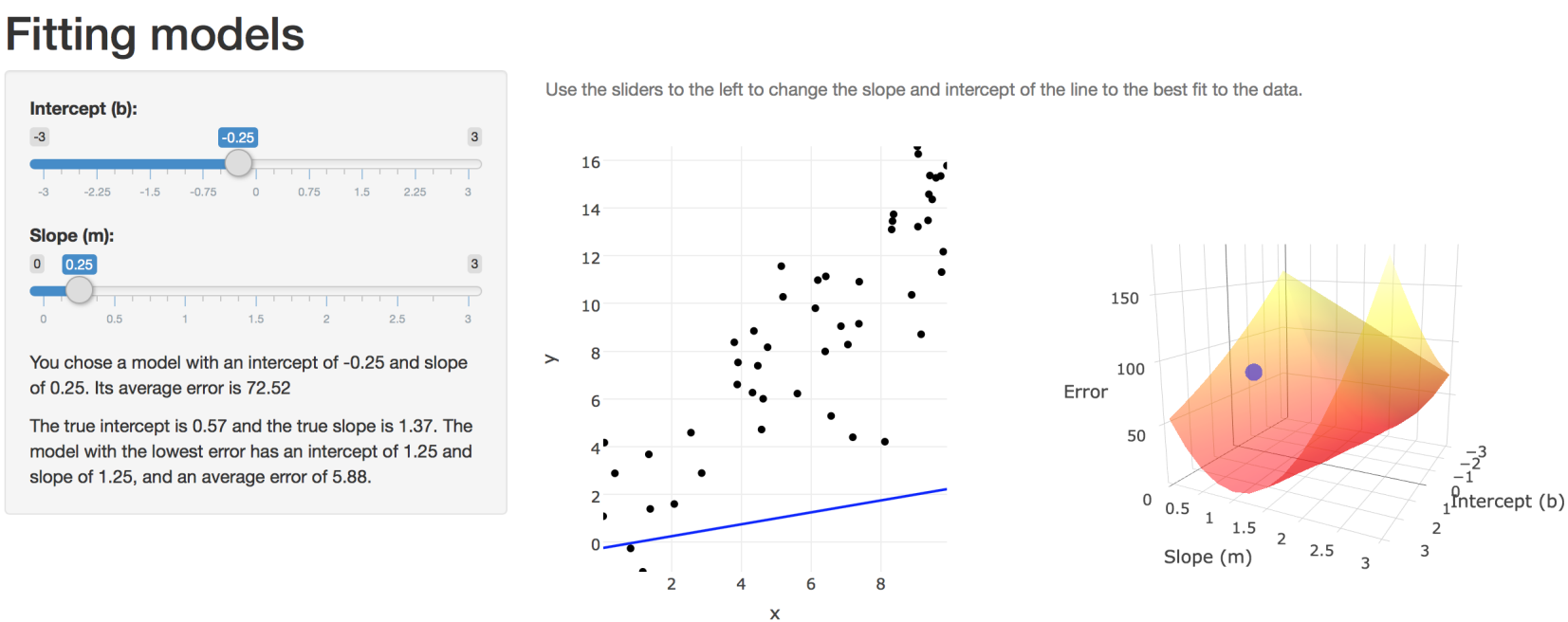
And we'll measure the **mean squared error** between the predicted and actual value of each observation:

$$\mathrm{MSE}(\mathrm{m}, \mathrm{b}) = \sum_i (\hat{y}_i - y_i)^2 = \sum_i (mx_i + b - y_i)^2$$

Linear regression

We'd like the **slope** m and **intercept** b with the **lowest error**:

$$(\hat{b}, \hat{m}) = \arg\min_{(b,m)} \mathrm{MSE}(\mathrm{m}, \mathrm{b}) = \sum_i (mx_i + b - y_i)^2$$

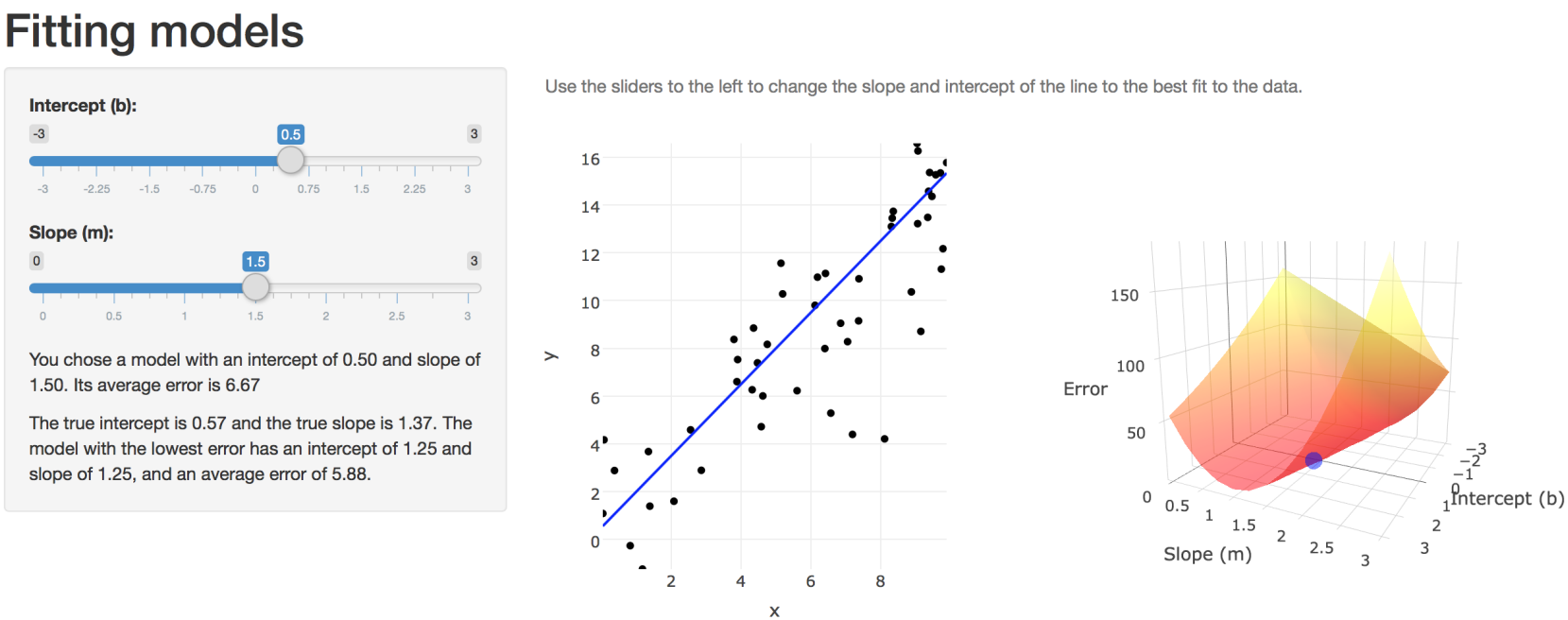


With just two parameters, we can manually search for the best fit

Linear regression

We'd like the **slope** m and **intercept** b with the **lowest error**:

$$(\hat{b}, \hat{m}) = \arg\min_{(b,m)} \mathrm{MSE}(\mathrm{m}, \mathrm{b}) = \sum_i (mx_i + b - y_i)^2$$



With just two parameters, we can manually search for the best fit

Linear regression

We'd like the **slope** m and **intercept** b with the **lowest error**:

$$(\hat{b}, \hat{m}) = \arg\min_{(b,m)} \mathrm{MSE}(\mathrm{m}, \mathrm{b}) = \sum_i (mx_i + b - y_i)^2$$

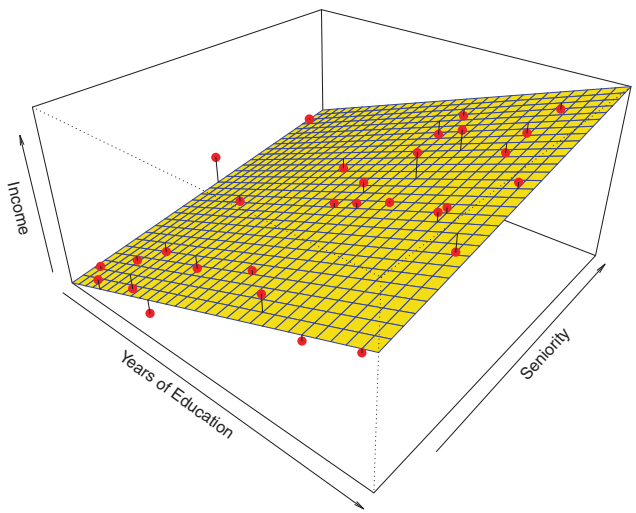
But notice the above is quadratic in m and b, so we can solve for the exact minimum:

$$\begin{aligned} \hat{m} &= \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ \hat{b} &= \bar{y} - \hat{m}\bar{x} \end{aligned}$$

Multiple linear regression

We can extend this to making predictions $\hat{y}$ from multiple features $x_1, x_2, \ldots$:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K = \beta \cdot x$$



e.g., predict income given education and time at company

Multiple linear regression

We can extend this to making predictions $\hat{y}$ from multiple features $x_1, x_2, \ldots$:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K = \beta \cdot x$$

And there's still a closed form solution:

$$\hat{\beta} = \arg\min_{\beta} \sum_i (\beta \cdot x_i - y_i)^2 = (X^T X)^{-1} X^T y$$

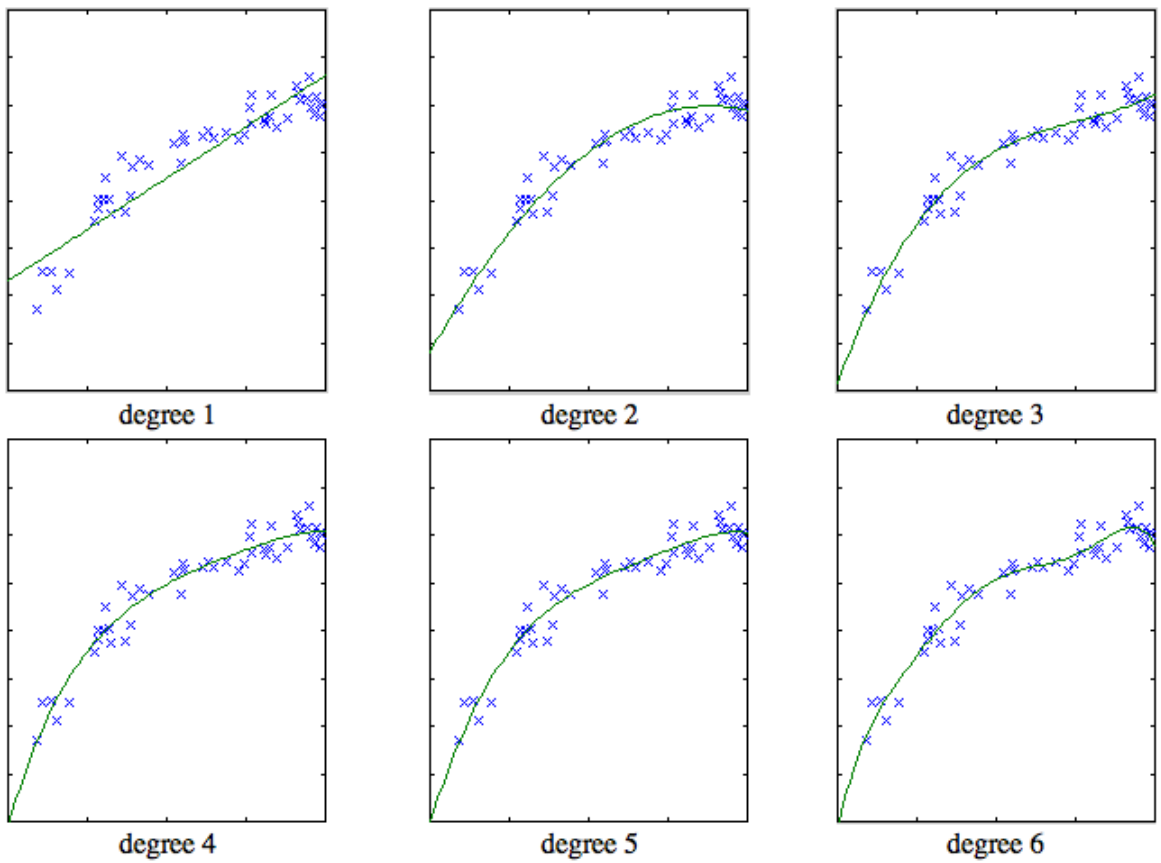
Although it quickly becomes computationally expensive with many features

Multiple linear regression

We can even use **non-linear features**, for instance to fit a polynomial:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_K x^K$$

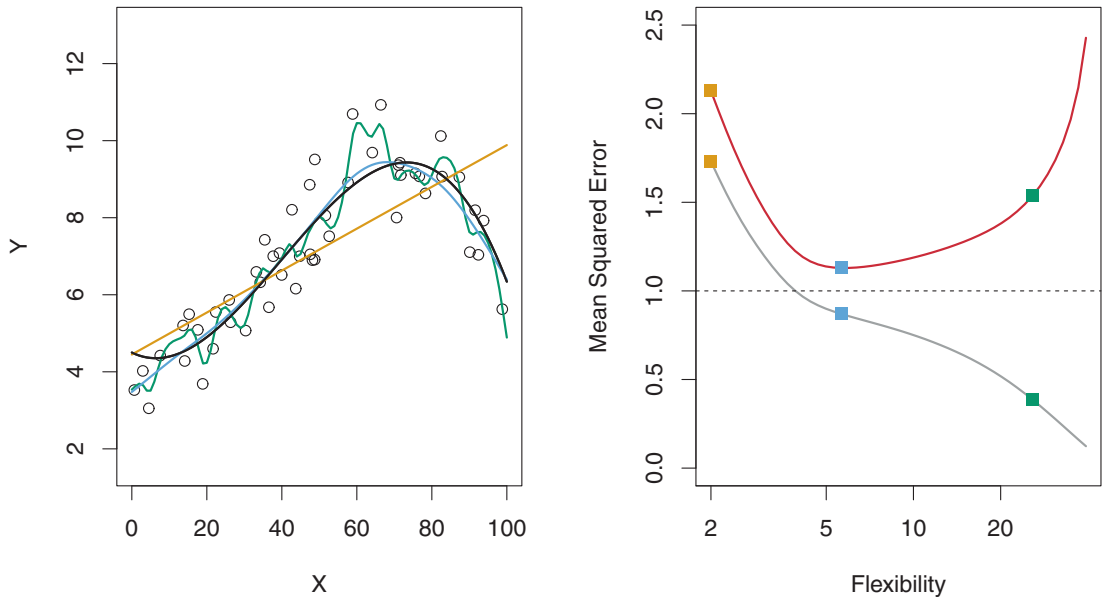
So **what degree polynomial** should we pick?



Cross-validation

- Randomly split our data into three sets
- For each polynomial degree k:
  - Fit a model to the training set
  - Evaluate on the validation set
- Select the model with the lowest validation error
- Quote final performance of this model on the test set

Cross-validation



- Randomly split our data into three sets
- For each polynomial degree k:
  - Fit a model to the training set
  - Evaluate on the validation set
- Select the model with the lowest validation error
- Quote final performance of this model on the test set

Bigger models $\neq$ Better models

Regression

Our models should be complex enough to explain the past, but simple enough to generalize to the future

Regression

Simple methods (e.g., linear models) work surprisingly well, especially with lots of data