# CSCE 421: Machine Learning

## Lecture 3: Linear Regression


Texas A&M University
Section 201/501
Bobak Mortazavi
Ryan King
Zhale Nowroozilarki

# Goals For This Lecture

- Motivate a simple supervised learning problem
- Introduce a linear machine learning method (Linear regression)
- Develop a Loss Function
- Ordinary Least Squares - Optimally solve the learning problem
- Interpret model
- Understanding Accuracy and Error
- Acknowledgements: example and figure sources: James, Witten, Hastie, Tibshirani (ISLR)

# Notation and Modeling

- $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$
- $\boldsymbol{x}_i$ a column vector of length p, with N samples
- $y_i$ a scalar
- for p = 1, linear regression is fitting line to data in 2-dimensional space
- in general, linear regression is about fitting a hyperplane to a scatter of points in p + 1 dimensional space

# Notation and Modeling

- Consider the p dimensional case
- The objective is determining intercept $w_0$ and p slope weights $w_i' s$ so that for all N datapoints:

$$w_0 + x_{1,i}w_1 + x_{2,i}w_2 + \cdots + x_{p,i}w_p \approx y_i$$

- Putting it into the vector form:

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \dots \\ w_p \end{bmatrix}, \dot{x}_i = \begin{bmatrix} 1 \\ x_{1,i} \\ \dots \\ x_{p,i} \end{bmatrix}$$

- $\dot{x_i}$ obtained by stacking a 1 on top of $x_p$
- Our linear equation would be

$$x_i^T \, w \ \approx \ y_i, i = 1, \dots, N$$

# 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
  - Radio
  - 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?

# 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:

$$w_0 + w_1 x \approx y$$

# Simple Linear Regression

We want to predict y based upon a single predictor x, we want to regress y on to x:

$$w_0 + w_1 x \approx y$$
$$w_0 + w_1 TV \approx Sales$$

# Parameters

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

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

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

# Estimating the Coefficients

- We do not know $w_0$ or $w_1$
- So, assume we have a training set $D = \{(x_1, y_1), \cdots, (x_N, y_N)\}$
- Assume $\mathrm{N} = 200$ markets of sales and tv budget
- Goal: set $\hat{w}_0$ and $\hat{w}_1$ so we are as close to $y_i$ from $x_i$ for all $i$

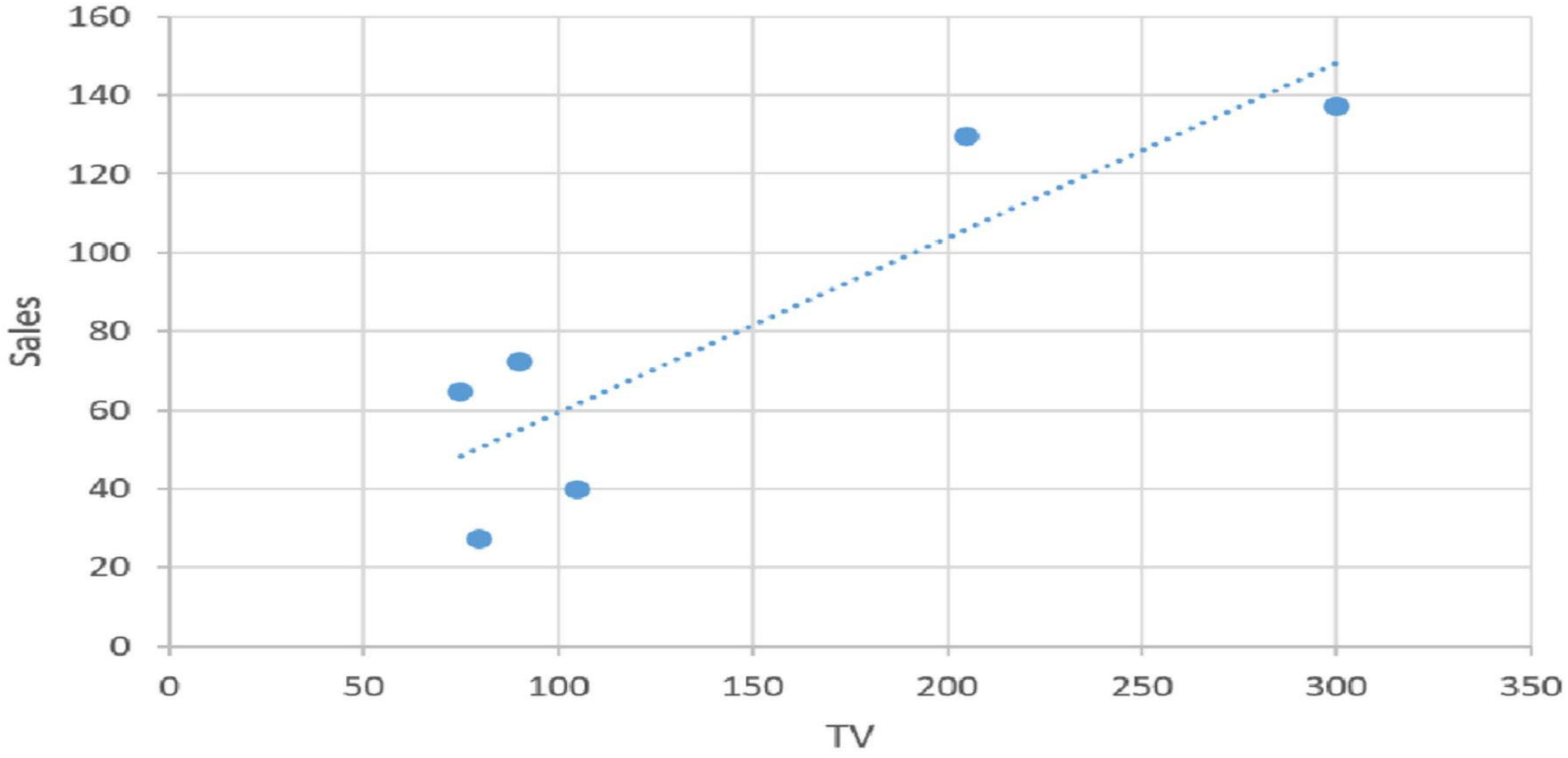
# Residual

- Let $\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i$ be the prediction for $y$ based on the i'th 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 while considering this total error

# Sum of Residual



# Least Squares

The residual sum of squares

$$RSS = e_1^2 + e_2^2 + \cdots + e_N^2$$
$$= (y_1 - \hat{w}_0 - \hat{w}_1 x_1)^2 + \cdots + (y_N - \hat{w}_0 - \hat{w}_1 x_N)^2$$

# Least Squares: Learning Coefficients

The residual sum of squares

$$RSS = e_1^2 + e_2^2 + \cdots + e_N^2$$
$$= (y_1 - \hat{w}_0 - \hat{w}_1 x_1)^2 + \cdots + (y_\mathrm{N} - \hat{w}_0 - \hat{w}_1 x_N)^2$$

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

$$RSS = \sum_{\mathrm{i}=1}^{N} (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$$

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

# Differentiation: $\widehat{w}_0$

$RSS = \sum_{\mathrm{i}=1}^{N}(y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$

# Differentiation: $\widehat{w}_0$

$RSS = \sum_{\mathrm{i}=1}^{N} (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$

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

# Differentiation: $\hat{w}_0$

$RSS = \sum_{\mathrm{i}=1}^{N}(y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$

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

$= -2\sum_{\mathrm{i}=1}^{N}(y_i - \hat{w}_0 - \hat{w}_1 x_i)$

# Differentiation: $\hat{w}_0$

$RSS = \sum_{\mathrm{i}=1}^{N} (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$

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

$= -2\sum_{\mathrm{i}=1}^{N} (y_i - \hat{w}_0 - \hat{w}_1 x_i)$

$= -2\sum_{\mathrm{i}=1}^{N} y_i + 2\sum_{\mathrm{i}=1}^{N} \hat{w}_0 + 2\hat{w}_1 \sum_{\mathrm{i}=1}^{N} x_i$

# Differentiation: $\widehat{w}_0$

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

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

$= -2\sum_{i=1}^{N} (y_i - \hat{w}_0 - \hat{w}_1 x_i)$

$= -2\sum_{i=1}^{N} y_i + 2\sum_{i=1}^{N} \hat{w}_0 + 2\hat{w}_1 \sum_{i=1}^{N} x_i$

**Note:** $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ is the sample mean

# Differentiation: $\hat{w}_0$

$RSS = \sum_{\mathrm{i}=1}^{N}(y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$

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

$= -2\sum_{\mathrm{i}=1}^{N}(y_i - \hat{w}_0 - \hat{w}_1 x_i)$

$= -2\sum_{\mathrm{i}=1}^{N} y_i + 2\sum_{\mathrm{i}=1}^{N} \hat{w}_0 + 2\hat{w}_1 \sum_{\mathrm{i}=1}^{N} x_i$

**Note:** $\bar{y} = \frac{1}{N}\sum_{\mathrm{i}=1}^{N} y_i$ is the sample mean

$= -2N\bar{y} + 2N\hat{w}_0 + 2N\hat{w}_1\bar{x}$

# Differentiation: $\hat{w}_0$

$RSS = \sum_{\mathrm{i}=1}^{N} (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$

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

$= -2\sum_{\mathrm{i}=1}^{N} (y_i - \hat{w}_0 - \hat{w}_1 x_i)$

$= -2\sum_{\mathrm{i}=1}^{N} y_i + 2\sum_{\mathrm{i}=1}^{N} \hat{w}_0 + 2\hat{w}_1 \sum_{\mathrm{i}=1}^{N} x_i$

**Note:** $\bar{y} = \frac{1}{N}\sum_{\mathrm{i}=1}^{N} y_i$ is the sample mean

$= -2N\bar{y} + 2N\hat{w}_0 + 2N\hat{w}_1\bar{x}$

To minimize, set $\frac{\partial RSS}{\partial \hat{w}_0} = 0$

# Differentiation: $\hat{w}_0$

$RSS = \sum_{\mathrm{i}=1}^{N} (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$

$$\frac{\partial RSS}{\partial \hat{w}_0} = -2N\bar{y} + 2N\hat{w}_0 + 2N\hat{w}_1\bar{x}$$

To minimize, set $\frac{\partial RSS}{\partial \hat{w}_0} = 0$

$$-2N\bar{y} + 2N\hat{w}_0 + 2N\hat{w}_1\bar{x} = 0$$

# Differentiation: $\hat{w}_0$

$RSS = \sum_{\mathrm{i}=1}^{N}(y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$

$$\frac{\partial RSS}{\partial \hat{w}_0} = -2N\bar{y} + 2N\hat{w}_0 + 2N\hat{w}_1\bar{x}$$

To minimize, set $\frac{\partial RSS}{\partial \hat{w}_0} = 0$

$$-2N\bar{y} + 2N\hat{w}_0 + 2N\hat{w}_1\bar{x} = 0$$

$$2N\hat{w}_0 = 2N\bar{y} - 2N\hat{w}_1\bar{x}$$

# Differentiation: $\widehat{w}_0$

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

$$\frac{\partial RSS}{\partial \hat{w}_0} = -2N\bar{y} + 2N\hat{w}_0 + 2N\hat{w}_1\bar{x}$$

To minimize, set $\frac{\partial RSS}{\partial \hat{w}_0} = 0$

$-2N\bar{y} + 2N\hat{w}_0 + 2N\hat{w}_1\bar{x} = 0$

$2N\hat{w}_0 = 2N\bar{y} - 2N\hat{w}_1\bar{x}$

$\cancel{2N}\hat{w}_0 = \cancel{2N}\bar{y} - \cancel{2N}\hat{w}_1\bar{x}$

# Differentiation: $\widehat{w}_0$

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

$$\frac{\partial RSS}{\partial \hat{w}_0} = -2N\bar{y} + 2N\hat{w}_0 + 2N\hat{w}_1\bar{x}$$

To minimize, set $\frac{\partial RSS}{\partial \hat{w}_0} = 0$

$-2N\bar{y} + 2N\hat{w}_0 + 2N\hat{w}_1\bar{x} = 0$

$2N\hat{w}_0 = 2N\bar{y} - 2N\hat{w}_1\bar{x}$

$\cancel{2N}\hat{w}_0 = \cancel{2N}\bar{y} - \cancel{2N}\hat{w}_1\bar{x}$

$\hat{w}_0^* = \bar{y} - \hat{w}_1\bar{x}$

# Differentiation: $\widehat{w}_1$

$RSS = \sum_{\mathrm{i}=1}^{N} (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$

# Differentiation: $\hat{w}_1$

$RSS = \sum_{\mathrm{i}=1}^{N} (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$

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

# Differentiation: $\widehat{w}_1$

$$RSS = \sum_{\mathrm{i}=1}^{N}(y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$$

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

$$= -2\sum_{\mathrm{i}=1}^{N}(y_i - \hat{w}_0 - \hat{w}_1 x_i)(x_i)$$

# Differentiation: $\hat{w}_1$

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

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

$= -2\sum_{i=1}^{N}(y_i - \hat{w}_0 - \hat{w}_1 x_i)(x_i)$

Set equal to 0

$-2\sum_{i=1}^{N} y_i x_i + 2\hat{w}_0 \sum_{i=1}^{N} x_i + 2\hat{w}_1 \sum_{i=1}^{N} x_i^2 = 0$

# Differentiation: $\hat{w}_1$

$$-2\sum_{i=1}^{N} y_i x_i + 2\hat{w}_0 \sum_{i=1}^{N} x_i + 2\hat{w}_1 \sum_{i=1}^{N} x_i^2 = 0$$

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

# Differentiation: $\hat{w}_1$

$$-2\sum_{i=1}^{N} y_i x_i + 2\hat{w}_0 \sum_{i=1}^{N} x_i + 2\hat{w}_1 \sum_{i=1}^{N} x_i^2 = 0$$

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

$$= -\sum_{i=1}^{N} y_i x_i + (\bar{y} - \hat{w}_1 \bar{x}) \sum_{i=1}^{N} x_i + \hat{w}_1 \sum_{i=1}^{N} x_i^2 = 0$$

# Differentiation: $\hat{w}_1$

$$-2\sum_{\mathrm{i}=1}^{N} y_i x_i + 2\hat{w}_0 \sum_{\mathrm{i}=1}^{N} x_i + 2\hat{w}_1 \sum_{\mathrm{i}=1}^{N} x_i^2 = 0$$

$$= -\cancel{2}\sum_{\mathrm{i}=1}^{N} y_i x_i + \cancel{2}\hat{w}_0 \sum_{\mathrm{i}=1}^{N} x_i + \cancel{2}\hat{w}_1 \sum_{\mathrm{i}=1}^{N} x_i^2 = 0$$

$$= -\sum_{\mathrm{i}=1}^{N} y_i x_i + (\bar{y} - \hat{w}_1 \bar{x}) \sum_{\mathrm{i}=1}^{N} x_i + \hat{w}_1 \sum_{\mathrm{i}=1}^{N} x_i^2 = 0$$

$$= -\sum_{\mathrm{i}=1}^{N} y_i x_i + \sum_{\mathrm{i}=1}^{N} x_i + \hat{w}_1 \sum_{\mathrm{i}=1}^{N} x_i^2 = 0$$

# Differentiation: $\widehat{w}_1$

$$-2\sum_{\mathrm{i}=1}^{N} y_i x_i + 2\hat{w}_0 \sum_{\mathrm{i}=1}^{N} x_i + 2\hat{w}_1 \sum_{\mathrm{i}=1}^{N} x_i^2 = 0$$

$$= -\cancel{2}\sum_{\mathrm{i}=1}^{N} y_i x_i + \cancel{2}\hat{w}_0 \sum_{\mathrm{i}=1}^{N} x_i + \cancel{2}\hat{w}_1 \sum_{\mathrm{i}=1}^{N} x_i^2 = 0$$

$$= -\sum_{\mathrm{i}=1}^{N} y_i x_i + (\bar{y} - \hat{w}_1 \bar{x}) \sum_{\mathrm{i}=1}^{N} x_i + \hat{w}_1 \sum_{\mathrm{i}=1}^{N} x_i^2 = 0$$

$$= -\sum_{i=1}^{N} y_i x_i + \bar{y} \sum_{i=1}^{N} x_i - \hat{w}_1 \bar{x} \sum_{i=1}^{N} x_i + \hat{w}_1 \sum_{i=1}^{N} x_i^2 = 0$$

$$\bar{y} \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i = \widehat{w}_1 \bar{x} \sum_{i=1}^{N} x_i - \widehat{w}_1 \sum_{i=1}^{N} x_i^2$$

$$\bar{y} \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i = \widehat{w}_1 \left(\bar{x} \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i^2\right)$$

$$\widehat{w}_1^* = \frac{\bar{y} \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i}{\bar{x} \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i^2}$$

# Differentiation: $\widehat{w}_1$

$$\widehat{w}_1^* = \frac{\bar{y}\ \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i}{\bar{x} \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i^2}$$

$$\widehat{w}_1^* = \frac{\bar{y}\ \bar{x} N - \sum_{i=1}^{N} y_i x_i}{\bar{x}^2 N - \sum_{i=1}^{N} x_i^2}$$

$$\widehat{w}_1^* = \frac{\sum_{i=1}^{N} y_i x_i - \bar{y}\ \bar{x} N}{\sum_{i=1}^{N} x_i^2 - \bar{x}^2 N}$$

# Differentiation: $\hat{w}_1$ - Numerator

$$\sum_{i=1}^{N} y_i x_i - \bar{y}\,\bar{x}N$$

$$\sum_{i=1}^{N} y_i x_i - \bar{y}\,\bar{x}N - \bar{y}\,\bar{x}N + \bar{y}\,\bar{x}N$$

$$\sum_{i=1}^{N} y_i x_i - \bar{y}\,\sum_{i=1}^{N} x_i - \bar{x}\sum_{i=1}^{N} y_i + \bar{y}\,\bar{x}N$$

$$\sum_{i=1}^{N} y_i x_i - \bar{y}\,\sum_{i=1}^{N} x_i - \bar{x}\sum_{i=1}^{N} y_i + \bar{y}\,\bar{x}\sum_{i=1}^{N} 1$$

$$\sum_{i=1}^{N} y_i x_i - \bar{y}\,\sum_{i=1}^{N} x_i - \bar{x}\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \bar{y}\,\bar{x}$$

$$\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} \bar{y} x_i - \sum_{i=1}^{N} \bar{x} y_i + \sum_{i=1}^{N} \bar{y}\,\bar{x}$$

$$\sum_{i=1}^{N} (y_i x_i - \bar{y} x_i + \bar{x} y_i + \bar{y}\,\bar{x})$$

$$\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

# Differentiation: $\hat{w}_1$

$$\hat{w}_1^* = \frac{\bar{y}\ \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i}{\bar{x}\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i^2}$$

$$\hat{w}_1^* = \frac{\bar{y}\ \bar{x}N - \sum_{i=1}^{N} y_i x_i}{\bar{x}^2 N - \sum_{i=1}^{N} x_i^2}$$

$$\hat{w}_1^* = \frac{\sum_{i=1}^{N} y_i x_i - \bar{y}\ \bar{x}N}{\sum_{i=1}^{N} x_i^2 - \bar{x}^2 N}$$

$$\hat{w}_1^* = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} x_i^2 - \bar{x}^2 N}$$

# Differentiation: $\hat{w}_1$ - Denominator

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

# Differentiation: $\widehat{w}_1$

$$\widehat{w}_1^* = \frac{\bar{y}\ \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i}{\bar{x} \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i^2}$$

$$\widehat{w}_1^* = \frac{\bar{y}\ \bar{x} N - \sum_{i=1}^{N} y_i x_i}{\bar{x}^2 N - \sum_{i=1}^{N} x_i^2}$$

$$\widehat{w}_1^* = \frac{\sum_{i=1}^{N} y_i x_i - \bar{y}\ \bar{x} N}{\sum_{i=1}^{N} x_i^2 - \bar{x}^2 N}$$

$$\widehat{w}_1^* = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} x_i^2 - \bar{x}^2 N}$$

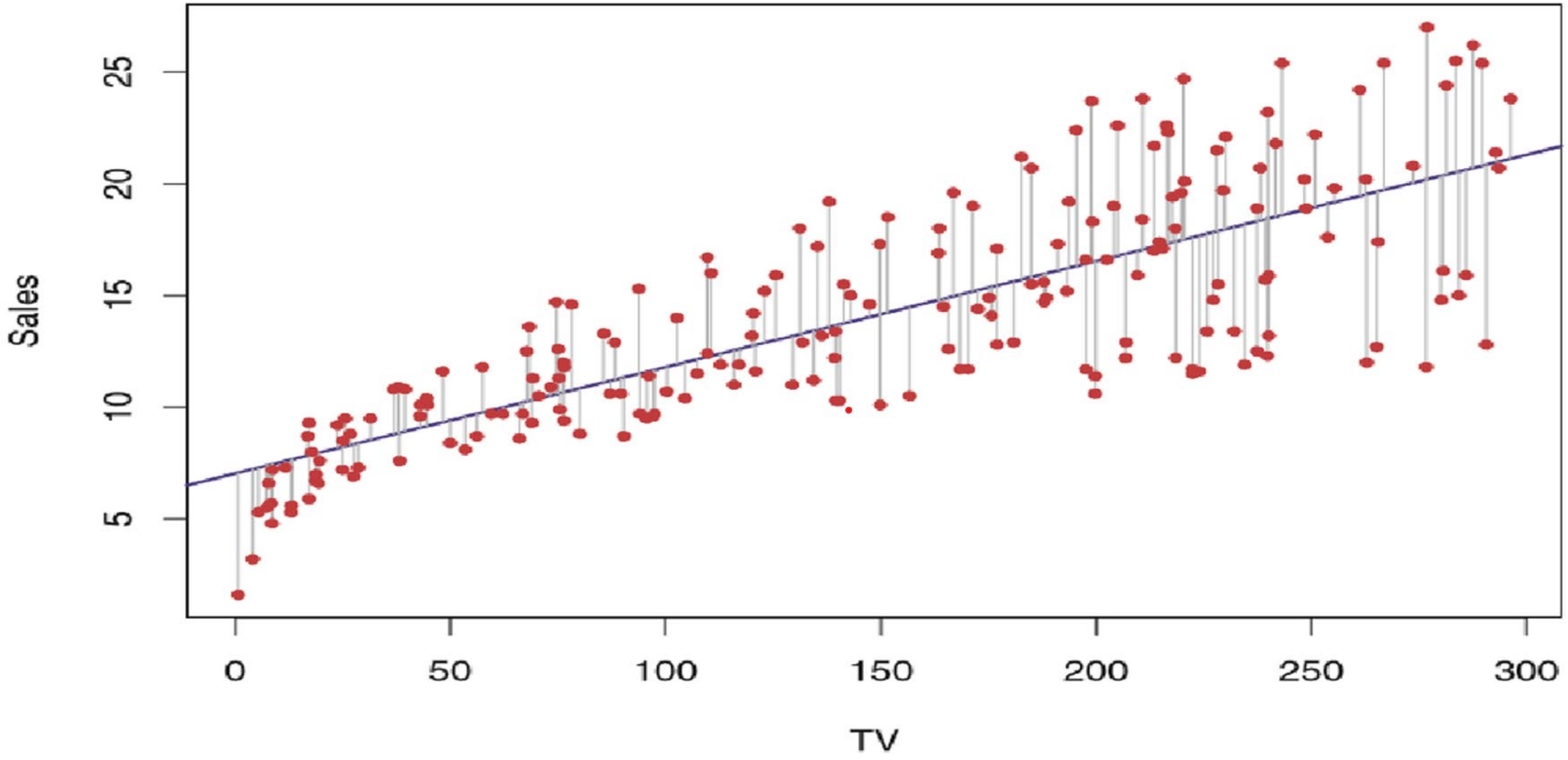
$$\widehat{w}_1^* = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

# Optimal Coefficients: $\widehat{\boldsymbol{w}}_0, \widehat{\boldsymbol{w}}_1$

$$\hat{w}_0^* = \bar{y} - \hat{w}_1\bar{x}$$

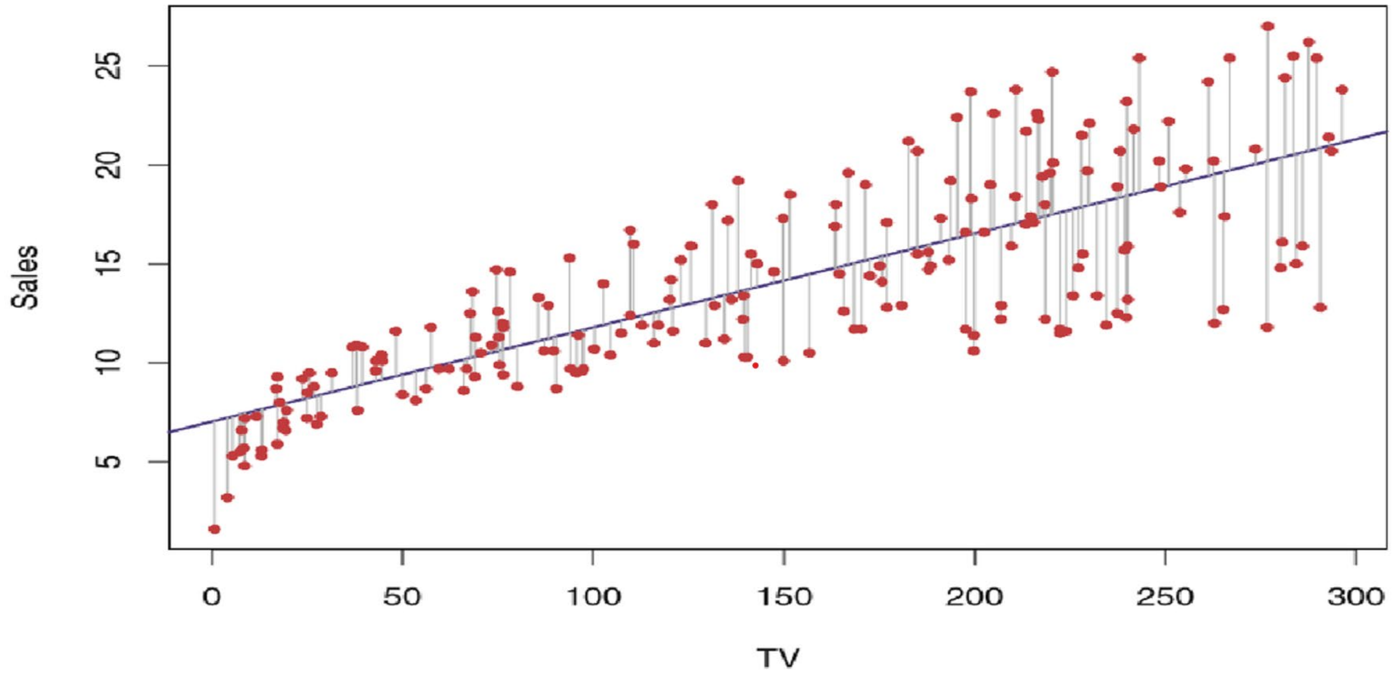
$$\widehat{w}_1^* = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

# Advertising Solution




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

# Advertising Solution




$\hat{w}_0 = 7.03$ and $\hat{w}_1 = 0.0475$. If we had no TV advertising, how many units would we sell? What if we had \$1000 budgeted for TV?

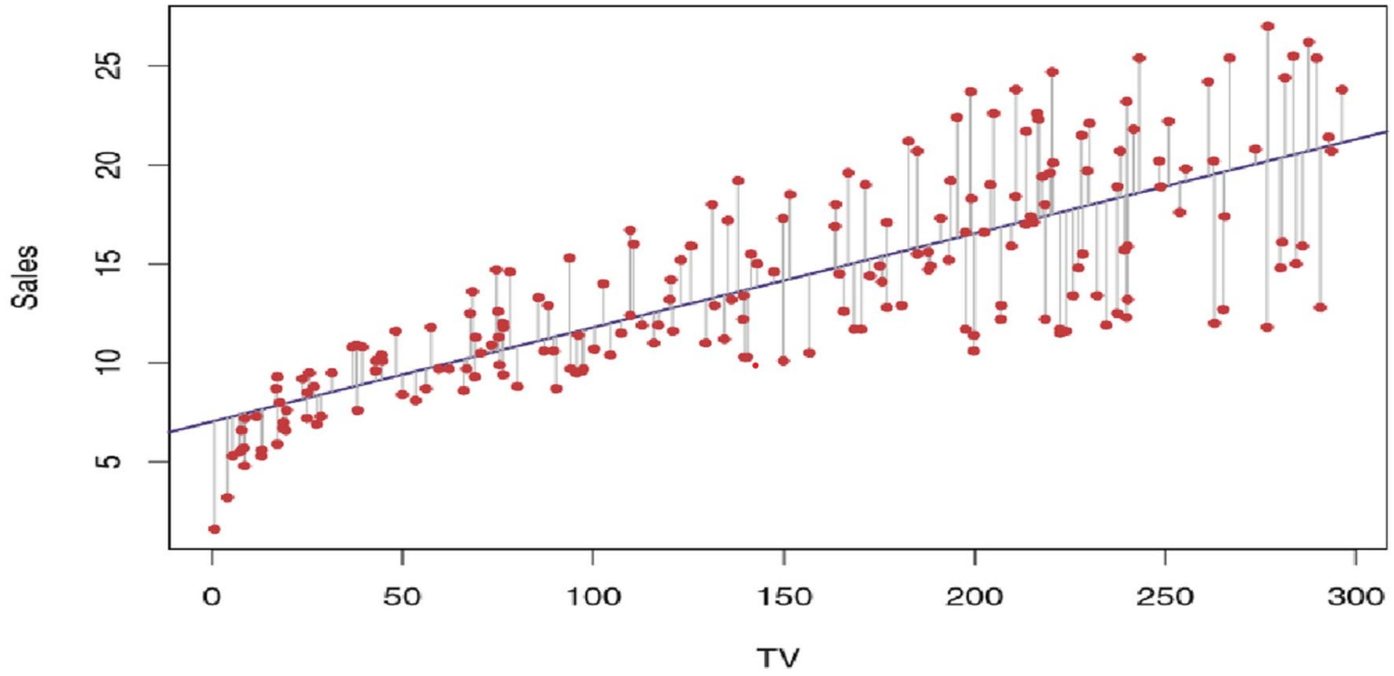
A. 703, 475 + 703

B. 7.03, 47.5 + 7.03

C. 47.5 + 7.03, 7.03

D. 475 + 703, 703

# Advertising Solution




$\hat{w}_0 = 7.03$ and $\hat{w}_1 = 0.0475$. If we had no TV advertising, how many units would we sell? What if we had $1000 budgeted for TV?

A. 703, 475 + 703

B. 7.03, 47.5 + 7.03

C. 47.5 + 7.03, 7.03

D. 475 + 703, 703

# Least Absolutes

- The residual sum of absolutes

$$RSS = |e_1| + |e_2| \cdots |e_N|$$

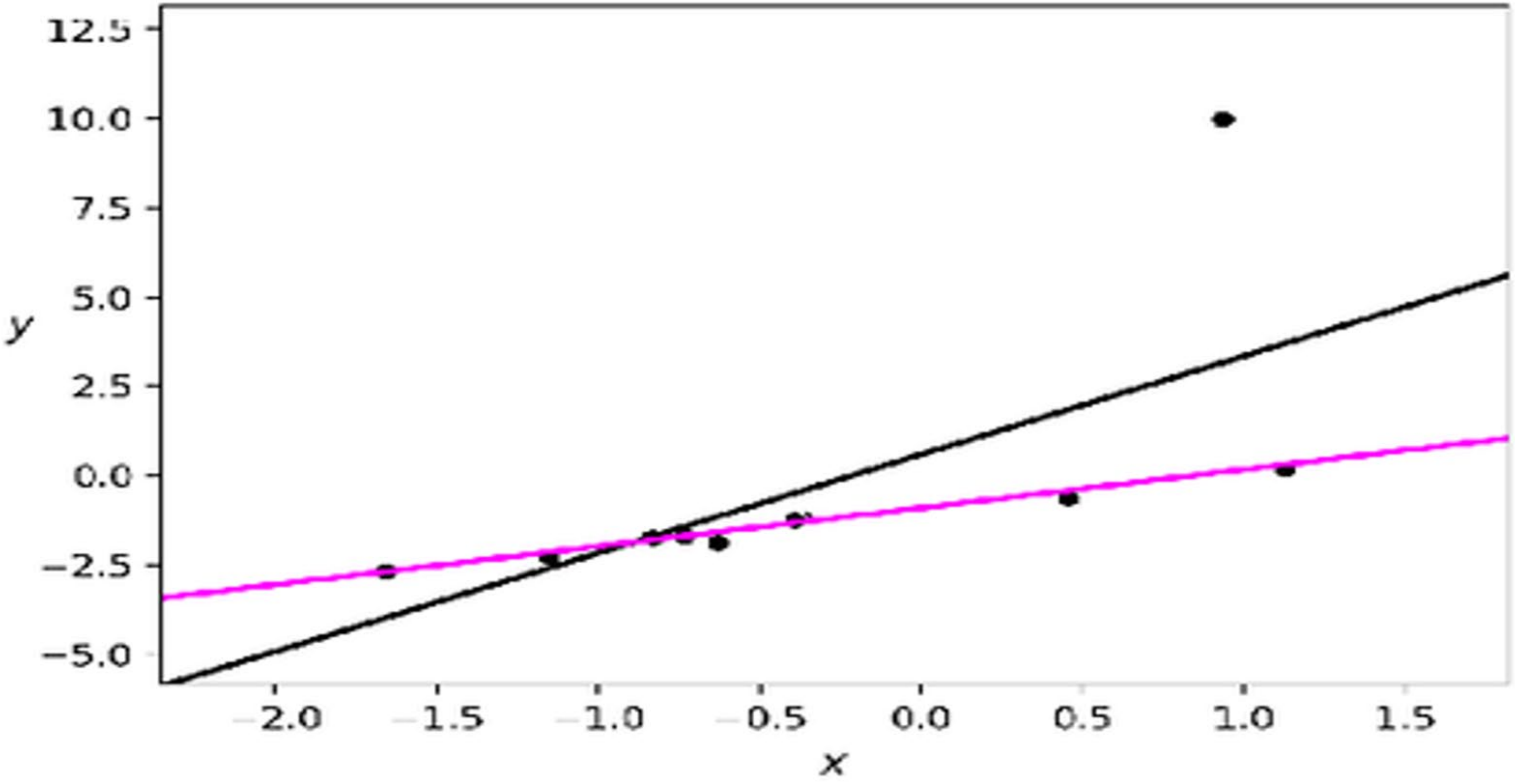
$$= |y_1 - \hat{w}_0 - \hat{w}_1 x_1| + |y_2 - \hat{w}_0 - \hat{w}_1 x_2| \cdots |y_N - \hat{w}_0 - \hat{w}_1 x_N|$$

# Least Absolutes

- Downside of least square cost:
  - Squaring errors larger than 1 emphasizes them
  - Forces the weights to minimize larger errors, typically those of outliers
  - Susceptible to overfitting to outliers
- Least absolute error partially addresses this problem

# Least Absolutes

- Black line fitted using least squares
- Pink line fitted using least absolute



# Accuracy of Coefficient Estimates

- Assume the true relationship is $y = f(x) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ (mean zero random error term)

- So, $y = w_0 + w_1 x + \epsilon$

# Accuracy of Coefficient Estimates

- Assume the true relationship is $y = f(x) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ (mean zero random error term)
- So, $y = w_0 + w_1 x + \epsilon$
- This is the population regression line which is the best linear approximation to the true relationship between x and y.

# Accuracy of Coefficient Estimates

- Assume the true relationship is $y = f(x) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ (mean zero random error term)

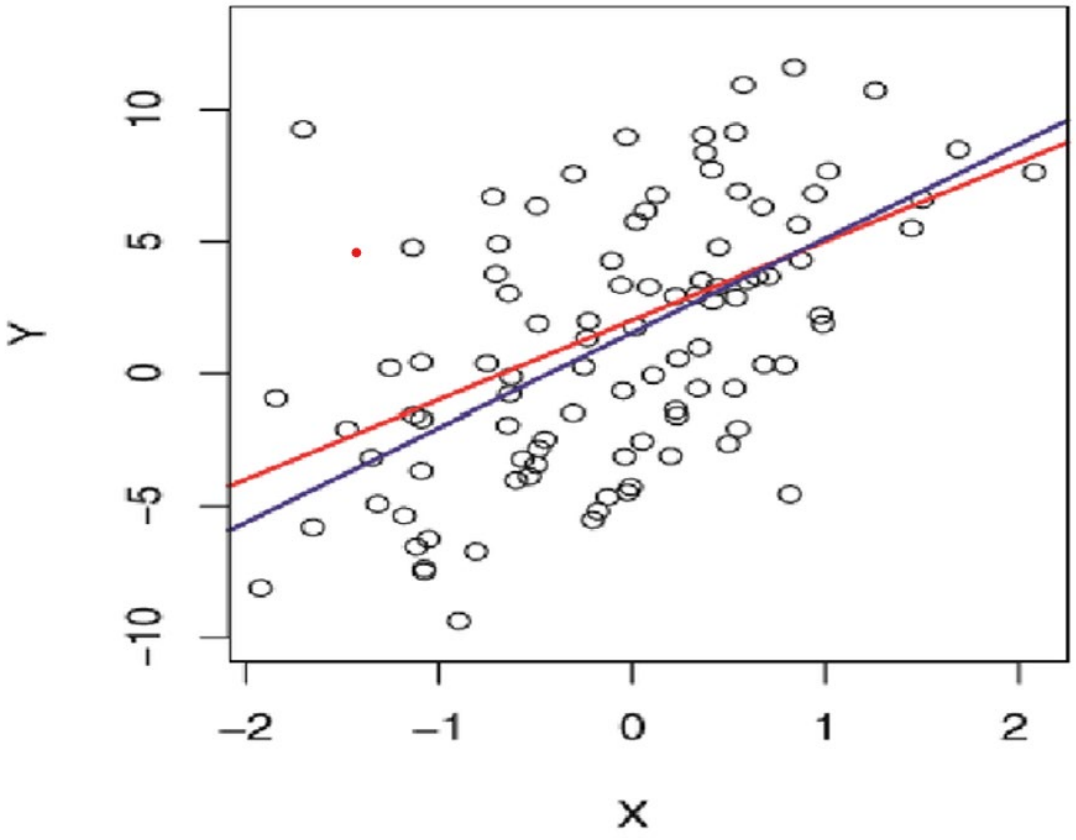
- So, $y = w_0 + w_1 x + \epsilon$

- 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.

- $\hat{\mu} = \bar{y}$ - sample mean from observations recorded is close with lots of sampling. Same $\widehat{w}_0$ and $\widehat{w}_1$ - 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

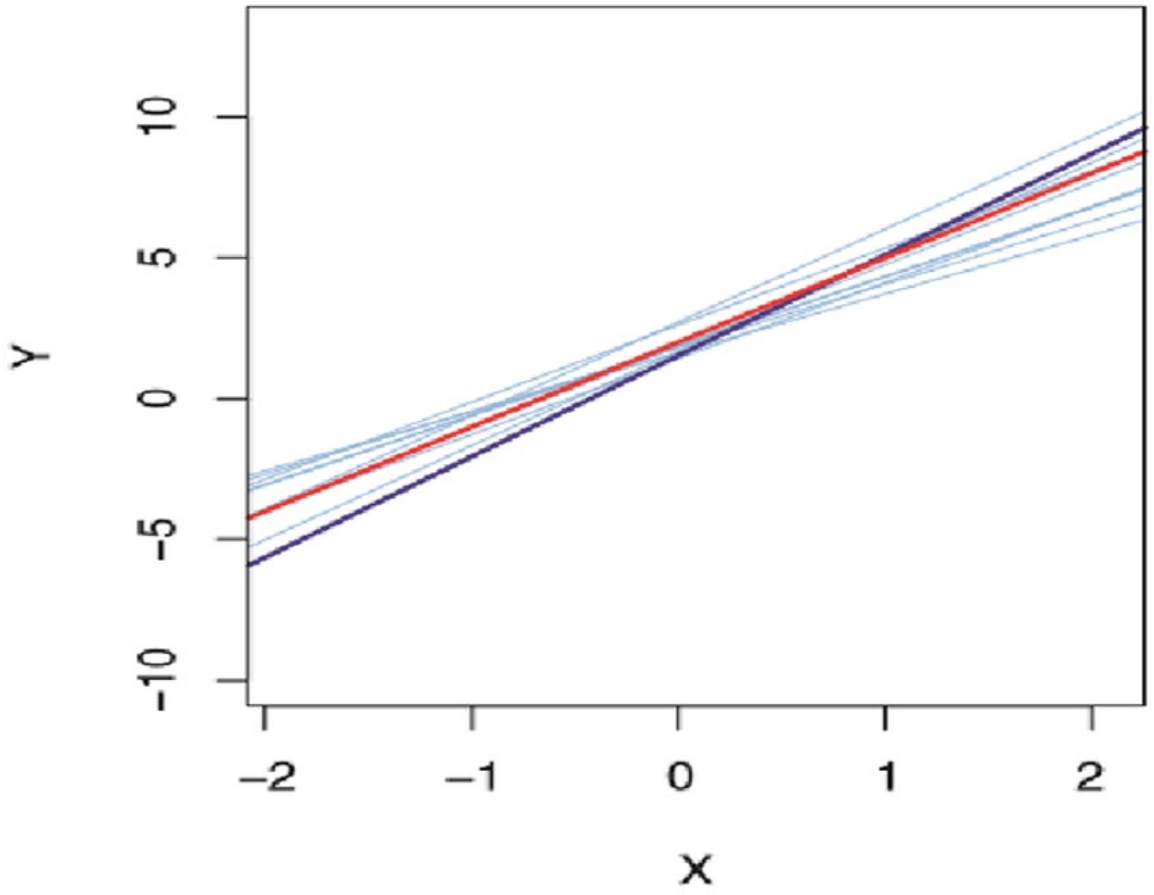
- Assume, for example $y = 2 + 3x + \epsilon$ and you sample this population with 100 random variables x to generate 100 y.
- $\hat{\mu} = \bar{y}$ - sample mean from observations recorded is close with lots of sampling. Same $\widehat{w}_0$ and $\widehat{w}_1$ - is a good estimate with enough data.
- 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 $\mu$, this is unbiased since, on average, we expect them to be the same.
  - One set of $y_1, y_2, \cdots, y_N$ might result in $\hat{\mu}$ that underestimates $\mu$
  - Another that overestimates $\mu$
  - etc

# Accuracy of Coefficient Estimates

- Same with $\hat{w}_0$ and $\hat{w}_1$ - average enough samples and enough regressions to get to the true $w_0$ and $w_1$

# 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 $\hat{w}_0$ and $\hat{w}_1$ - average enough samples and enough regressions to get to the true $w_0$ and $w_1$
- So, we ask, how accurate is the sample mean $\hat{\mu}$ from the estimate of $\mu$ – how far off is a single estimate?

# Accuracy of Coefficient Estimates

- Same with $\hat{w}_0$ and $\hat{w}_1$ - average enough samples and enough regressions to get to the true $w_0$ and $w_1$

- So, we ask, how accurate is the sample mean $\hat{\mu}$ from the estimate of $\mu$ – how far off is a single estimate?

- We need to calculate the standard error of $\hat{\mu}$, $\mathrm{SE}(\hat{\mu})$

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

- Where $\sigma^2$ 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 $\mu$ – larger N, smaller error

# Accuracy of Coefficient Estimates

- In the same vein – How close can we make $\hat{w}_0$ and $\hat{w}_1$ to $w_0$ and $w_1$?

# Accuracy of Coefficient Estimates

- In the same vein – How close can we make $\widehat{w}_0$ and $\widehat{w}_1$ to $w_0$ and $w_1$?

$$\mathrm{SE}(\widehat{w}_0)^2 = \sigma^2\left(\frac{1}{N} + \frac{x^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)$$

$$\mathrm{SE}(\widehat{w}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \sigma^2 = Var(\epsilon)$$

- We assume $\epsilon_i$ are uncorrelated with common variance $\sigma^2$ (Often not true but a good approximation)

# Accuracy of Coefficient Estimates

- In the same vein – How close can we make $\widehat{w}_0$ and $\widehat{w}_1$ to $w_0$ and $w_1$?

$$\mathrm{SE}(\widehat{w}_0)^2 = \sigma^2 \left( \frac{1}{N} + \frac{x^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \right)$$

$$\mathrm{SE}(\widehat{w}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \sigma^2 = Var(\epsilon)$$

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

# Accuracy of Coefficient Estimates

- In the same vein – How close can we make $\widehat{w}_0$ and $\widehat{w}_1$ to $w_0$ and $w_1$?

$$\mathrm{SE}(\widehat{w}_0)^2 = \sigma^2\left(\frac{1}{N} + \frac{x^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right)$$

$$\mathrm{SE}(\widehat{w}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \sigma^2 = Var(\epsilon)$$

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

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

# Coefficient Estimates: Confidence Intervals

$$\mathrm{SE}(\hat{w}_0)^2 = \sigma^2 \left( \frac{1}{N} + \frac{x^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \right)$$

$$\mathrm{SE}(\hat{w}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \sigma^2 = Var(\epsilon)$$

$$\hat{w} \pm SE(\hat{w})$$

# Hypothesis Testing

- Standard Errors let us hypothesis test
- Most common is the Null Hypothesis
- $H_0$: There is no relation between x and y
- Alternatively, we have $H_a$: There is some relationship between x and y
- Mathematically, this is like testing
  - $H_0$: $\mathrm{w}_1 = 0$ Therefore $y = w_0 + \epsilon$
  - $H_a$: $w_1 \neq 0$ therefore determine that $\widehat{w}_1$ is sufficiently far from 0
- The important question becomes – how far is far enough?

# T-Statistic

- T-statistic $t_w = \frac{\hat{w}_1 - w}{SE(\hat{w}_1)}$
- T-statistic $t_w = \frac{\hat{w}_1 - 0}{SE(\hat{w}_1)}$ for $H_0$

# T-Statistic

- T-statistic $t_w = \frac{\hat{w}_1 - w}{SE(\hat{w}_1)}$
- T-statistic $t_w = \frac{\hat{w}_1 - 0}{SE(\hat{w}_1)}$ for $H_0$
- If no relationship between x and y exists, we expect a t-distribution with P-2 degrees of freedom
- Compute the probability of observing any number equal to ---t--- or larger in absolute value, assuming $w_1 = 0$
- 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 P=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 P=30 the t-statistic for the null hypothesis are around 2 and 2.75 respectively.

We conclude $w_0 \neq 0$ and $w_1 \neq 0$

# 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 $w_0$ and $w_1$, 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 of 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 $TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

# $R^2$ Statistic

- Proportion of variance explained, always between 0 and 1, independent of scale of y
- Total sum of squares $TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

- 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

- $R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$

- $R^2$ is the proportion of variability in y that can be explained using $\boldsymbol{x}$.
- $R^2$ close to 1 – large proportion of variation explained by the regression
- $R^2$ close to 0 – regression id not explain the variation – perhaps because model is wrong. $\sigma^2$ 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

$$Cor(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

- 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

- Understanding key notation
- Important questions to ask for supervised learning problem
- Ordinary Least Squares
- Simple Linear Regression
- Optimizing RSS
- Next Time: Multiple Linear Regression and Coding