CSCE 633: Machine Learning

Lecture 4: Linear Regression

Texas A&M University

9-2-19

Before we begin

- HW 1 will be posted on ecampus next week **START EARLY**
- Projects! Lot's of questions
- Can I come up with my own project?
- Do I need to work in teams?
- I don't have a project idea, how do I find a team?
- More formal details to come

Goals of this lecture

- Simple Linear Regression
- Multiple Linear Regression
- Convexity

Predicting Quantitative Response

- $D = \{(X_i, y_i)\}_{i=1}^{n}$
- y_i can be
  - Categorical $y_i \in \{1, 2, \cdots, C\}$
  - Binary $y_i \in \{0, 1\}$
  - $y_i \in \mathbb{R}$
- There are many algorithms that predict quantitative response
- Many are generalizations of Linear Regression

Before we begin: Notation

- n vs. N
- p vs. D
- $\mathbf{w}$ vs. $\boldsymbol{\beta}$
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$

An Important Example: Advertising

- How do I make a useful Market Plan for the coming fiscal year to increase sales?
- My budget includes advertising in TV
- advertising in radio
- advertising in newspapers

An Important Example: Advertising

- How do I make a useful Market Plan for the coming fiscal year to increase sales?
- My budget includes advertising in TV
- advertising in radio
- advertising in newspapers
- How much should I add or subtract from each to increase sales?

Important Questions to Ask

- Is there a relationship between budget and sales?
- If there is a relationship, how strong is it?
- Which of the three media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Supervised Learning: Regression

- input $\mathbf{x}$: advertising media budgets (TV, Radio, Newspaper)
- output y: sales
- model parameters $\mathbf{w}$
- Deterministic (parametric) linear model

$$y = f(\mathbf{x}|\mathbf{w}) = \mathbf{w}^T\mathbf{x}$$

- Deterministic non-linear model

$$y = f(\mathbf{x}|\mathbf{w}) = \mathbf{w}^T\phi(\mathbf{x})$$

- Non-Deterministic (probabilistic) non-linear model

$$y = f(\mathbf{x}|\mathbf{w}) = \mathbf{w}^T\phi(\mathbf{x}) + \epsilon,\ \epsilon \sim N(\mu, \sigma^2)$$

Simple Linear Regression

We want to predict Y based upon a single predictor X

Simple Linear Regression

We want to predict Y based upon a single predictor X, we want to regress Y on to X:

$$Y \approx \beta_0 + \beta_1 X$$

Simple Linear Regression

We want to predict Y based upon a single predictor X, we want to regress Y on to X:

$$Y \approx \beta_0 + \beta_1 X$$

$$Sales \approx \beta_0 + \beta_1 TV$$

Parameters

We want to learn (trained by existing data) the parameters of the model, also known as the coefficients, β

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$

Estimating the Coefficients

- We do not know β_0 or β_1
- So, assume we have a training set $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$
- Assume $n = 200$ markets of sales and tv budget.
- Goal: set $\hat{\beta}_0$ and $\hat{\beta}_1$ so we are as close to y_i from x_i for all i

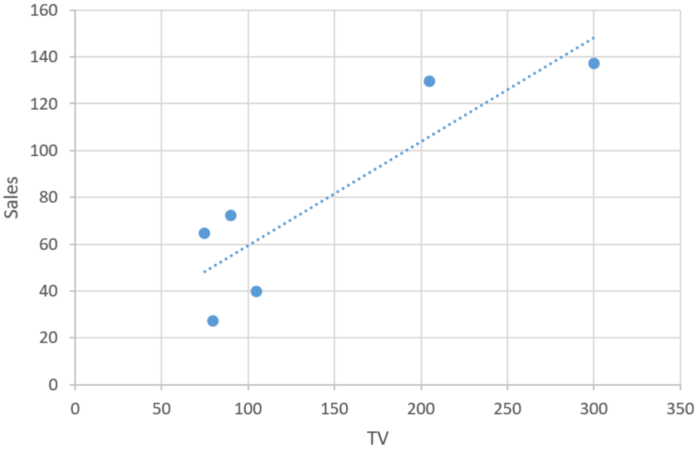
Residual

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for y based on the ith value of x
- Then the residual error is

$$e_i = y_i - \hat{y}_i$$

So we can define total error as $\sum_{i=1}^{n} e_i$ and want to fit a model to minimize this error

Sum of Residuals



Least Squares

The residual sum of squares

$$\begin{aligned} RSS &= e_1^2 + e_2^2 + \cdots + e_n^2 \\ &= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \end{aligned}$$

Least Squares: Learning Coefficients

The residual sum of squares

$$\begin{aligned} RSS &= e_1^2 + e_2^2 + \cdots + e_n^2 \\ &= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \end{aligned}$$

if RSS is our total sum of squared error, what do we need to learn?

Differentiation

To minimize RSS, need to differentiate with respect to both unknowns

$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

- Calculate $\frac{\partial RSS}{\partial \hat{\beta}_0}$
- Calculate $\frac{\partial RSS}{\partial \hat{\beta}_1}$

Differentiation: $\hat{\beta}_0$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $\frac{\partial RSS}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1)$

Differentiation: $\hat{\beta}_0$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $\frac{\partial RSS}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1)$
- $= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$, where $e_i = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$

Differentiation: $\hat{\beta}_0$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $\frac{\partial RSS}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1)$
- $= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$, where $e_i = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$
- $= -2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} \hat{\beta}_0 + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i$

Differentiation: $\hat{\beta}_0$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $\frac{\partial RSS}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1)$
- $= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$, where $e_i = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$
- $= -2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} \hat{\beta}_0 + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i$
- Note: $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean

Differentiation: $\hat{\beta}_0$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $\frac{\partial RSS}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1)$
- $= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$, where $e_i = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$
- $= -2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} \hat{\beta}_0 + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i$
- $= -2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\overline{x}$

Differentiation: $\hat{\beta}_0$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $\frac{\partial RSS}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1)$
- $= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$, where $e_i = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$
- $= -2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} \hat{\beta}_0 + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i$
- $= -2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\overline{x}$
- To minimize, set $\frac{\partial RSS}{\partial \hat{\beta}_0} = 0$

Differentiation: $\hat{\beta}_0$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $= -2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\overline{x}$
- To minimize, set $\frac{\partial RSS}{\partial \hat{\beta}_0} = 0$
- $-2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\overline{x} = 0$

Differentiation: $\hat{\beta}_0$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $= -2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\overline{x}$
- To minimize, set $\frac{\partial RSS}{\partial \hat{\beta}_0} = 0$
- $-2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\overline{x} = 0$
- $2n\hat{\beta}_0 = 2n\overline{y} - 2n\hat{\beta}_1\overline{x}$

Differentiation: $\hat{\beta}_0$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $= -2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\overline{x}$
- To minimize, set $\frac{\partial RSS}{\partial \hat{\beta}_0} = 0$
- $-2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\overline{x} = 0$
- $2n\hat{\beta}_0 = 2n\overline{y} - 2n\hat{\beta}_1\overline{x}$
- $\cancel{2n}\hat{\beta}_0 = \cancel{2n}\overline{y} - \cancel{2n}\hat{\beta}_1\overline{x}$

Differentiation: $\hat{\beta}_0$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $= -2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\overline{x}$
- To minimize, set $\frac{\partial RSS}{\partial \hat{\beta}_0} = 0$
- $-2n\overline{y} + 2n\hat{\beta}_0 + 2n\hat{\beta}_1\overline{x} = 0$
- $2n\hat{\beta}_0 = 2n\overline{y} - 2n\hat{\beta}_1\overline{x}$
- $\cancel{2n}\hat{\beta}_0 = \cancel{2n}\overline{y} - \cancel{2n}\hat{\beta}_1\overline{x}$
- $\boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}}$

Differentiation: $\hat{\beta}_1$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $\frac{\partial RSS}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i)$

Differentiation: $\hat{\beta}_1$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $\frac{\partial RSS}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i)$
- $= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i)$

Differentiation: $\hat{\beta}_1$

- $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $\frac{\partial RSS}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i)$
- $= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i)$
- $= -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}_0 \sum_{i=1}^{n} x_i + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$

Differentiation: $\hat{\beta}_1$

- $= -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}_0 \sum_{i=1}^{n} x_i + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $= -\cancel{2}\sum_{i=1}^{n} y_i x_i + \cancel{2}\hat{\beta}_0 \sum_{i=1}^{n} x_i + \cancel{2}\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$

Differentiation: $\hat{\beta}_1$

- $= -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}_0 \sum_{i=1}^{n} x_i + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $= -\cancel{2}\sum_{i=1}^{n} y_i x_i + \cancel{2}\hat{\beta}_0 \sum_{i=1}^{n} x_i + \cancel{2}\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $-\sum_{i=1}^{n} y_i x_i + (\overline{y} - \hat{\beta}_1 \overline{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$

Differentiation: $\hat{\beta}_1$

- $= -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}_0 \sum_{i=1}^{n} x_i + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $= -\cancel{2}\sum_{i=1}^{n} y_i x_i + \cancel{2}\hat{\beta}_0 \sum_{i=1}^{n} x_i + \cancel{2}\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $-\sum_{i=1}^{n} y_i x_i + (\overline{y} - \hat{\beta}_1 \overline{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $-\sum_{i=1}^{n} y_i x_i + \overline{y} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \overline{x} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$

Differentiation: $\hat{\beta}_1$

- $= -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}_0 \sum_{i=1}^{n} x_i + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $= -\cancel{2}\sum_{i=1}^{n} y_i x_i + \cancel{2}\hat{\beta}_0 \sum_{i=1}^{n} x_i + \cancel{2}\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $-\sum_{i=1}^{n} y_i x_i + (\overline{y} - \hat{\beta}_1 \overline{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $-\sum_{i=1}^{n} y_i x_i + \overline{y} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \overline{x} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $\overline{y} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i = \hat{\beta}_1 \overline{x} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$

Differentiation: $\hat{\beta}_1$

- $= -2\sum_{i=1}^{n} y_i x_i + 2\hat{\beta}_0 \sum_{i=1}^{n} x_i + 2\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $= -\cancel{2}\sum_{i=1}^{n} y_i x_i + \cancel{2}\hat{\beta}_0 \sum_{i=1}^{n} x_i + \cancel{2}\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $-\sum_{i=1}^{n} y_i x_i + (\overline{y} - \hat{\beta}_1 \overline{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $-\sum_{i=1}^{n} y_i x_i + \overline{y} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \overline{x} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$
- $\overline{y} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i = \hat{\beta}_1 \overline{x} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$
- $\overline{y} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i = \hat{\beta}_1 (\overline{x} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2)$

Differentiation: $\hat{\beta}_1$

- $\hat{\beta}_1 = \frac{\overline{y}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i}{\overline{x}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2}$

Differentiation: $\hat{\beta}_1$

- $\hat{\beta}_1 = \frac{\overline{y}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i}{\overline{x}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2}$
- $\hat{\beta}_1 = \frac{\overline{yx} n - \sum_{i=1}^{n} y_i x_i}{\overline{x}^2 n - \sum_{i=1}^{n} x_i^2}$

Differentiation: $\hat{\beta}_1$

- $\hat{\beta}_1 = \frac{\overline{y}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i}{\overline{x}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2}$
- $\hat{\beta}_1 = \frac{\overline{y}\overline{x}n - \sum_{i=1}^{n} y_i x_i}{\overline{x}^2 n - \sum_{i=1}^{n} x_i^2}$
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \overline{y}\overline{x}n}{\sum_{i=1}^{n} x_i^2 - \overline{x}^2 n}$

Differentiation: $\hat{\beta}_1$: Numerator

- $\sum_{i=1}^{n} x_i y_i - \overline{y}\overline{x}n$

Differentiation: $\hat{\beta}_1$: Numerator

- $\sum_{i=1}^{n} x_i y_i - \overline{yx}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{yx}n + \overline{yx}n - \overline{yx}n$

Differentiation: $\hat{\beta}_1$: Numerator

- $\sum_{i=1}^{n} x_i y_i - \overline{yx}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{yx}n - \overline{yx}n + \overline{yx}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + \overline{yx}n$

Differentiation: $\hat{\beta}_1$: Numerator

- $\sum_{i=1}^{n} x_i y_i - \overline{yx}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{yx}n - \overline{yx}n + \overline{yx}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + \overline{yx}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + \overline{yx}\sum_{i=1}^{n} 1$

Differentiation: $\hat{\beta}_1$: Numerator

- $\sum_{i=1}^{n} x_i y_i - \overline{yx} n$
- $\sum_{i=1}^{n} x_i y_i - \overline{yx} n - \overline{yx} n + \overline{yx} n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i - \overline{x} \sum_{i=1}^{n} y_i + \overline{yx} n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i - \overline{x} \sum_{i=1}^{n} y_i + \overline{yx} \sum_{i=1}^{n} 1$
- $\sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i - \overline{x} \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \overline{yx}$

Differentiation: $\hat{\beta}_1$: Numerator

- $\sum_{i=1}^{n} x_i y_i - \overline{yx}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{yx}n - \overline{yx}n + \overline{yx}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + \overline{yx}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + \overline{yx}\sum_{i=1}^{n} 1$
- $\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \overline{yx}$
- $\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\overline{y} - \sum_{i=1}^{n} \overline{x}y_i + \sum_{i=1}^{n} \overline{xy}$

Differentiation: $\hat{\beta}_1$: Numerator

- $\sum_{i=1}^{n} x_i y_i - \overline{y}\overline{x}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y}\overline{x}n - \overline{y}\overline{x}n + \overline{y}\overline{x}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + \overline{y}\overline{x}n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + \overline{y}\overline{x}\sum_{i=1}^{n} 1$
- $\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \overline{y}\overline{x}$
- $\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\overline{y} - \sum_{i=1}^{n} \overline{x}y_i + \sum_{i=1}^{n} \overline{x}\overline{y}$
- $\sum_{i=1}^{n}(x_i y_i - x_i\overline{y} - \overline{x}y_i + \overline{x}\overline{y})$

Differentiation: $\hat{\beta}_1$: Numerator

- $\sum_{i=1}^{n} x_i y_i - \overline{yx} n$
- $\sum_{i=1}^{n} x_i y_i - \overline{yx} n - \overline{yx} n + \overline{yx} n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i - \overline{x} \sum_{i=1}^{n} y_i + \overline{yx} n$
- $\sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i - \overline{x} \sum_{i=1}^{n} y_i + \overline{yx} \sum_{i=1}^{n} 1$
- $\sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i - \overline{x} \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \overline{yx}$
- $\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \overline{y} - \sum_{i=1}^{n} \overline{x} y_i + \sum_{i=1}^{n} \overline{xy}$
- $\sum_{i=1}^{n} (x_i y_i - x_i \overline{y} - \overline{x} y_i + \overline{xy})$
- $\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$

Differentiation: $\hat{\beta}_1$

- $\hat{\beta}_1 = \frac{\overline{y}\sum_{i=1}^n x_i - \sum_{i=1}^n y_i x_i}{\overline{x}\sum_{i=1}^n x_i - \sum_{i=1}^n x_i^2}$
- $\hat{\beta}_1 = \frac{\overline{yx} n - \sum_{i=1}^n y_i x_i}{\overline{x}^2 n - \sum_{i=1}^n x_i^2}$
- $\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \overline{yx} n}{\sum_{i=1}^n x_i^2 - \overline{x}^2 n}$
- $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n x_i^2 - \overline{x}^2 n}$

Differentiation: $\hat{\beta}_1$: Denominator

Denominator for $\frac{\partial RSS}{\partial \hat{\beta}_1} = 0$

$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 - n\bar{x}^2 + n\bar{x}^2$

$= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$

$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + \bar{x}^2 \sum_{i=1}^{n} 1$

$= \sum_{i=1}^{n} x_i^2 - 2\bar{x} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2$

$= \sum_{i=1}^{n} \left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right)$

$= \sum_{i=1}^{n} \left(x_i^2 - \bar{x}x_i - \bar{x}x_i + \bar{x}^2\right)$

$= \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})$

$= \sum_{i=1}^{n} (x_i - \bar{x})^2$

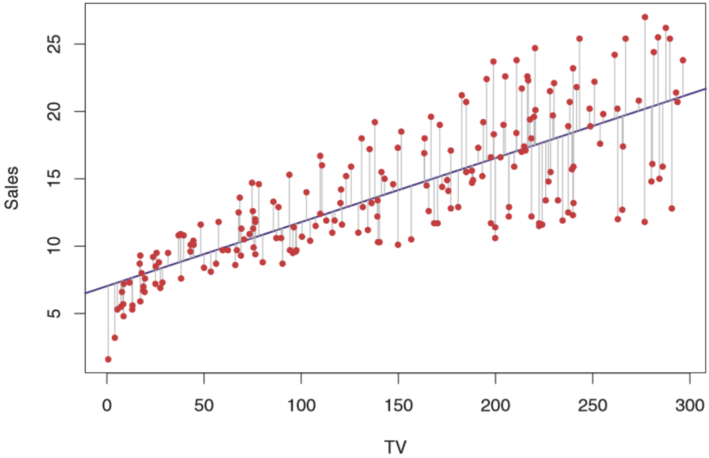
Differentiation: $\hat{\beta}_1$

- $\hat{\beta}_1 = \frac{\overline{y}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i}{\overline{x}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2}$
- $\hat{\beta}_1 = \frac{\overline{y}\overline{x}n - \sum_{i=1}^{n} y_i x_i}{\overline{x}^2 n - \sum_{i=1}^{n} x_i^2}$
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \overline{y}\overline{x}n}{\sum_{i=1}^{n} x_i^2 - \overline{x}^2 n}$
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} x_i^2 - \overline{x}^2 n}$
- $\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$

Optimal Coefficents: $\hat{\beta}_0$, $\hat{\beta}_1$

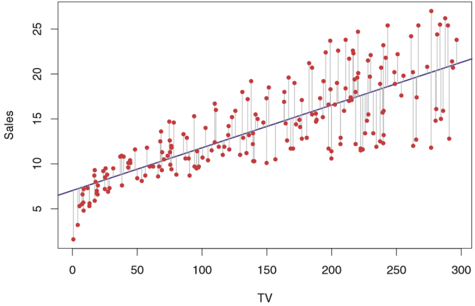
- $\boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}}$
- $\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$

Advertising Solution



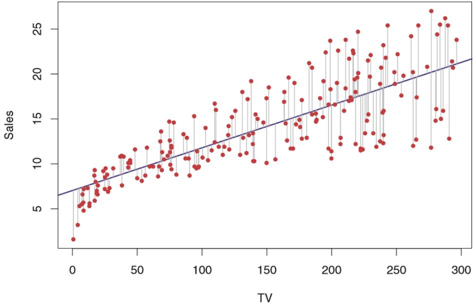
-
- $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.0475$
- Source: ISLR

Advertising Solution



-
- $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.0475$, if we had no TV advertising, how many units would we sell? What if we had 1000 budgeted for TV?
- A) 703, 475
- B) 7.03, 47.5
- C) 47.5, 7.03
- D) 475, 703

Advertising Solution



-
- $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.0475$, if we had no TV advertising, how many units would we sell? What if we had 1000 budgeted for TV?
- A) 703, 475
- B) 7.03, 47.5
- C) 47.5, 7.03
- D) 475, 703

Accuracy of Coefficient Estimates

- Remember, the true relationship is $Y = f(X) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$
- So, $Y = \beta_0 + \beta_1 X + \epsilon$

Accuracy of Coefficient Estimates

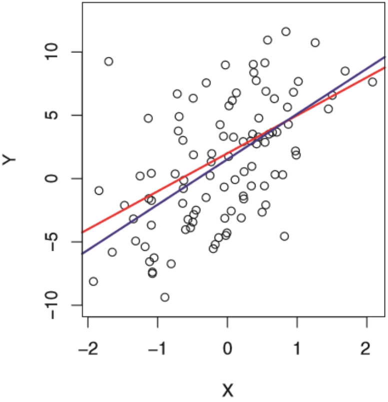
- Remember, the true relationship is $Y = f(X) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$
- So, $Y = \beta_0 + \beta_1 X + \epsilon$
- This is the *population regression line* which is the best linear approximation to the true relationship between X and Y

Accuracy of Coefficient Estimates

- Remember, the true relationship is $Y = f(X) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$
- So, $Y = \beta_0 + \beta_1 X + \epsilon$
- This is the *population regression line* which is the best linear approximation to the true relationship between X and Y
- Assume, for example $Y = 2 + 3X + \epsilon$ and you sample this population with 100 random variables X to generate 100 Y

Accuracy of Coefficient Estimates

- Assume, for example $Y = 2 + 3X + \epsilon$ and you sample this population with 100 random variables X to generate 100 Y



-

Accuracy of Coefficient Estimates

- Assume, for example $Y = 2 + 3X + \epsilon$ and you sample this population with 100 random variables X to generate 100 Y - repeating the process
- $\hat{\mu} = \overline{y}$ - sample mean from observations recorded is close with lots of sampling. Same $\hat{\beta}_0$ and $\hat{\beta}_1$ - is a good estimate with enough data.
- linear regression versus estimation of the mean of a random variable leads to concept of bias

Accuracy of Coefficient Estimates

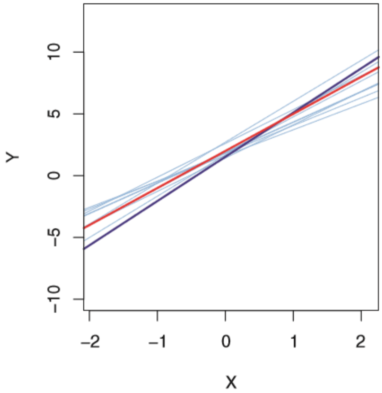
- Assume, for example $Y = 2 + 3X + \epsilon$ and you sample this population with 100 random variables X to generate 100 Y - repeating the process
- $\hat{\mu} = \overline{y}$ - sample mean from observations recorded is close with lots of sampling. Same $\hat{\beta}_0$ and $\hat{\beta}_1$ - is a good estimate with enough data.
- linear regression versus estimation of the mean of a random variable leads to concept of bias
- If we use the sample mean $\hat{\mu}$ to estimate true μ, this is unbiased since, on average, we expect them to e the same.
  - one set of $y_1, y_2, \cdots, y_n$ might result in $\hat{\mu}$ that underestimates μ
  - Another that overestimates μ
  - etc.

Accuracy of Coefficient Estimates

- Same with $\hat{\beta}_0$ and $\hat{\beta}_1$ - average enough samples and enough regressions to get to true β_0 and β_1

Accuracy of Coefficient Estimates

- Assume, for example $Y = 2 + 3X + \epsilon$ and you sample this population with 100 random variables X to generate 100 Y - repeating the process



-

Accuracy of Coefficient Estimates

- Same with $\hat{\beta}_0$ and $\hat{\beta}_1$ - average enough samples and enough regressions to get to true β_0 and β_1
- So we ask, how accurate is the sample mean $\hat{\mu}$ from the estimate of μ - how far off is a single estimate?

Accuracy of Coefficient Estimates

- Same with $\hat{\beta}_0$ and $\hat{\beta}_1$ - average enough samples and enough regressions to get to true β_0 and β_1
- So we ask, how accurate is the sample mean $\hat{\mu}$ from the estimate of μ - how far off is a single estimate?
- We need to calculate the standard error of $\hat{\mu}$, $SE(\hat{\mu})$

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

- Where σ^2 is the standard deviation of each of the realizations of y_i of Y (the n observations must be uncorrelated)
- Average amount $\hat{\mu}$ differs from μ - larger n, smaller error

Accuracy of Coefficient Estimates

- In the same vein - How close can we make $\hat{\beta}_0$ and $\hat{\beta}_1$ to β_0 and β_1?

Accuracy of Coefficient Estimates

- In the same vein - How close can we make $\hat{\beta}_0$ and $\hat{\beta}_1$ to β_0 and β_1?

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right)$$
$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},\ \sigma^2 = Var(\epsilon)$$

- We assume ϵ_i are uncorrelated with common variance σ^2 (Often not true but a good approximation)

Accuracy of Coefficient Estimates

- In the same vein - How close can we make $\hat{\beta}_0$ and $\hat{\beta}_1$ to β_0 and β_1?

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right)$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},\ \sigma^2 = Var(\epsilon)$$

- We assume ϵ_i are uncorrelated with common variance σ^2 (Often not true but a good approximation)
- When x_i are spread out, and smaller, we have more leverage to estimate the slope, reducing $SE(\hat{\beta}_1)$
- $SE(\hat{\beta}_0) = SE(\hat{\mu})$ if $\overline{x} = 0$

Accuracy of Coefficient Estimates

- In the same vein - How close can we make $\hat{\beta}_0$ and $\hat{\beta}_1$ to β_0 and β_1?

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right)$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},\ \sigma^2 = Var(\epsilon)$$

- We assume ϵ_i are uncorrelated with common variance σ^2 (Often not true but a good approximation)
- When x_i are spread out, and smaller, we have more leverage to estimate the slope, reducing $SE(\hat{\beta}_1)$
- $SE(\hat{\beta}_0) = SE(\hat{\mu})$ if $\overline{x} = 0$
- σ^2 is not known either but can be estimated from data. the estimate, σ is the residual standard error:

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

Coefficient Estimates: Confidence Intervals

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right)$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},\ \sigma^2 = Var(\epsilon)$$

$$\hat{\beta} \pm 2SE(\hat{\beta})$$

Hypothesis Testing

- Standard Errors let us hypothesis test.
- Most common is the Null Hypothesis
- H_0: There is no relationship between X and Y
- Alternatively we have H_a: There is some relationship between X and Y
- Mathematically, this is like testing H_0: $\beta_1 = 0$ therefore $Y = \beta_0 + \epsilon$
- H_a: $\beta_1 \neq 0$ therefore determine that $\hat{\beta}_1$ is sufficiently far from 0
- The important question becomes - how far is far enough?

T-Statistic

t-statistic $t_\beta = \frac{\hat{\beta}_1 - \beta}{SE(\hat{\beta}_1)}$

t-statistic $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ for H_0

T-Statistic

t-statistic $t_\beta = \frac{\hat{\beta}_1 - \beta}{SE(\hat{\beta}_1)}$

t-statistic $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ for H_0

- If no relationship between X and Y exists, we expect a t-distribution with n-2 degrees of freedom
- Compute the probability of observing any number equal to —t— or larger in absolute value, assuming $\beta_1 = 0$
- This probability is called the p-value
- A small p-value - it is unlikely to observe a substantial association between predictor and response due to chance
- Therefore a small p-value means there is an association between X and Y so we can *reject the null hypothesis*
- The cutoff is usually 5% or 1%

Advertising Example

If n $=$ 30

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

With $n = 30$ the t-statistic for the null hypothesis are around 2 and 2.75 respectively

We conclude $\beta_0 \neq 0$ and $\beta_1 \neq 0$

Important Questions to Ask

- Is there a relationship between budget and sales?
- If there is a relationship, how strong is it?
- Which of the three media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Accuracy of Simple Linear Regression

- Once we reject the null hypothesis for β_0 and β_1, it is natural to ask how well the model fits the data
- One measure is the residual standard error
$$RSE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$
- Measure of lack of fit, it is an absolute measure. It is not always clear what a good value for RSE is
- Another possible measurement is the R^2 statistic

R^2 Statistic

- Proportion of variance explained, always between 0 and 1, independent of scale of Y
- Total sum of squares $TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$
- $R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$

R^2 Statistic

- Proportion of variance explained, always between 0 and 1, independent of scale of Y
- Total sum of squares $TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$
- $R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$
- TSS measures the total variance in response Y (amount inherent in response before the regression is performed)
- RSS amount left unexplained after the regression

R^2 Statistic

- $R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$
- R^2 is the proportion of variability in Y that can be explained using X
- R^2 close to 1 - large proportion of variation explained by the regression
- R^2 close to 0 - regression did not explain the variation - perhaps because model is wrong, σ^2 is too high, or possibly both?
- R^2 is a measure of the linear relationship between X and Y
- Still. What is a good value for R^2?

R^2 Statistic: Correlation

- $Cor(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$
- This is also a measure of the linear relationship between X and Y
- $r = Cor(X, Y)$
- in Simple linear regression, $R^2 = r^2$. In multiple regression however r^2 does not extend

Takeaways and Next Time

- Ordinary Least Squares Optimization
- Linear Regression
- Next Time: More variables!
- example and figure sources: James, Witten, Hastie, Tibshirani (ISLR)