

Simple and Multiple Linear Regression

Fatemeh Asgarinejad
February 6, 2025

Announcements + schedule + links

- Lectures
- Office Hours
- GitHub (code + slides) | [Google Colab](#)
 - Slides: <https://github.com/Fasgarinejad/Machine-Learning> or scan
- **Announcements**



Course Plan: Where We Are Now

Week	Date	Topic	Activities	Week	Date	Topic	Activities				
1	Tuesday – Feb 4	Introduction to AI		10	Tuesday – April 8		Quiz 4– April 8				
	Thursday – Feb 6	Linear regression 			Thursday – April 10						
2	Tuesday – Feb 11	Linear classifier	Quiz 1 – Feb 11	11	Tuesday – April 15						
	Thursday – Feb 13				Thursday – April 17						
3	Tuesday – Feb 18			12	Tuesday – April 22		Quiz 5– April 22				
	Thursday – Feb 20				Thursday – April 24						
4	Tuesday – Feb 25		Quiz 2– Feb 25	13	Tuesday – April 29						
	Thursday – Feb 27				Thursday – May 1						
5	Tuesday – March 4			14	Tuesday – May 6		Quiz 6– May 6				
	Thursday – March 6				Thursday – May 8						
6	Tuesday – March 11		Quiz 3– March 13	15	Tuesday – May 13						
	Thursday – March 13				Thursday – May 15						
7	Tuesday – March 18		Review–March 18 Exam1–March 20	16	Final						
	Thursday – March 20										
8	Tuesday – March 25	Spring Break									
	Thursday – March 27										
9	Tuesday – April 1										
	Thursday – April 3										

Today's Learning Goals: Linear Regression

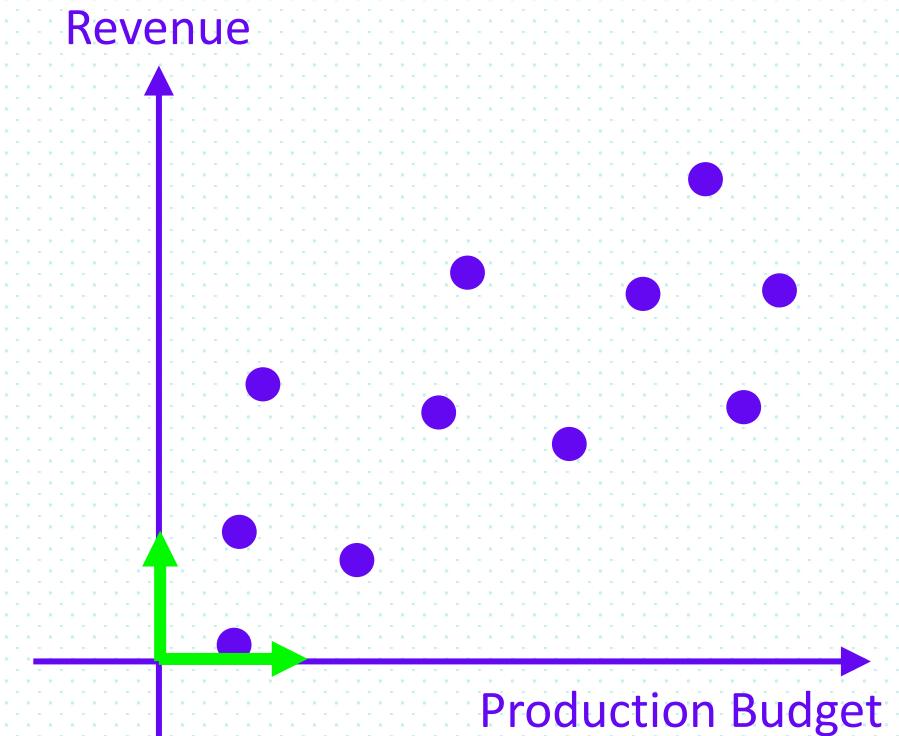
- Introduction to Linear Regression
- Simple Linear Regression
- Fitting the Linear Model
 - Parametrizing a Line
 - The Line Fitting Problem
- Model Evaluation:
 - Minimizing MSE Loss Function
 - R-Squared
- Multiple Linear Regression
- Regularization
- Assumptions in Linear Regression

Data and Python Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
```

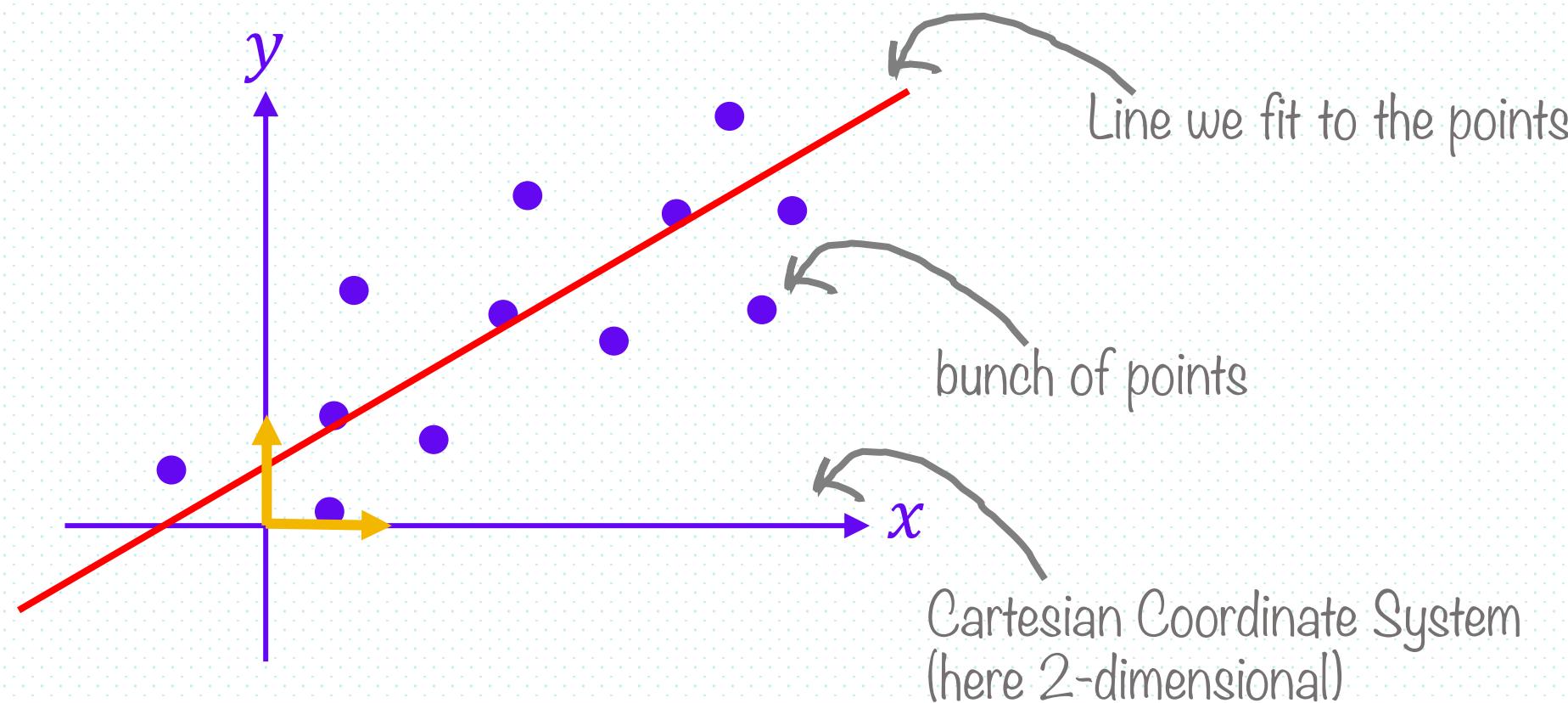
Data is from **THE NUMBERS**

Can the **Production Budget** for movies be used to predict what their **revenue** will be like?



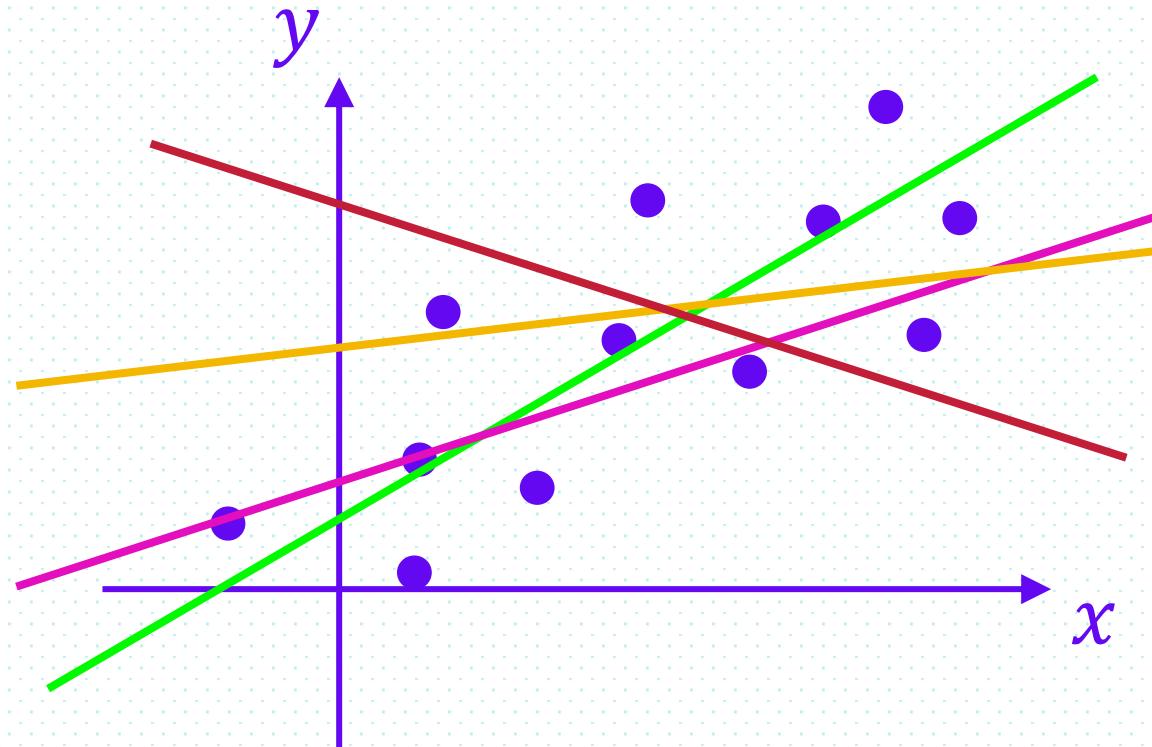
Linear Regression

Linear Regression is all about fitting a line to a bunch of points.



Linear Regression

Linear Regression is all about fitting a line to a bunch of points.



What are the x-axis and y-axis in the cartesian coordinate system?



What these points represent?



How to fit a line? And why should I care about fitting a line?



How to know the fitted line is good? And what it means to have a good line.

We want to do prediction: example (Movie Revenue)

Release Date	Movie	Worldwide Revenue
1 Dec 16, 2015	Star Wars Ep. VII: The Force Awakens	\$2,056,046,835
2 Dec 9, 2022	Avatar: The Way of Water	\$2,315,589,775
3 Jun 28, 2023	Indiana Jones and the Dial of Destiny	\$383,963,057
4 Apr 23, 2019	Avengers: Endgame	\$2,748,242,781
5 May 21, 2025	Mission: Impossible—The Final Reckoning	
6 May 20, 2011	Pirates of the Caribbean: On Stranger Tides	\$1,045,713,802
7 Apr 22, 2015	Avengers: Age of Ultron	\$1,395,316,979
8 May 17, 2023	Fast X	\$714,375,114
9 May 23, 2018	Solo: A Star Wars Story	\$393,151,347
10 Apr 25, 2018	Avengers: Infinity War	\$2,048,359,754
11 May 24, 2007	Pirates of the Caribbean: At World's End	\$960,996,492
12 Nov 13, 2017	Justice League	\$655,945,209
13 Jul 11, 2023	Mission: Impossible Dead Reckoning Part One	\$566,644,143
14 Dec 14, 2016	Rogue One: A Star Wars Story	\$1,055,083,596
15 Dec 18, 2019	Star Wars: The Rise of Skywalker	

Predict the worldwide Revenue for “Star Wars: The rise of Skywalker” movie

Without further information, we might predict the **mean** $617,650,554.57$.

What is the error of this prediction?

error = $actual\ revenue - predicted\ revenue$

What is the squared error of this prediction?

squared error = $error^2$

= $(actual\ revenue - predicted\ revenue)^2$

squared error

$$= (1,069,951,814 - 617,650,554.57)^2$$

$$= 1.8769e+16$$

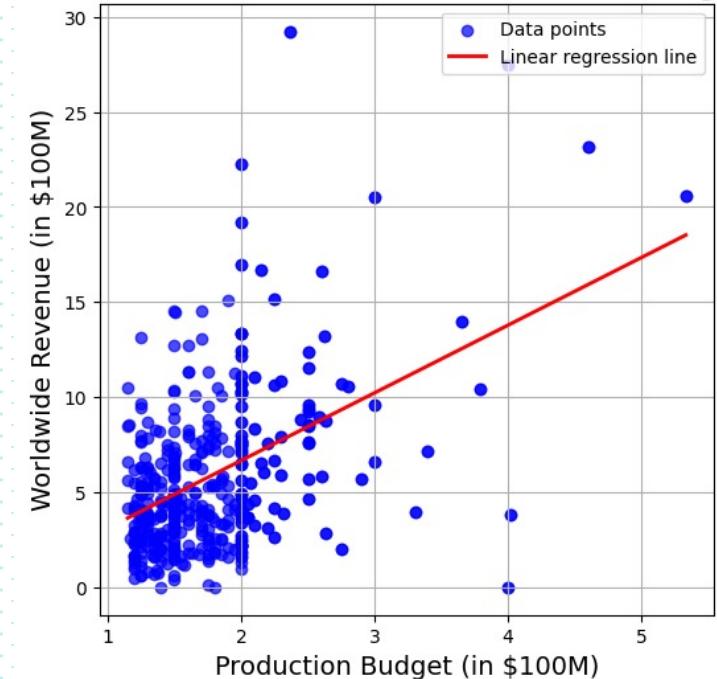
data from: <https://www.the-numbers.com/movie/budgets/all>

Better predictions with more information

Release Date	Movie	Production Budget	Worldwide Revenue
1 Dec 16, 2015	Star Wars Ep. VII: The Force Awakens	\$533,200,000	\$2,056,046,835
2 Dec 9, 2022	Avatar: The Way of Water	\$460,000,000	\$2,315,589,775
3 Jun 28, 2023	Indiana Jones and the Dial of Destiny	\$402,300,000	\$383,963,057
4 Apr 23, 2019	Avengers: Endgame	\$400,000,000	\$2,748,242,781
5 May 21, 2025	Mission: Impossible—The Final Reckoning	\$400,000,000	
6 May 20, 2011	Pirates of the Caribbean: On Stranger Tides	\$379,000,000	\$1,045,713,802
7 Apr 22, 2015	Avengers: Age of Ultron	\$365,000,000	\$1,395,316,979
8 May 17, 2023	Fast X	\$340,000,000	\$714,375,114
9 May 23, 2018	Solo: A Star Wars Story	\$330,400,000	\$393,151,347
10 Apr 25, 2018	Avengers: Infinity War	\$300,000,000	\$2,048,359,754
11 May 24, 2007	Pirates of the Caribbean: At World's End	\$300,000,000	\$960,996,492
12 Nov 13, 2017	Justice League	\$300,000,000	\$655,945,209
13 Jul 11, 2023	Mission: Impossible Dead Reckoning Part Or	\$290,000,000	\$566,644,143
14 Dec 14, 2016	Rogue One: A Star Wars Story	\$280,200,000	\$1,055,083,596
15 Dec 18, 2019	Star Wars: The Rise of Skywalker	\$275,000,000	

We also get Production Budget of movies

Worldwide Revenue vs Production Budget



This is a **regression problem** with:
Predictor variable: Production Budget
Response variable: Worldwide Revenue

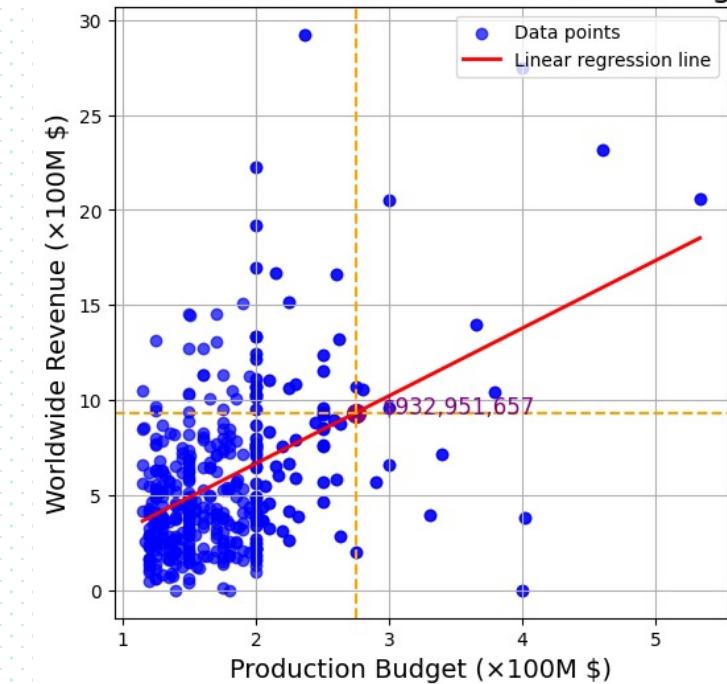
data from: <https://www.the-numbers.com/movie/budgets/all>

Better predictions with more information

Release Date	Movie	Production Budget	Worldwide Revenue
1 Dec 16, 2015	Star Wars Ep. VII: The Force Awakens	\$533,200,000	\$2,056,046,835
2 Dec 9, 2022	Avatar: The Way of Water	\$460,000,000	\$2,315,589,775
3 Jun 28, 2023	Indiana Jones and the Dial of Destiny	\$402,300,000	\$383,963,057
4 Apr 23, 2019	Avengers: Endgame	\$400,000,000	\$2,748,242,781
5 May 21, 2025	Mission: Impossible—The Final Reckoning	\$400,000,000	
6 May 20, 2011	Pirates of the Caribbean: On Stranger Tides	\$379,000,000	\$1,045,713,802
7 Apr 22, 2015	Avengers: Age of Ultron	\$365,000,000	\$1,395,316,979
8 May 17, 2023	Fast X	\$340,000,000	\$714,375,114
9 May 23, 2018	Solo: A Star Wars Story	\$330,400,000	\$393,151,347
10 Apr 25, 2018	Avengers: Infinity War	\$300,000,000	\$2,048,359,754
11 May 24, 2007	Pirates of the Caribbean: At World's End	\$300,000,000	\$960,996,492
12 Nov 13, 2017	Justice League	\$300,000,000	\$655,945,209
13 Jul 11, 2023	Mission: Impossible Dead Reckoning Part Or	\$290,000,000	\$566,644,143
14 Dec 14, 2016	Rogue One: A Star Wars Story	\$280,200,000	\$1,055,083,596
15 Dec 18, 2019	Star Wars: The Rise of Skywalker	\$275,000,000	

Prediction using linear regression

Worldwide Revenue vs Production Budget



squared error

$$\begin{aligned} &= (1,069,951,814 - 932,951,657)^2 \\ &= 1.8769043e+16 \ll 2.0457643e+17 \end{aligned}$$

Better predictions with more information



Release Date	Movie	Production Budget
1 Dec 16, 2015	Star Wars Ep. VII: The Force Awakens	\$533M
2 Dec 9, 2022	Avatar: The Way of Water	\$460M
3 Jun 28, 2023	Indiana Jones and the Dial of Destiny	\$402M
4 Apr 23, 2019	Avengers: Endgame	\$400M
5 May 21, 2025	Mission: Impossible—The Final Reckoning	\$400M
6 May 20, 2011	Pirates of the Caribbean: On Stranger Tides	\$379M
7 Apr 22, 2015	Avengers: Age of Ultron	\$365M
8 May 17, 2023	Fast X	\$340M
9 May 23, 2018	Solo: A Star Wars Story	\$330M
10 Apr 25, 2018	Avengers: Infinity War	\$300M
11 May 24, 2007	Pirates of the Caribbean: At World's End	\$300M
12 Nov 13, 2017	Justice League	\$300M
13 Jul 11, 2023	Mission: Impossible Dead Reckoning Part One	\$290M
14 Dec 14, 2016	Rogue One: A Star Wars Story	\$280M
15 Dec 18, 2019	Star Wars: The Rise of Skywalker	\$275M



Is the line above for movies with high Production Budget?

ession

on Budget

nts
egression line

•
•

—

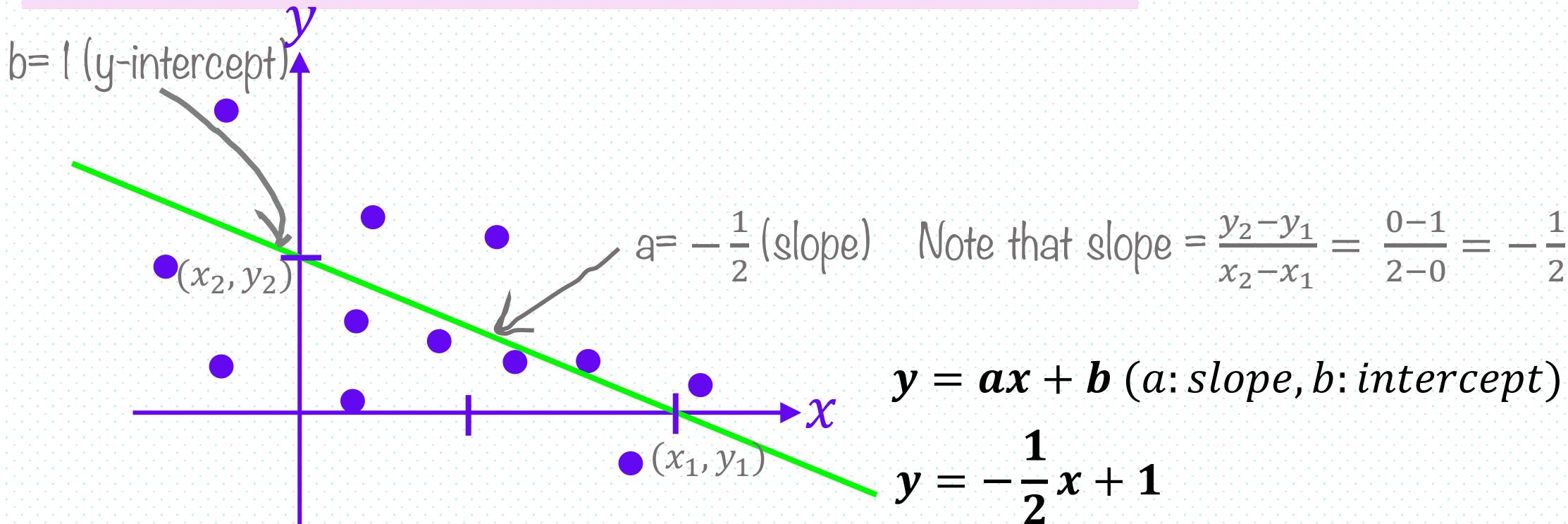
5
41)

g above line for
or movies with

Parametrizing a Line

A line can be parametrized as $y = ax + b$ (a : slope, b : intercept)

What is the corresponding equation for the following line?



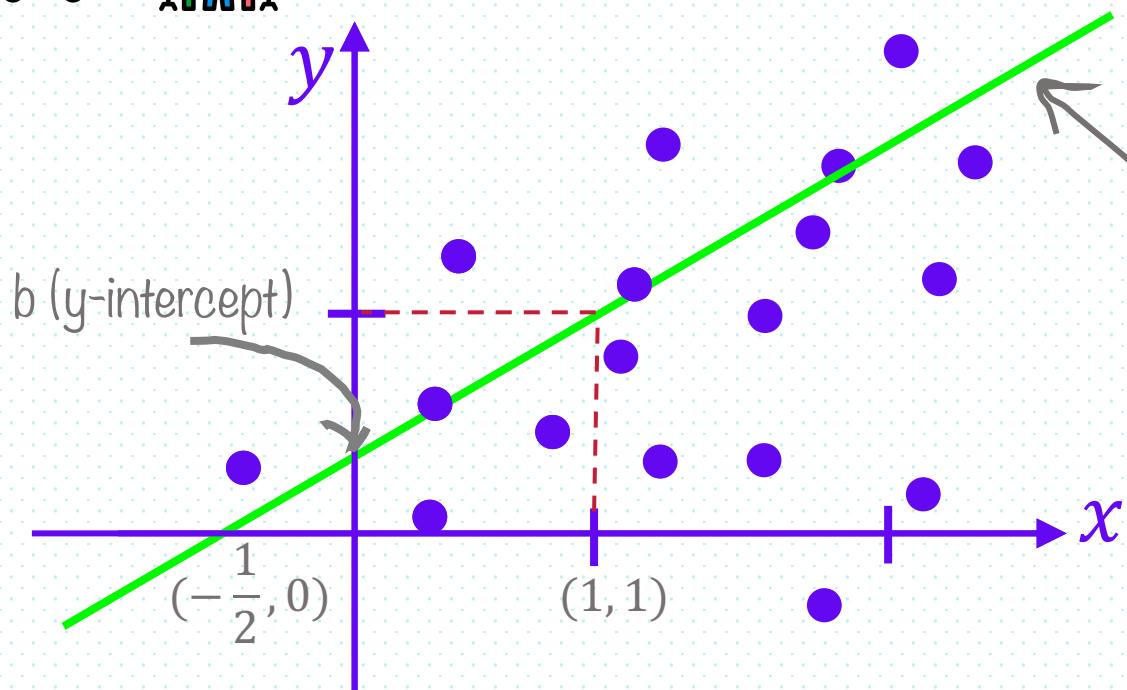


Parametrizing a Line

A line can be parametrized as $y = ax + b$ (a : slope, b : intercept)



What is the corresponding equation for the following line?



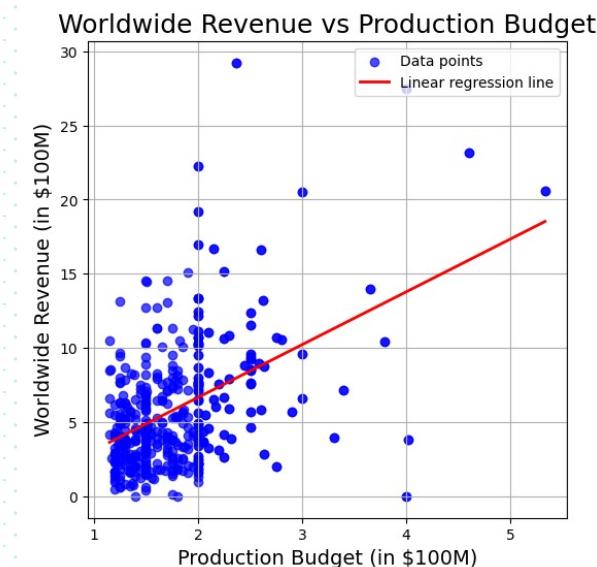
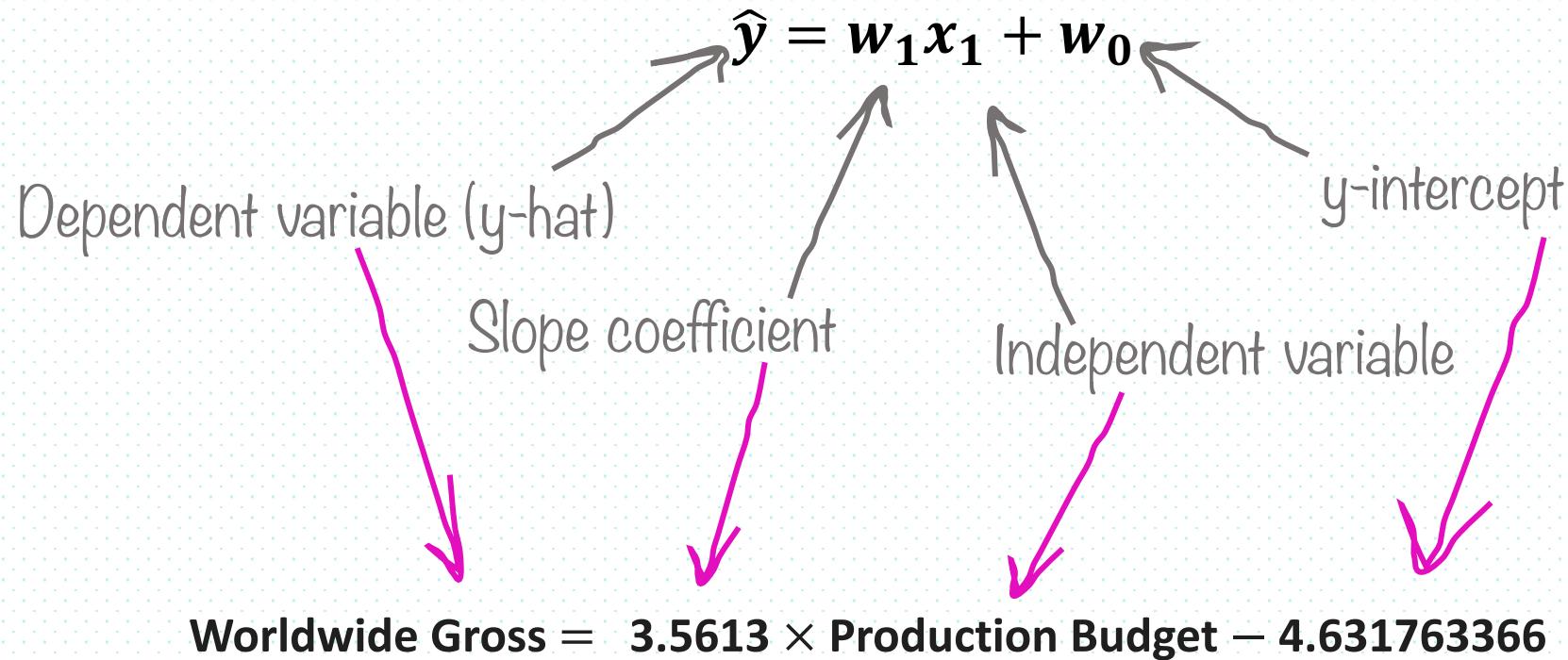
$$a = \frac{2}{3} \quad a = \text{slope} = \frac{2}{3} = \frac{y_{\text{intercept}} - 1}{0 - 1} \rightarrow b = \frac{1}{3}$$

$y = ax + b$ (a : slope, b : intercept)

$$y = \frac{2}{3}x + \frac{1}{3}$$

Simple Linear Regression

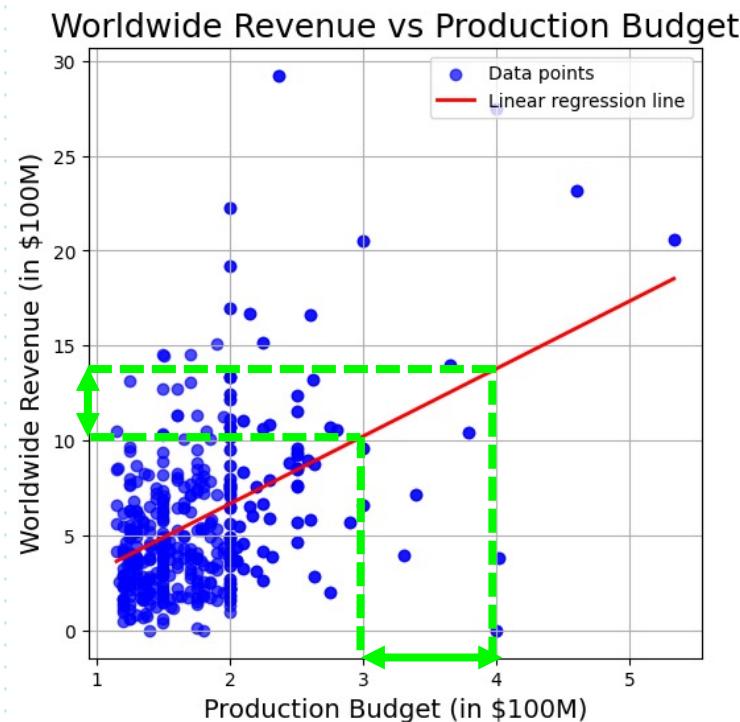
Linear Regression is all about fitting a line to a bunch of points.



Parametrizing a Line

$$\hat{y} = w_1 \times x_1 + w_0$$

Worldwide Gross = **3.5613** \times Production Budget – 4.6317633.66

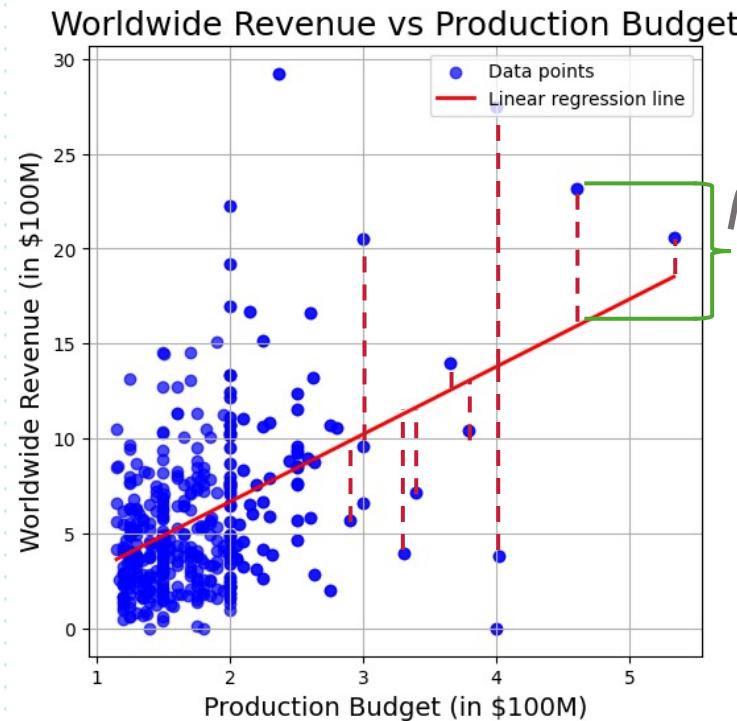


The above linear equation means that, for any **unit** increase in Production Budget, there will be **3.5613 unit** increase in Worldwide Revenue



The Line Fitting Problem

What is a good line $\hat{y} = w_1 \times x_1 + w_0$ to fit to the data?



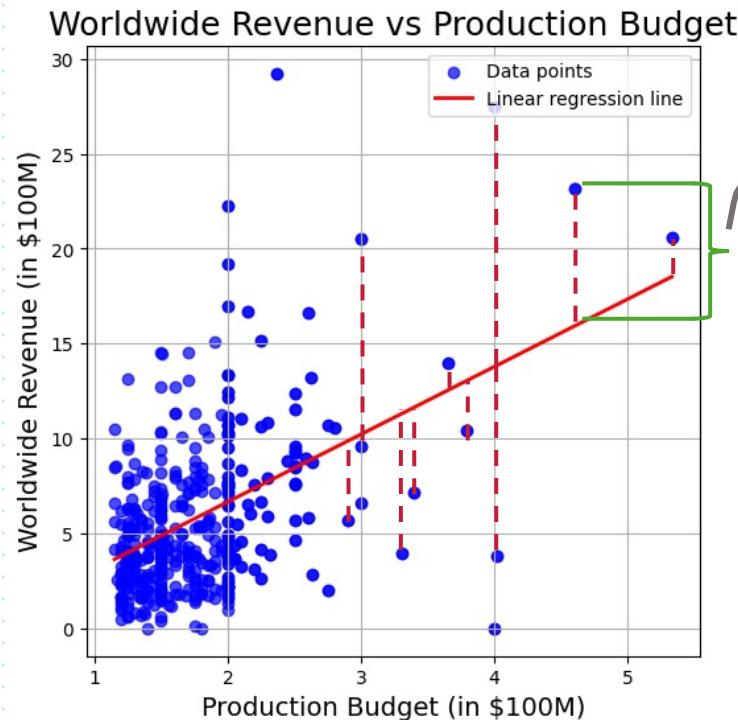
residual (error): $\epsilon_{data\ i} = y_{data\ i} - \hat{y}_{data\ i}$

We want b_0 and b_1 in $\hat{y} = w_1 \times x_1 + w_0$ such that the residual² (squared error) = $(y_{data\ i} - \hat{y}_{data\ i})^2$ is minimized.



The Line Fitting Problem

What is a good line $\hat{y} = w_1 \times x_1 + w_0$ to fit to the data?



residual (error): $\epsilon_{data\ i} = y_{data\ i} - \hat{y}_{data\ i}$

We want b_0 and b_1 in $\hat{y} = w_1 \times x_1 + w_0$ such that the residual² (squared error) = $(y_{data\ i} - \hat{y}_{data\ i})^2$ is minimized.

And we have n points, so we average residual² (squared error)s

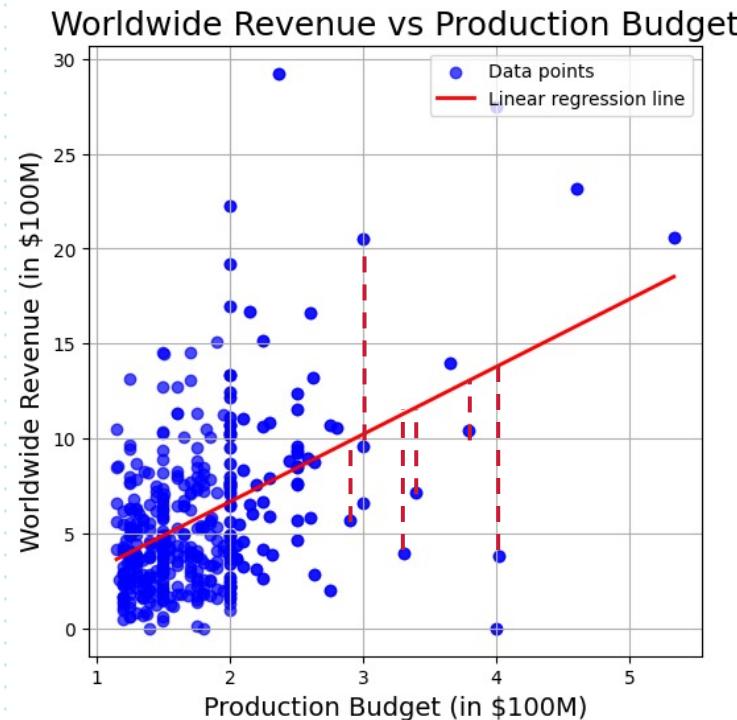
$$\text{Mean Squared Error (MSE)} = \frac{1}{n} \sum_{i=1}^n (y_{data\ i} - \hat{y}_{data\ i})^2$$



Why are we squaring residuals?

The Line Fitting Problem

What is a good line $\hat{y} = w_1 \times x_1 + w_0$ to fit to the data?



$$\text{Mean Squared Error (MSE)} = \frac{1}{n} \sum_{i=1}^n (y_{\text{data } i} - \hat{y}_{\text{data } i})^2$$

We choose MSE as our **Loss Function** and our goal is to minimize it.

Loss Function measures the difference between a model's predicted output and the actual target and evaluates how well our model fits our dataset.

Minimizing MSE Loss Function

$$\text{Mean Squared Error (MSE)} = \frac{1}{n} \sum_{i=1}^n (y_{data\ i} - \widehat{y_{data\ i}})^2$$

We choose MSE as our **Loss Function** and our goal is to minimize it.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_{data\ i} - \widehat{y_{data\ i}})^2 = \frac{1}{n} \sum_{i=1}^n (y_{data\ i} - (w_1 x_1^i + w_0))^2 \quad \text{Remember } \widehat{y} = w_1 \times x_1 + w_0$$

To minimize MSE (Loss **L**), Compute the gradient (vector of partial derivatives, set it to zero, and solve:

$$\frac{d\mathbf{L}}{dw_1} = \frac{d\mathbf{L}}{dw_0} = 0$$

Remember $\frac{du^2}{dx} = 2u \frac{du}{dx}$

Minimizing Loss Function

$$\text{Mean Squared Error (MSE)} = \frac{1}{n} \sum_{i=1}^n (y_{data\ i} - \widehat{y_{data\ i}})^2 = \frac{1}{n} \sum_{i=1}^n (y_{data\ i} - (w_1 x_1^i + w_0))^2$$

To minimize MSE (Loss \mathbf{L}), compute the gradient (vector of partial derivatives, set it to zero, and solve:

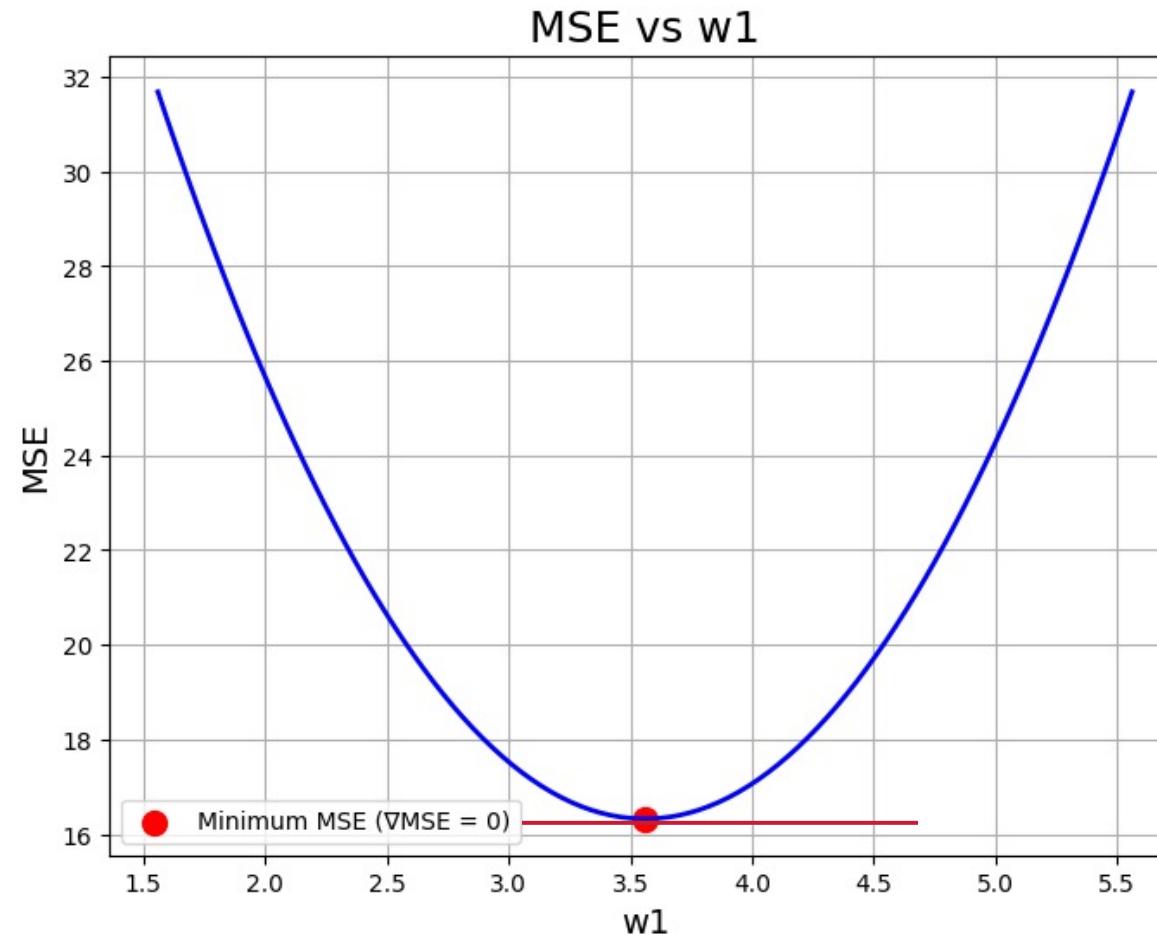
$$\frac{d\mathbf{L}}{dw_1} = \frac{d\mathbf{L}}{dw_0} = 0$$

$$\frac{d\mathbf{L}}{dw_0} = \frac{2}{n} \sum_{i=1}^n (y_i - (w_1 x_1^i + w_0)) \times (-1) = 0 \rightarrow \sum_{i=1}^n y_i = w_1 \times \sum_{i=1}^n x_i + n w_0 \rightarrow w_0 = \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} w_1 \times \sum_{i=1}^n x_i$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \rightarrow \mathbf{w_0} = \bar{y} - \mathbf{w_1} \bar{x} \quad (\text{best intercept we can get in linear regression})$$

$$\frac{d\mathbf{L}}{dw_1} = 0 \rightarrow w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (\text{best coefficient we can get in linear regression})$$

Minimizing MSE Loss Function using Gradient Descent



Example of finding the best line to fit data



What is the best line $\hat{y} = w_1 \times x_1 + w_0$ to fit to the following data?

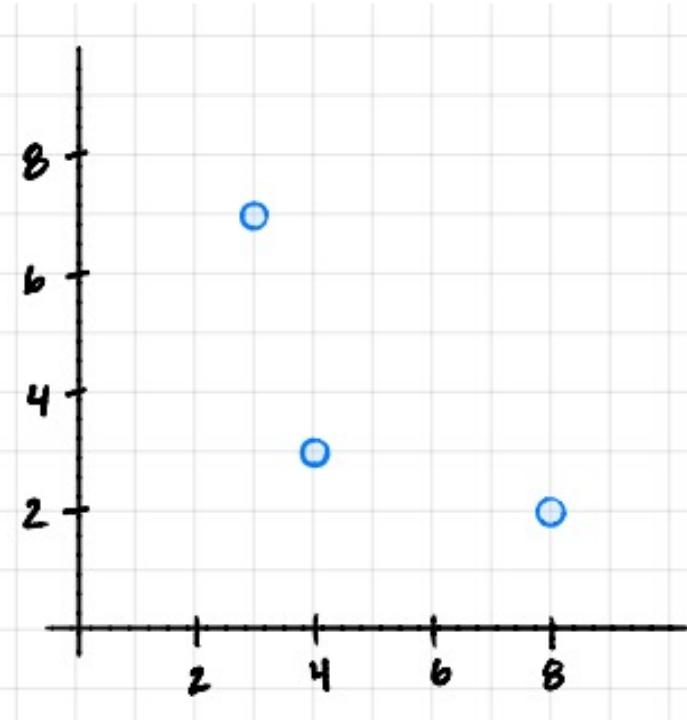
Hints: $w_0 = \bar{y} - w_1 \bar{x}$
best intercept

$$w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

best coefficient

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{3}(3 + 4 + 8) = 5$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{3}(7 + 3 + 2) = 4$$



x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
3	7	-2	3	-6	4
4	3	-1	-1	1	1
8	2	3	-2	-6	9

Example of finding the best line to fit data



What is a best line $\hat{y} = w_1 \times x_1 + w_0$ to fit to the following data?

Hints: $w_0 = \bar{y} - w_1 \bar{x}$
best intercept

$$w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

best coefficient

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{3}(3 + 4 + 8) = 5$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{3}(7 + 3 + 2) = 4$$

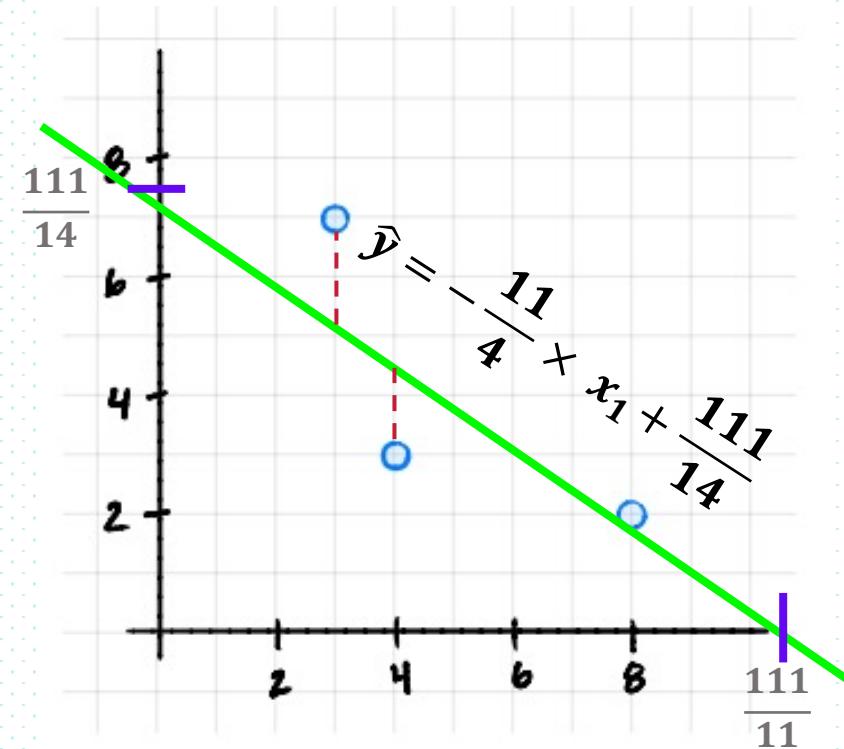
x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
3	7	(3-5)	(7-4)	(3-5)(7-4)=-6	4
4	3	(4-5)	(3-4)	(4-5)(3-4)=1	1
8	2	(8-5)	(2-4)	(8-5)(2-4)=-6	9

$$w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{(-6 + 1 - 6)}{4 + 1 + 9} = -\frac{11}{14}$$
$$w_0 = \bar{y} - w_1 \bar{x} = 4 - \left(-\frac{11}{14}\right) \times 5 = \frac{111}{14} \rightarrow \hat{y} = -\frac{11}{4} \times x_1 + \frac{111}{14}$$

Example of finding the best line to fit data



What is a best line $\hat{y} = w_1 \times x_1 + w_0$ to fit to the following data?



Let's plot $\hat{y} = -\frac{11}{4} \times x_1 + \frac{111}{14}$

Slope coefficient

y-intercept

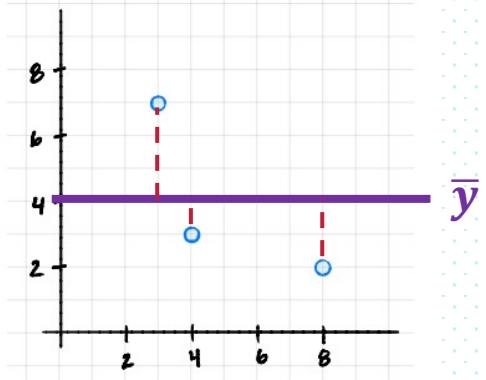
Other Metrics: R^2 (R – squared)

$$R^2(y, \hat{y}) = 1 - \frac{\text{Residual Sum of Squares}}{\text{Total Sum of Squares}}$$

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



Do we aim to have a higher R^2 or a lower?



R^2 is a normalized version of MSE , and is often easier to interpret as it does not depend on the scale of the data.

Multiple Linear Regression

Linear Regression is all about fitting a line to a bunch of points.

$$\hat{y} = w_d x_d + \dots + w_2 x_2 + w_1 x_1 + w_0$$

Dependent variable (y-hat)

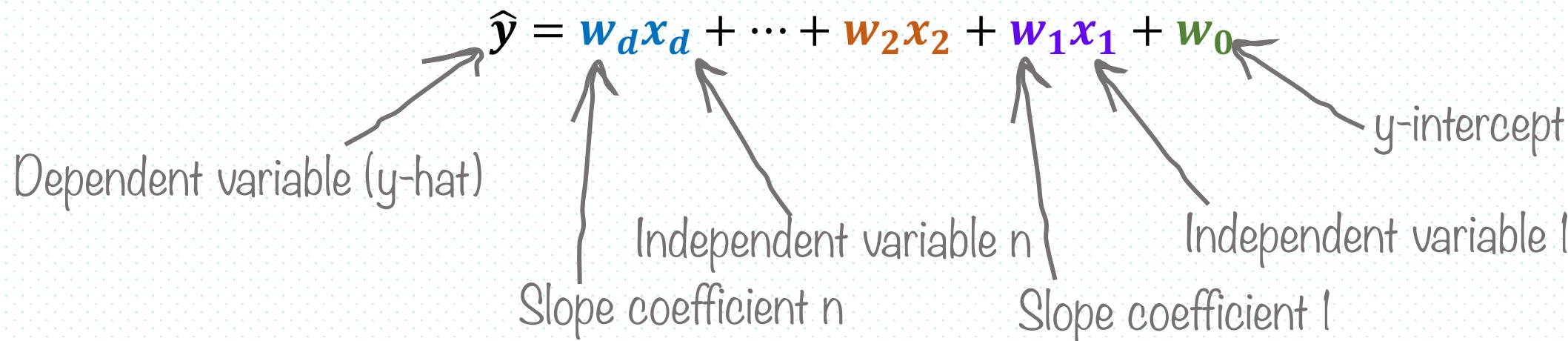
Independent variable n

Slope coefficient n

Independent variable 1

Slope coefficient 1

y-intercept



Worldwide Gross = $w_d \times$ Production Budget + $\dots + w_2 \times$ Genre + $w_1 \times$ Theater Count + w_0

Multiple Linear Regression

For each movie we have:

6 features $x = (x_1, x_2, \dots, x_6)$

Production Budget, Awards/Nominations, Theater Count, Star Power, Release Date and Genre.

$$\text{Worldwide Gross} = w_d \times \text{Production Budget} + \dots + w_2 \times \text{Genre} + w_1 \times \text{Theater Count} + w_0$$

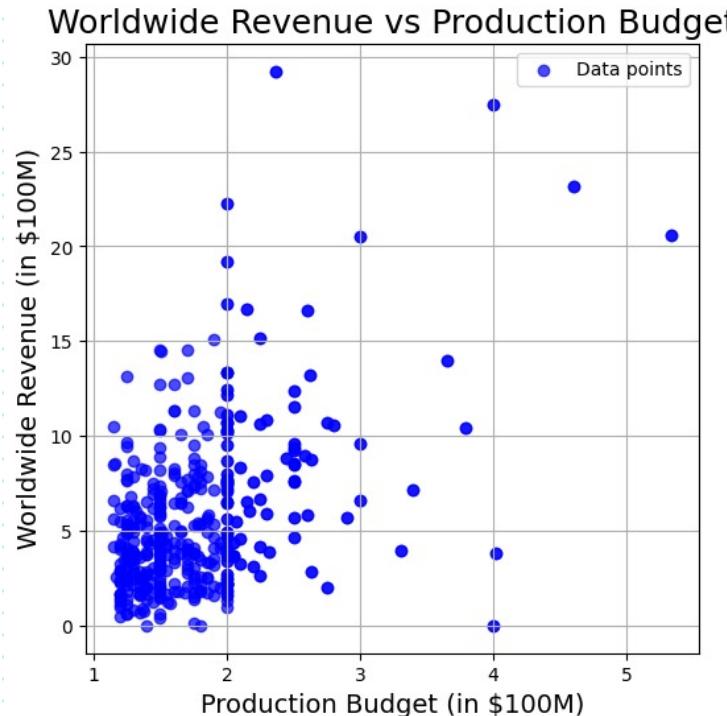
This is also a **regression problem** with:

Predictor (independent) variables: Production Budget, Awards/Nominations, Theater Count, Star Power, Release Date and Genre. $x \in \mathbb{R}^6$

Response (dependent) variable: Worldwide Revenue ($y \in \mathbb{R}$)



Back to the Movies Revenues



No predictor variables (average) $MSE = 2.0457643e+17$

One predictor variables (Production Budget) $MSE = 1.8769043e+16$

Exercise: Two predictor variables (Production Budget, Theater Count) $MSE =$



Do more features result in better prediction?

Link for using

Least-Squares Regression

Linear function of 6 variables: for $x \in \mathbb{R}^6$ where $w = (w_1, w_2, \dots, w_6)$

$$f(x) = y = w_6 \times x_6 + \cdots + w_2 \times x_2 + w_1 \times x_1 + w_0 = w \cdot x + w_0$$

Minimizing the Least Squares

$$L(w, w_0) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y^{(i)} - (w \cdot x^{(i)} + w_0))^2 = \sum_{i=1}^n (y^{(i)} - (\tilde{w} \cdot \tilde{x}^{(i)}))^2$$

$$\mathbf{w} \cdot \mathbf{x}^{(i)} + w_0 = w_0 * 1 + w_1 * x_2^i + \dots + w_d * x_d^i =$$

$$= (w_0, \dots, w_d) \cdot (1, \dots, x_d^i) = (w_0, w) \cdot (1, x) = \tilde{w} \cdot \tilde{x}^{(i)}$$

Least-Squares Regression

Goal: Finding $\tilde{w} \in R^{n+1}$ that minimizes $L(\tilde{w}) = \sum_{i=1}^n (y^{(i)} - (\tilde{w} \cdot \tilde{x}^{(i)}))^2$

$$x = \begin{bmatrix} \leftarrow \tilde{x}^{(1)} \rightarrow \\ \leftarrow \tilde{x}^{(2)} \rightarrow \\ \vdots \\ \leftarrow \tilde{x}^{(n)} \rightarrow \end{bmatrix} n * d \quad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix} n * 1$$

$$L(\tilde{w}) = \sum_{i=1}^n (y^{(i)} - (\tilde{w} \cdot \tilde{x}^{(i)}))^2 = \|y - X\tilde{w}\|^2$$

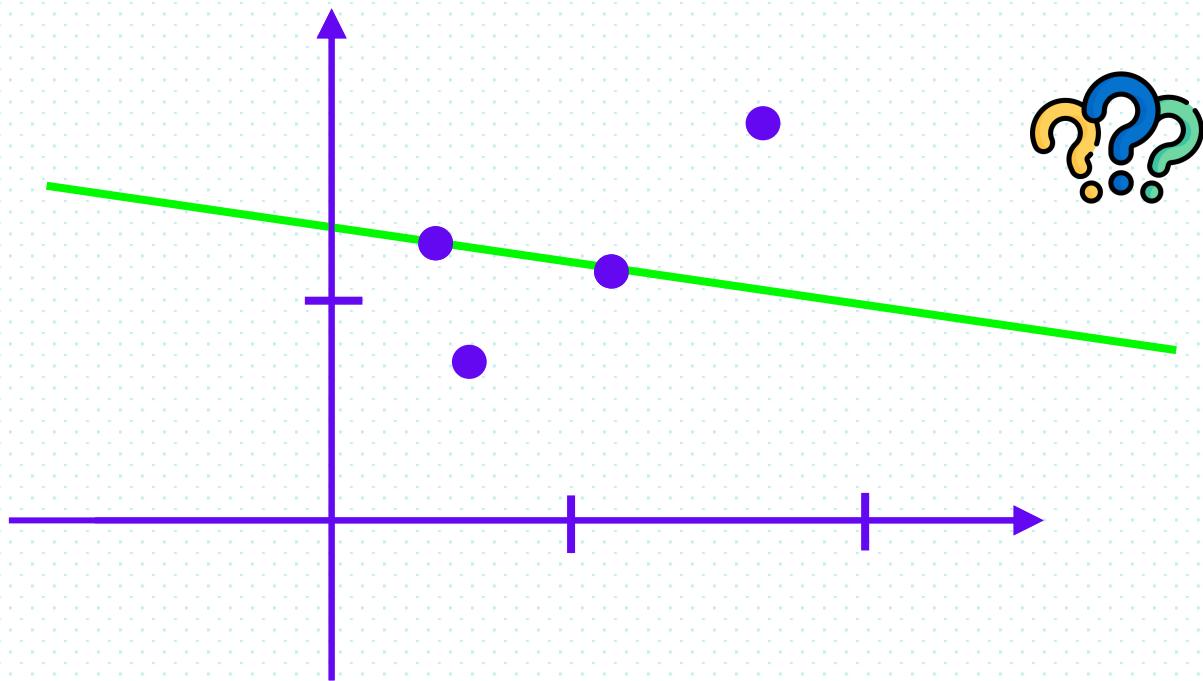
The Loss function is minimized at $\tilde{w} = (X^T X)^{-1} (X^T y)$

Least-Squares Regression

Is training loss a good estimate of future performance?

If n is large enough: maybe

Otherwise: probably an underestimate of future's error



n is small, training error is 0

If training error is not a good measure of future test error, how do we get a good estimate of future error?

Better error estimates

Recall: k-fold cross-validation

Divide the data into k equal-sized groups S_1, S_2, \dots, S_k

For $i = 1$ to k :

 Train a regressor on all data except S_i

 Let E_i be its error on S_i

Error estimate: average of E_1, E_2, \dots, E_k



If the training error is really not a very good measure of future error, why we use an optimization criterion **based exclusive on training error?**

Ridge Regression

Minimize Squared Loss **plus** a term that penalizes “complex” w :

$$L(w, w_0) = \sum_{i=1}^n (y^{(i)} - (w \cdot x^{(i)} + w_0))^2 + \lambda \|w\|^2$$



Adding a penalty term like this is called **regularization**.

Regularizer

$\lambda = 0 \rightarrow$ (Least Squares Solution)

$\lambda \rightarrow \infty \quad w \rightarrow 0$

$$w = (X^T X + \lambda I)^{-1} (X^T y)$$

Lasso Regression

Minimize Squared Loss **plus** a term that penalizes “complex” w :

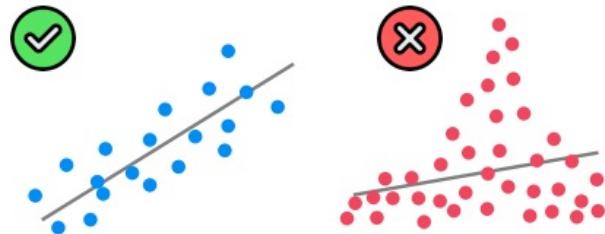
$$L(w, w_0) = \sum_{i=1}^n (y^{(i)} - (w \cdot x^{(i)} + w_0))^2 + \lambda \|w\|$$

Lasso tends to produce sparse w

Assumptions of Linear Regression (SuperDataScience)

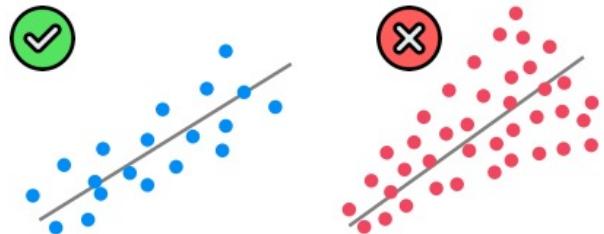
1. Linearity

(Linear relationship between Y and each X)



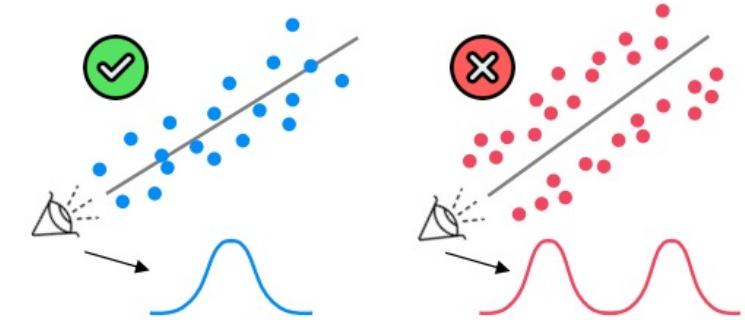
2. Homoscedasticity

(Equal variance)



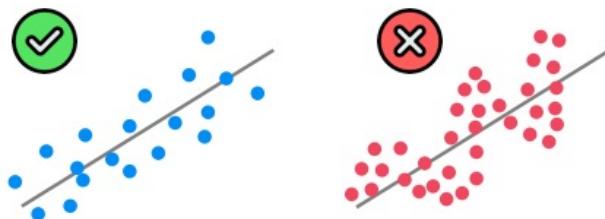
3. Multivariate Normality

(Normality of error distribution)



4. Independence

(of observations. Includes "no autocorrelation")



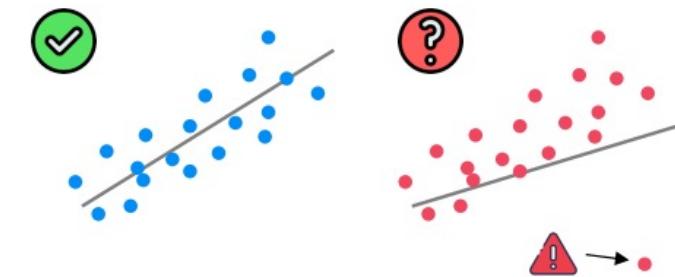
5. Lack of Multicollinearity

(Predictors are not correlated with each other)

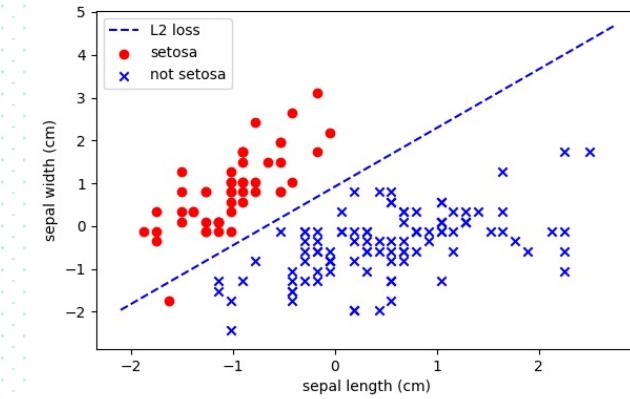
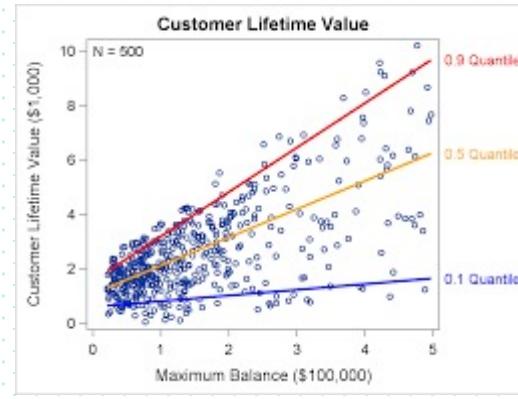
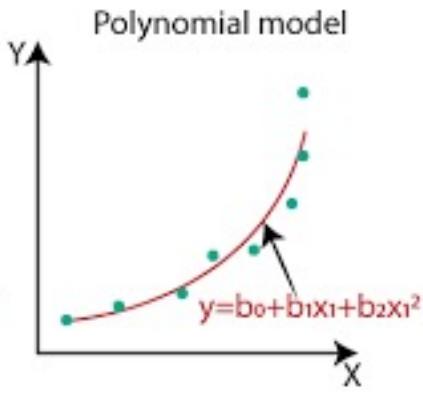
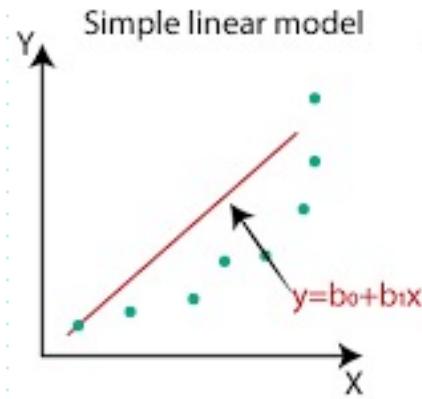


6. The Outlier Check

(This is not an assumption, but an "extra")



Going Beyond Linear Regression

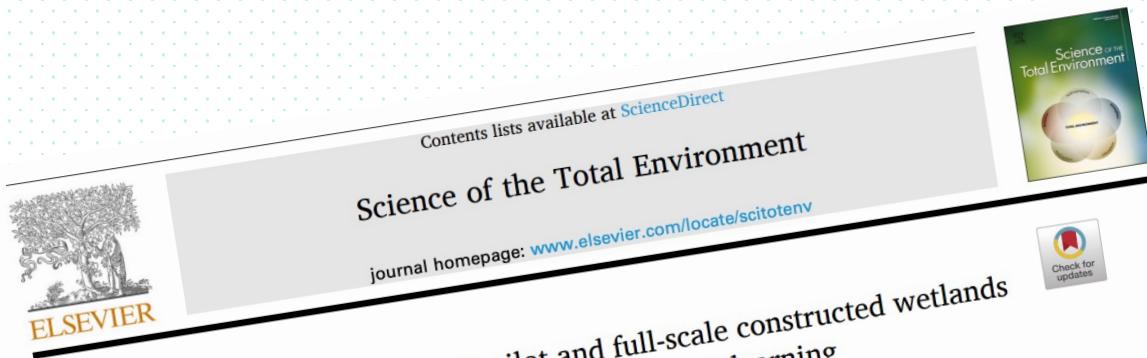


Some Variants of Linear Regression

Linear Models for classification

When the world is all about LLMs and GenAI, are we still using Linear Regression?

Yes! Linear regression is still widely used because it's interpretable, efficient, and works well for many problems where complex models are unnecessary.

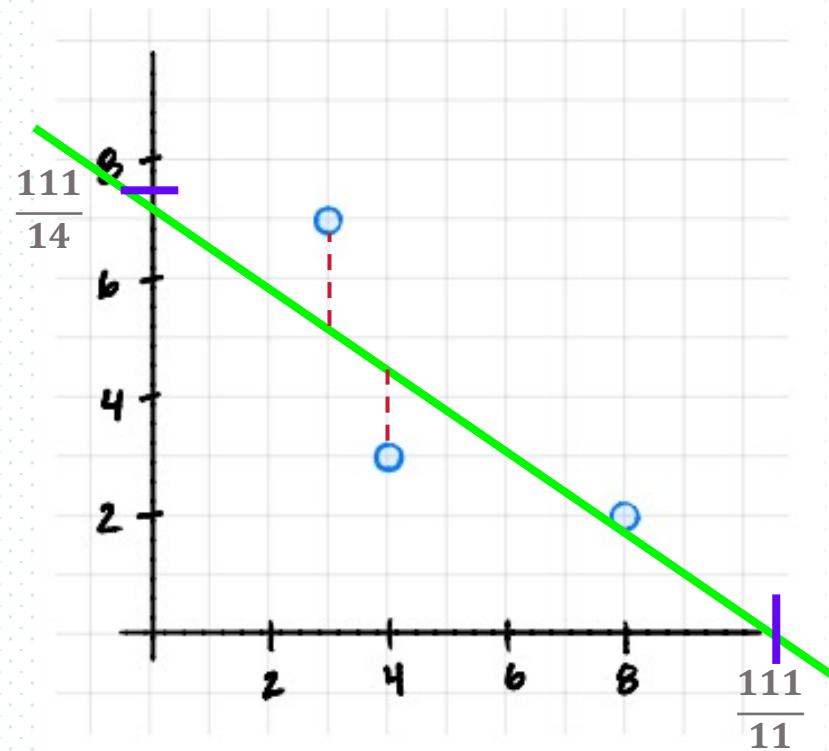


Estimating ammonium changes in pilot and full-scale constructed wetlands using kinetic model, linear regression, and machine learning
X. Cuong Nguyen^{a,b}, T. Phuong Nguyen^c, V. Son Lam^d, Phuoc-Cuong Le^e, T. Dieu Hien Vo^f,
Thu-Huong Thi Hoang^g, W. Jin Chung^h, S. Woong Chang^{h,*}, D. Duc Ng^h

Using of Ellipsoid Method and Linear Regression with L₁-Regularization for Medical Data Investigation

Petro Stetsyuk¹, Viktor Stovba¹, Ivan Senko¹, and Illya Chaikovsky¹

Exit Ticket – Due Saturday at 23:59 PM



What is the R^2 of the above fitted line?



Inspired by Resources:

Fundamentals of Machine Learning, UC San Diego, Professor Sanjoy Dasgupta

Fundamentals of Machine Learning, Sharif University of Technology, Professor Ali Sharifi Zarchi

Theoretical Foundations of Data Science, UC San Diego, Dr Albuyeh

Dataset: The-Numbers.com

Remember $w_0 = \bar{y} - w_1 \bar{x}$

Backup: Minimizing Loss Function

$$\frac{dL}{dw_1} = \frac{2}{n} \sum_{i=1}^n (y_i - (w_1 x_1^i + w_0)) \times (x_1^i) = 0 \rightarrow \sum_{i=1}^n y_i x_1^i - w_1 \times \sum_{i=1}^n (x_1^i)^2 - w_0 \times \sum_{i=1}^n x_1^i = 0$$

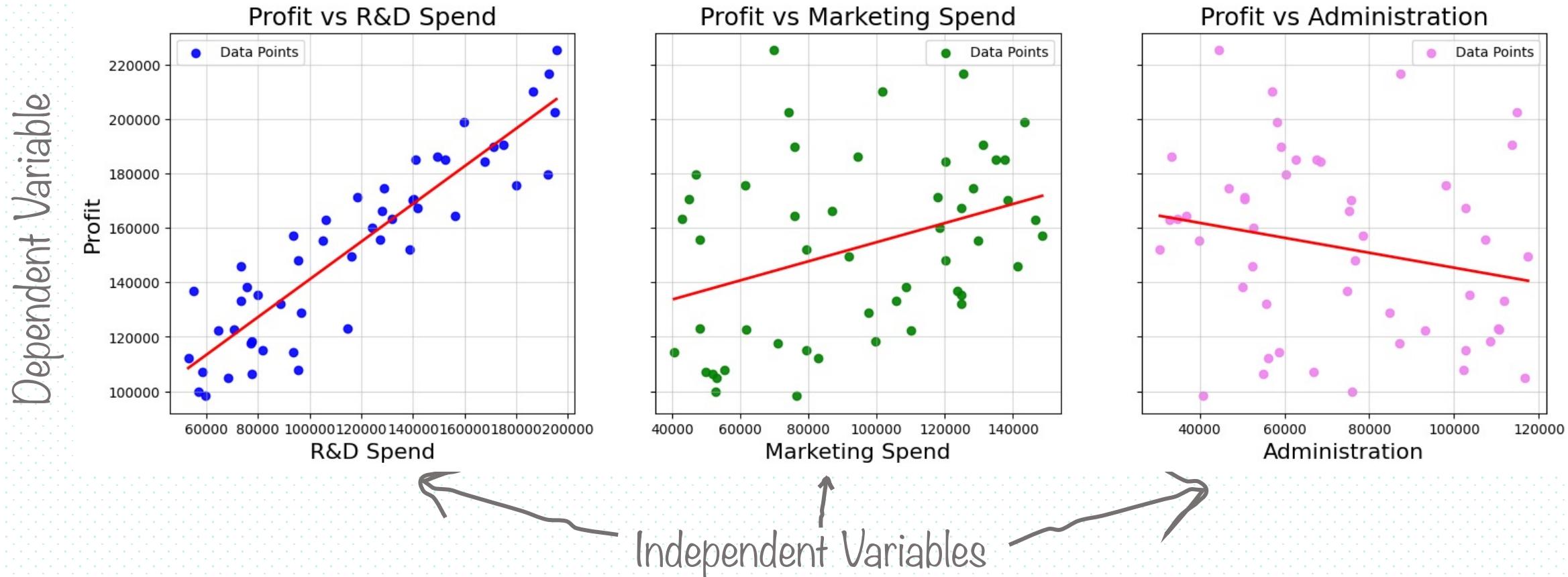
$$\sum_{i=1}^n y_i x_1^i - w_1 \times \sum_{i=1}^n (x_1^i)^2 - (\bar{y} - w_1 \bar{x}) \times \sum_{i=1}^n x_1^i = 0 \rightarrow \sum_{i=1}^n y_i x_1^i - w_1 \times \sum_{i=1}^n (x_1^i)^2 - \bar{y} \sum_{i=1}^n x_1^i + w_1 \bar{x} \sum_{i=1}^n x_1^i = 0$$

$$\sum_{i=1}^n y_i x_1^i - \bar{y} \sum_{i=1}^n x_1^i = w_1 \times \sum_{i=1}^n (x_1^i)^2 - w_1 \bar{x} \sum_{i=1}^n x_1^i \quad w_1 = \frac{\sum_{i=1}^n y_i x_1^i - \bar{y} \sum_{i=1}^n x_1^i}{\sum_{i=1}^n (x_1^i)^2 - \bar{x} \sum_{i=1}^n x_1^i}$$

$$w_1 = \frac{\sum_{i=1}^n y_i x_1^i - \bar{y} \sum_{i=1}^n x_1^i}{\sum_{i=1}^n (x_1^i)^2 - \bar{x} \sum_{i=1}^n x_1^i} = \frac{\sum_{i=1}^n y_i x_1^i - n \bar{y} \bar{x}}{\sum_{i=1}^n (x_1^i)^2 - n \bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{Cov(x, y)}{\sigma^2(x)}$$

Linear Regression: Another Example

Linear Regression is all about fitting a line to a bunch of points.



Ridge Regression

Minimize Squared Loss **plus** a term that penalizes “complex” w :

$$MSE = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - (w \cdot x^{(i)} + w_0))^2 + \lambda \|w\|$$

Adding a penalty term like this is called **regularization**.

$\lambda = 0$ (Least Squares Solution)

$\lambda \rightarrow \infty$ $w = 0$

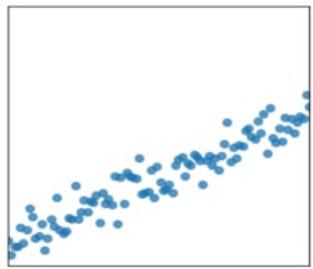
λ	training MSE	test MSE
0.00001	0.00	585.81
0.0001	0.00	564.28
0.001	0.00	404.08
0.01	0.01	83.48
0.1	0.03	19.26
1.0	0.07	7.02
10.0	0.35	2.84
100.0	2.40	5.79
1000.0	8.19	10.97
10000.0	10.83	12.63

Put predictor vectors in matrix X and responses in vector y .

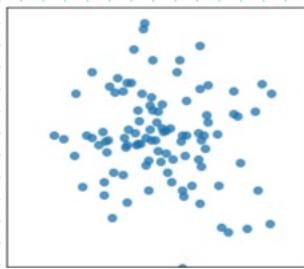
$$w = (X^T X + \lambda I)^{-1} (X^T y)$$

Pattens in Scatter Plots: Connection with Correlation

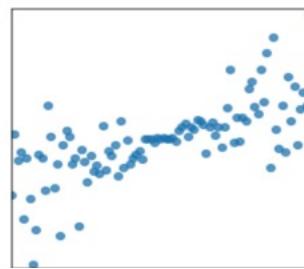
Correlation is the **linear association** of two variables x and y .



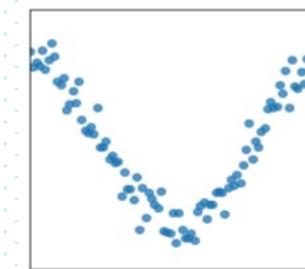
$$r = 0.949$$



$$r = -0.121$$



$$r = 0.704$$



$$r = 0.052$$

Correlation measures how tightly clustered a scatter plot is around a straight line.

Correlation ranges between -1 and 1.

Correlation is the average of the product of x and y , when both are in standard units.

$$r = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

Resources

Fundamentals of Statistics and Probability, Sharif University of Technology, Professor Ali Sharifi Zarchi

Statistics and Data Science, UC San Diego, Professor Alon Orlitsky

https://link.springer.com/content/pdf/10.1007/978-3-031-38747-0_3.pdf

eigenvector, eigenvalue

Hd why map?

absolute slide 18

slide 19, what does negative intercept mean?

some questions from solution manual

gradient descent with regression

logistic regression

plane? vs a line

why squared?

when we cannot use linear regression

what to do with outliers?

what it means to fit a line?

other metrics and how they are different?

popular variants: **Regularized Linear Regression**

<https://learning.edx.org/course/course-v1:UCSanDiegoX+DSE220x+1T2025/block>

Predicting Blood Pressure based on age, weight, and lifestyle factors. **Estimating Risk of Heart Disease** using cholesterol levels, BMI, and smoking habits. **Forecasting Diabetes Progression** based on glucose levels and insulin resistance. **Survival Prediction for Cancer Patients** based on tumor size, stage, and treatment response.

- 1.Understand computational issues related to data mining on a large scale.
- 2.Enumerate standard algorithms for supervised and unsupervised learning
- 3.Perform exploratory data analysis and visualization
- 4.Formulate techniques for outlier detection and data cleaning for domain-specific data sets