MS&E 125: Intro to Applied Statistics

Linear regression

Professor Udell

Management Science and Engineering
Stanford

March 23, 2023

Outline

Motivation: linear models

Linear models can be used for

- **prediction:** given a set of input variables, predict a value for the output variable
- **understanding:** how are the input variables related to the output variable, and to each other?
- **inference:** how much do the input variables affect the output variable?
- **counterfactuals:** what would happen if we changed the input variables?
- **control:** how can we change the input variables to achieve a desired output?

Outline

Regression setup

we want to predict output given inputs

- input variables $x\mathbf{R}^p$
  - also called "predictors", "independent variables", "covariates"
  - a row of a data table
- output variable $y \in \mathbf{R}$
  - also called "outcome", "response", "dependent variable", "label", "target" ...

Regression setup

we want to predict output given inputs

- input variables $x\mathbf{R}^p$
  - also called "predictors", "independent variables", "covariates"
  - a row of a data table
- output variable $y \in \mathbf{R}$
  - also called "outcome", "response", "dependent variable", "label", "target" ...

example: to predict the cost of an insurance claim,

- y is the cost of an insurance claim.
- entries of x are the properties of the insured and his/her vehicle, e.g., credit score, age of the vehicle, ...

Demo: simple linear regression

https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/regression.ipynb

Simple linear regression

simple linear regression: $p = 1$

- predict
$$\hat{y} = \beta_0 + \beta_1 x$$
- $\beta_0, \beta_1 \in \mathbf{R}$ are called **regression coefficients**
- $\hat{y}$ is called the **prediction** for input x

Predictions: example

In the fathers and sons dataset, we found

$$\hat{y} = 34 + 0.5x$$

where x is the height of the father in inches.

Predictions: example

In the fathers and sons dataset, we found

$$\hat{y} = 34 + 0.5x$$

where x is the height of the father in inches.

Q: What do the numbers 34 and .5 mean?

Predictions: example

In the fathers and sons dataset, we found

$$\hat{y} = 34 + 0.5x$$

where x is the height of the father in inches.

Q: What do the numbers 34 and .5 mean?
A: A father with height 0 inches has a son with height 34 inches. For each inch of height, the son is expected to be 0.5 inches taller.

Outline

Residuals

look at **residual** r to understand how well the model fits the data

$$r = y - \hat{y} = y - \beta_0 - \beta_1 x_1$$

pick β so the residuals are small

Dataset

to find the best line, we need a dataset! suppose we have

- n data points $(x^{(1)}, y_1), \ldots, (x^{(n)}, y_1)$
  - also called **dataset**, **examples**, **observations**, **samples** or **measurements**
- each $x_i \in \mathbf{R}^p$ is a vector of p input variables
  - a

  row from the data table
- each $y_i \in \mathbf{R}$ is a scalar output variable

Linear regression: two perspectives

how to choose β?

- **optimization perspective:** find β to minimize the sum of squared errors

$$\text{minimize} \qquad \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **statistical perspective:** find the line that maximizes the likelihood of the data

theorem: for appropriate assumptions, the two perspectives give the same answer
(coming in a few slides, or see All of Statistics ch. 14)

Least squares fitting

Least squares fitting

Q: Suppose $x_i = 0$ for every i. What is β_0?

Least squares fitting

Q: Suppose $x_i = 0$ for every i. What is β_0?
A: Set derivative to zero; solution is the average of the y_is.

Least squares fitting

Q: Suppose $x_i = 0$ for every i. What is β_0?
A: Set derivative to zero; solution is the average of the y_is.
Q: Given $x_i \in \mathbf{R}$, what is β_1?

Least squares fitting

Q: Suppose $x_i = 0$ for every i. What is β_0?
A: Set derivative to zero; solution is the average of the y_is.
Q: Given $x_i \in \mathbf{R}$, what is β_1?
A: Set derivative to zero; solution is the slope of the line of best fit.

Solve for β_0

$$\text{minimize} \qquad \sum i = 1^n (y_i - \beta_0 - \beta_1 x_i)^2$$

take derivative wrt β_0 and set to zero:

$$\begin{aligned}
\sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) &= 0 \\
\sum_{i=1}^{n} y_i &= \beta_0 n - \beta_1 \sum_{i=1}^{n} x_i \\
\frac{1}{n} \sum_{i=1}^{n} y_i &= \beta_0 - \beta_1 \frac{1}{n} \sum_{i=1}^{n} x_i
\end{aligned}$$

$\implies$ the model goes through the point of averages

Solve for β_1

$$\text{minimize} \qquad \sum i = 1^n (y_i - \beta_0 - \beta_1 x_i)^2$$

take derivative wrt β_1 and set to zero:

$$\begin{aligned} \sum_{i=1}^n -2(y_i - \beta_0 - \beta_1 x_i)x_i &= 0 \\ \sum_{i=1}^n x_i y_i &= \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 \end{aligned}$$

$$\beta_1 = \frac{\sum_{i=1}^n x_i y_i - \beta_0 \sum_i = 1^n x_i}{\sum_{i=1}^n x_i^2}$$

interpretation:

- suppose x and y have been standardized so that $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i = 0$ and $\frac{1}{n}\sum_{i=1}^n x_i^2 = \frac{1}{n}\sum_{i=1}^n y_i^2 = 1$.
- then $\beta_1 = \frac{1}{n}\sum_{i=1}^n x_i y_i$ is the **correlation** between x and y

Outline

Linear regression model

probabilistic model for linear regression: suppose the xs are fixed, and ys are generated by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i is a random variable with mean 0 and variance σ^2.

Linear regression model

probabilistic model for linear regression: suppose the xs are fixed, and ys are generated by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i is a random variable with mean 0 and variance σ^2.

under the model, the likelihood of observing residual $r = y - \hat{y}$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{r^2}{2\sigma^2}\right)$$

Demo: are errors iid normal?

https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/regression.ipynb

Maximum likelihood

the **likelihood function** gives the probability of the data given the parameters:

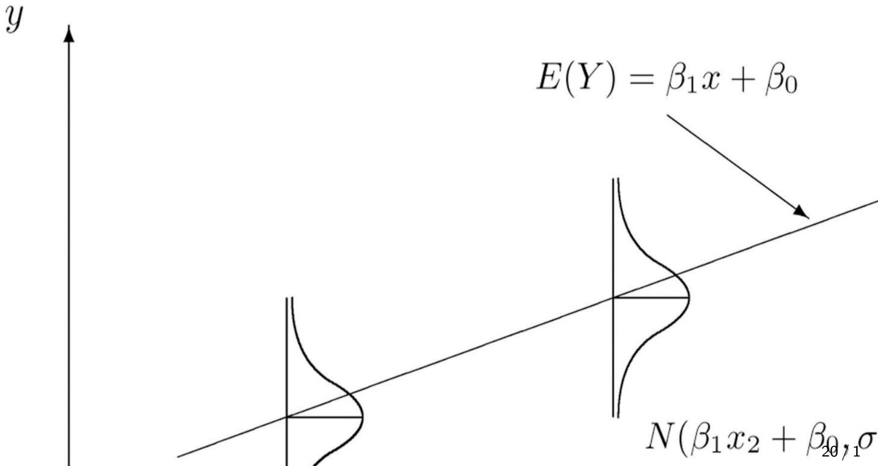
$$\ell(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)$$

maximum likelihood chooses β_0 and β_1 to maximize the likelihood function:

$$\begin{aligned}
\hat{\beta}_0, \hat{\beta}_1 &= \underset{\beta_0, \beta_1}{\operatorname{argmax}}\, \ell(\beta_0, \beta_1) \\
&= \underset{\beta_0, \beta_1}{\operatorname{argmax}}\, \log \ell(\beta_0, \beta_1) \\
&= \underset{\beta_0, \beta_1}{\operatorname{argmax}} \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right) \\
&= \underset{\beta_0, \beta_1}{\operatorname{argmax}} -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\
&= \operatorname{argmin} \frac{1}{\;} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2
\end{aligned}$$

Probabilistic interpretation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$



Estimation puts a hat on it

statisticians use hats to denote estimates:

- $\hat{\beta}_0$ is the estimate of β_0
- $\hat{y}$ is the estimate of y

these estimates are **random quantities** that depend on the data

`https://colab.research.google.com/github/stanford-mse-125/demos/blob/main/regression-uncertainty.ipynb`

Properties of the estimator

putting it together, we have found:

$$\hat{\beta}_1 = \rho(x, y)\hat{\sigma}_y / \hat{\sigma}_x, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

- $\rho(x, y)$ is the correlation between x and y
- $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the sample standard deviations of x and y
- $\bar{x}$ and $\bar{y}$ are the sample means of x and y

under the normal model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

these estimates are unbiased:

$$\mathbb{E}[\hat{\beta}_1] = \beta_1, \qquad \mathbb{E}[\hat{\beta}_0] = \beta_0$$

see All of Statistics ch 14 for formulas for the variance of the estimates

Outline

Matrix notation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

rewrite using linear algebra:

- form **response vector** $y \in \mathbf{R}^n$: each outcome y_i is an entry of y
    - a

  lso called **target vector**
- form **design matrix** $X \in \mathbf{R}^{n \times p}$: each example $x^{(i)}$ is a row of X
    - also called **feature matrix**
    - if the model includes a constant term, the 0th column of $X \in \mathbf{R}^{n \times p+1}$ is all ones

$$\begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Least squares in matrix notation

rewrite error:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \|y - X\beta\|^2$$

interpretation:

- $X\beta$ is a linear combination of the columns of X
- we seek the linear combination that best matches y