Master of Science in Business Analytics

Machine Learning I
OPAN 6602

Week 1 Live Session

Instructor:
Tommy Jones, PhD

GEORGETOWN UNIVERSITY / McDONOUGH SCHOOL of BUSINESS

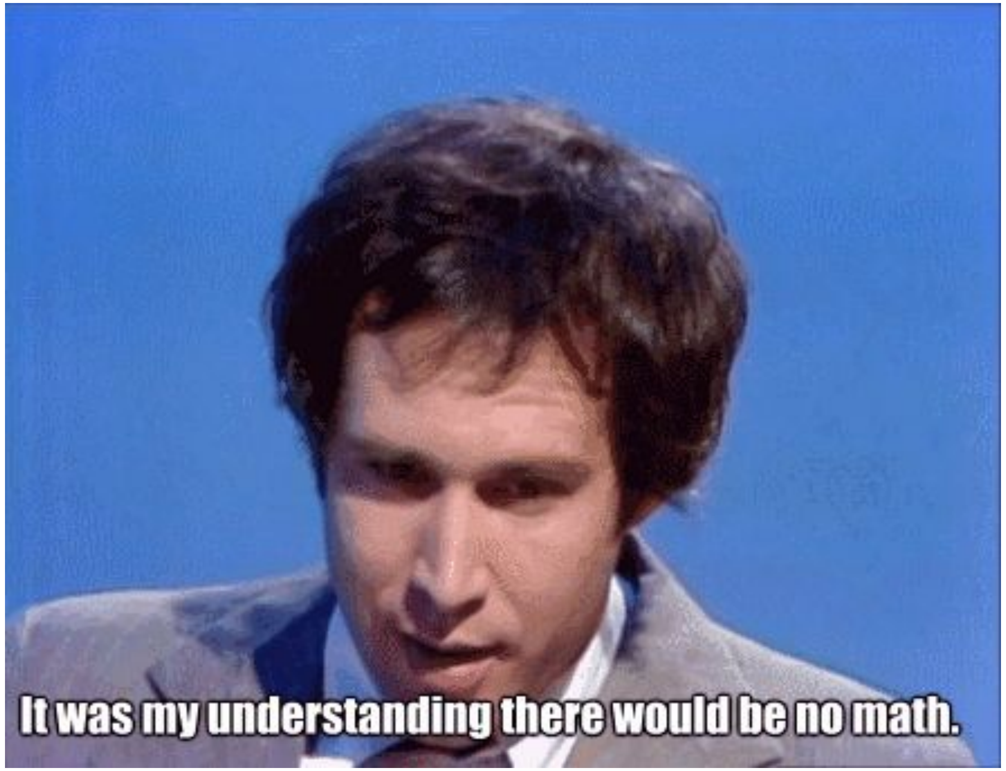
“All models are wrong. Some Models are useful.”

– George Box

“An approximate answer to the right question is worth far more than a precise answer to the wrong one.”

– John Tukey

Welcome to Foundations of Machine Learning I



Objectives:

- Know when to apply which model
- Know when a model is or is not working well
- Know how to run these models in Python

Week 1 Live Session Agenda

1. **Course Introduction**
2. Machine Learning Overview
3. Linear Regression Review
4. Coding Examples

October 30	Live Session 1: Linear Regression I
November 6	Live Session 2: Linear Regression II
November 12	*Project 1 Due*
November 13	Live Session 3: Logistic Regression
November 20	Live Session 4: Model Validation & Feature Engineering
December 3	*Project 2 Due*
December 4	Live Session 5: Biased Regression Models
December 11	Live Session 6: Decision Trees
December 18	*Final Project Due*

Course Introduction

- Machine Learning Concepts
  - Focus on supervised learning
- Tools: Python, Google Colab
- Assignments and Grades
  - Project 1: Multiple Linear Regression (Individual) 30%
  - Project 2: Logistic Regression (Individual) 30%
  - Project 3: (Group) 35% (25% report, 10% peer-evaluation)
  - Class participation (Individual) 5%

Office Hours

Me:

Mondays 4:30 - 6:00 PM Eastern, by appointment.
https://calendar.app.google/YGm3wdBi51mRBka5A

Manav (TA):

Sundays 1 - 2:30pm ET
Tuesdays 7:30 - 9pm ET
https://georgetown.zoom.us/my/manavarora

Tour of Course Canvas

Week 1 Live Session Agenda

1. Course Introduction
2. **Machine Learning Overview**
3. Linear Regression Review
4. Coding Examples

Defining Machine Learning

Machine Learning (from getting started video)
"A subfield of Artificial Intelligence where algorithms 'learn' patterns in data to solve problems."

Statistical Learning (from ISL pp. 16-17)
"Statistical Learning [is] a set of approaches for estimating f" where f defines the relationship between variable(s), Y, and variable(s), X, with some error, ϵ as in the equation below.

$$Y = f(X) + \epsilon$$

Types of Machine Learning

Supervised
We want to learn patterns in the data **based on a known outcome** we are trying to predict.

Unsupervised
We want to learn patterns in the data **without or independent of a known outcome** we are trying to predict.

Reinforcement
Used primarily in robotics. An algorithm interacts with its environment within a structure of rewards and penalties.

Comprehension Check:
Is this supervised or unsupervised ML?

Competition: October 2006.

Using a training data set of 400,000 Netflix customers' ratings for 18,000 movies, participants were asked to build a model predicting customer ratings.

The test set was a set of 1 million customer-movie pairs that are missing in the training data,

Ratings ranged between 1 (terrible) and 5 (amazing).

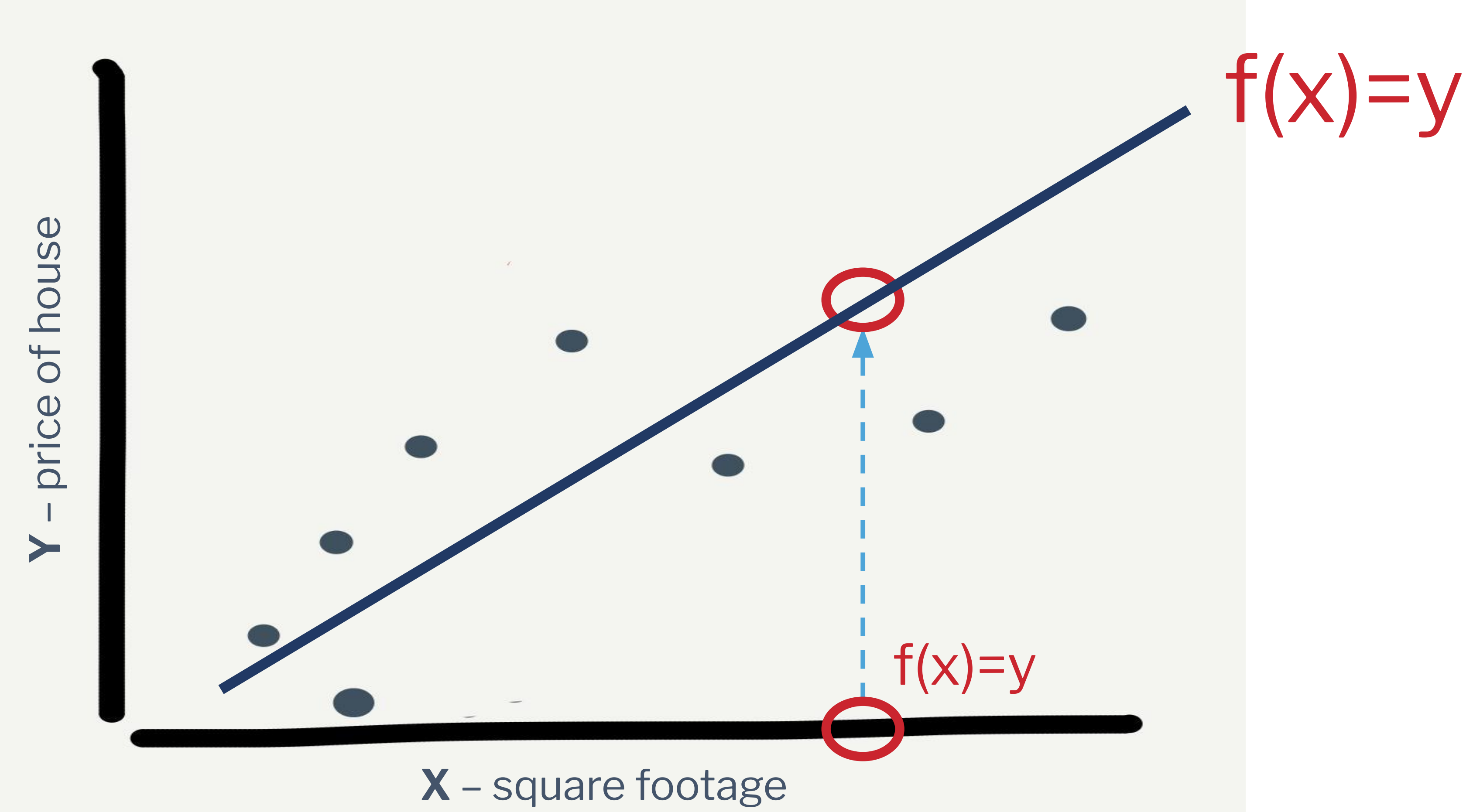
Netflix's original algorithm achieved a RMSE of 0.953. The first team to achieve a 10% improvement to Netflix's algorithm wins one million dollars.

Comprehension Check:
Is this supervised or unsupervised ML?

Unsupervised ML

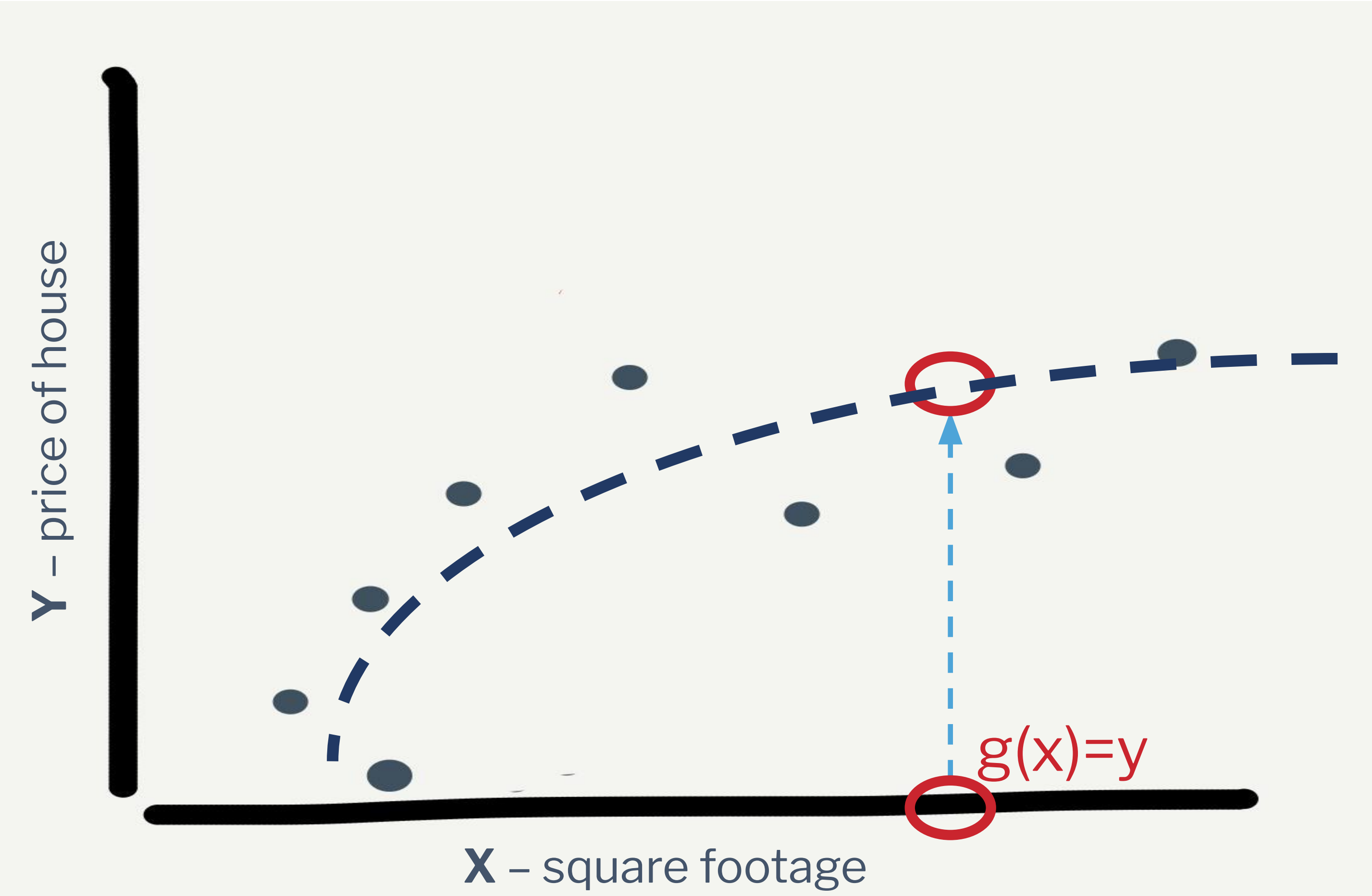
Supervised ML

Supervised Learning: Numeric Prediction (Regression)



X	Y
850	650,220
1,300	753,475
2,400	999,876
…	…

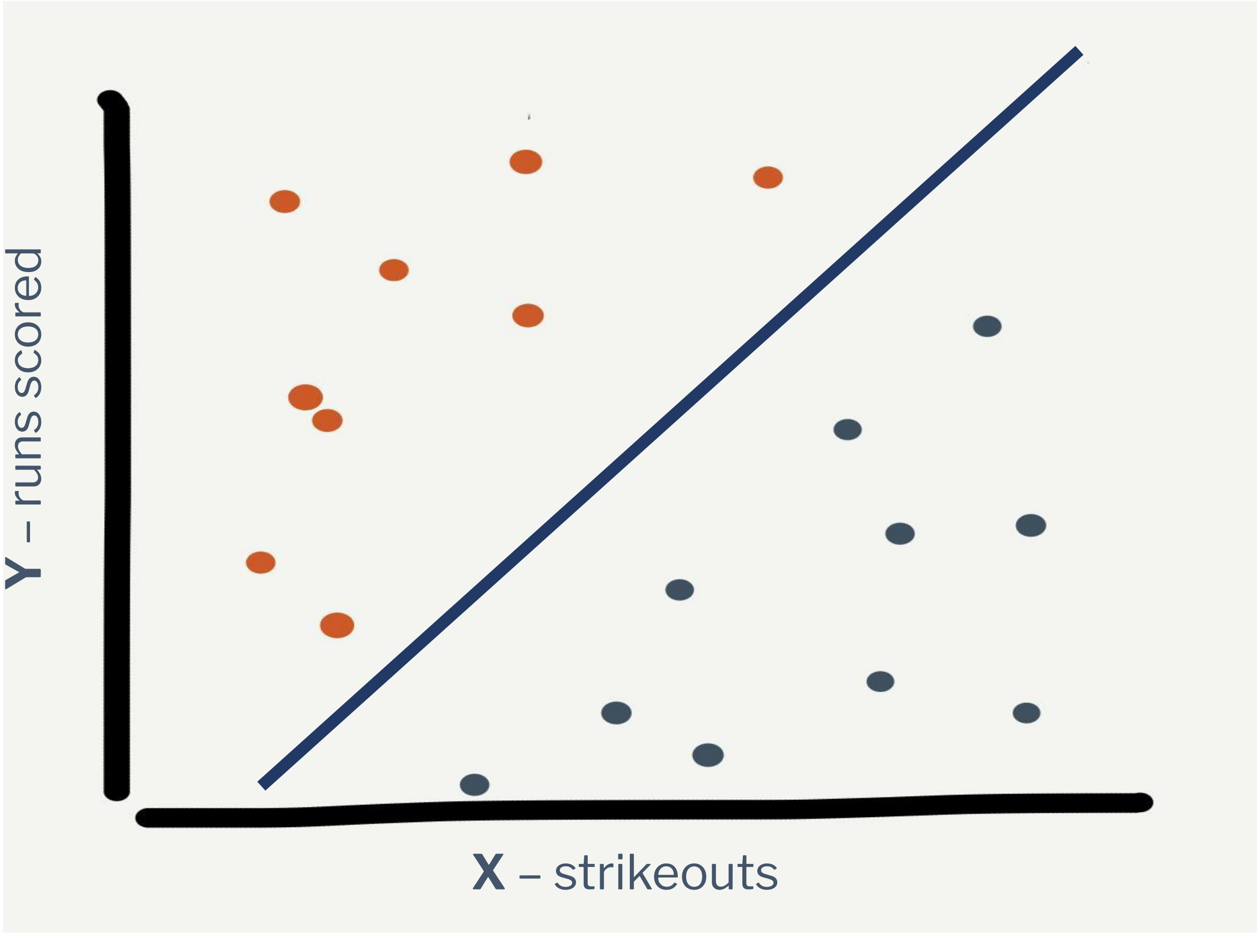
Supervised Learning: Numeric Prediction (Regression)



g(x)=y

X	Y
850	650,220
1,300	753,475
2,400	999,876
…	…

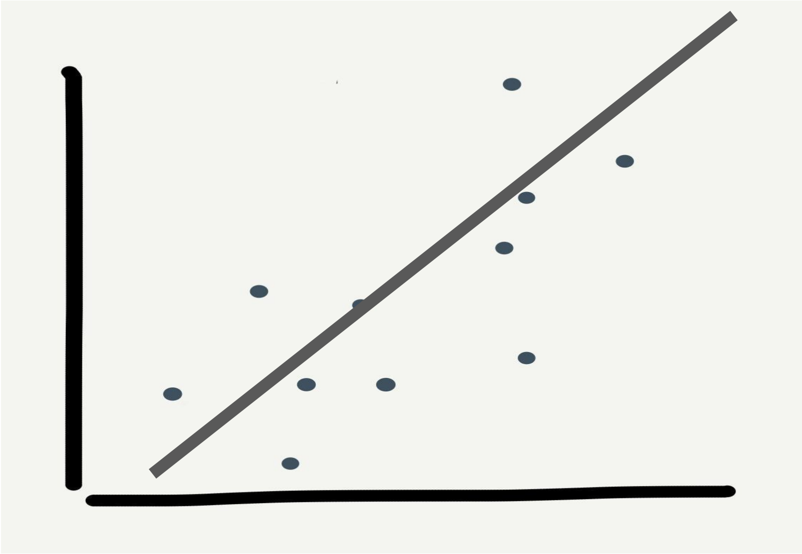
Supervised Learning: Classification



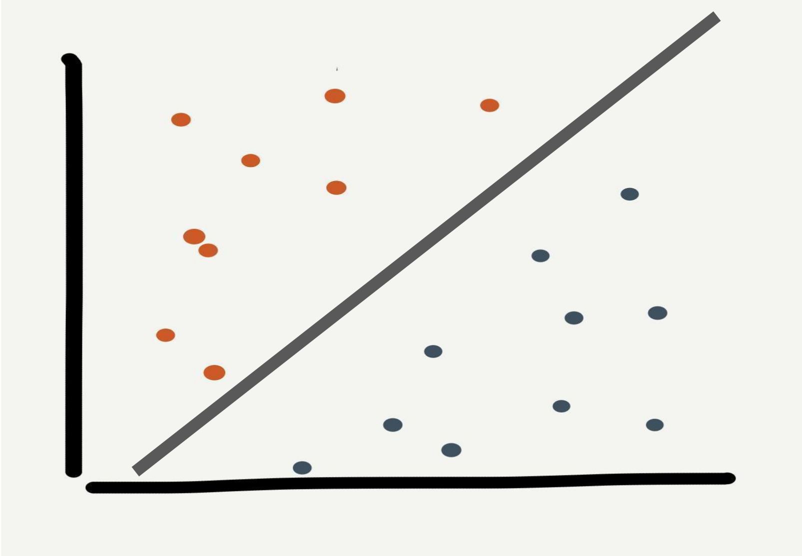
Win
Loss

X	Y	Class
0	5	Win
7	3	Loss
5	10	Win
…	…	…

Numeric Prediction



Classification



Unsupervised ML & Reinforcement Learning

Unsupervised

- Clustering (including embedding)
- Association Rules
- Anomaly Detection

Reinforcement

- An "agent" (the model) interacts with an environment and receives feedback in the form of rewards or penalties

Covered in later courses

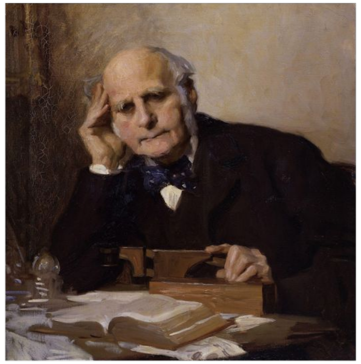
Week 1 Live Session Agenda

1. Course Introduction
2. Machine Learning Overview
3. **Linear Regression Review**
4. Coding Examples

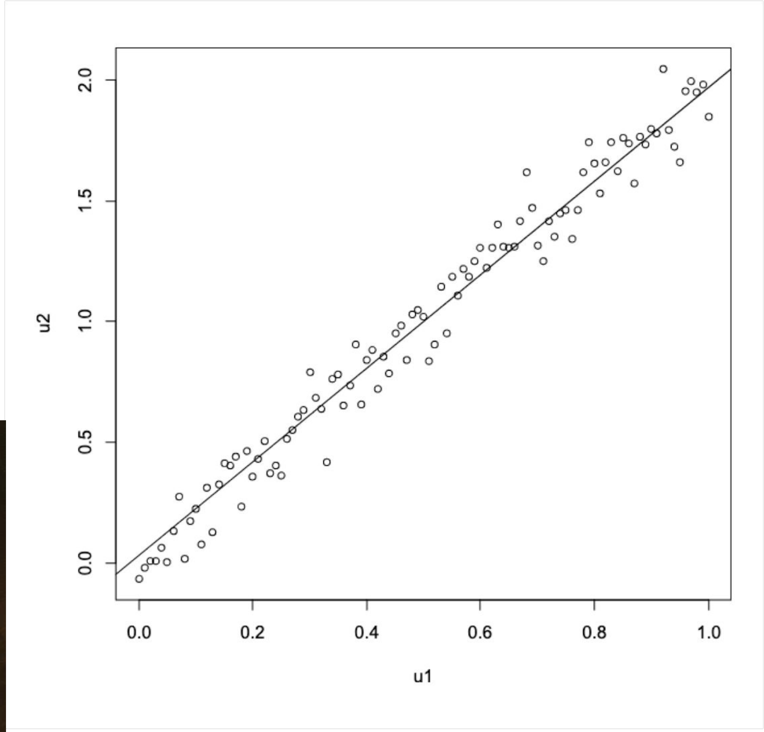
Linear Regression (OLS) - The Oldest ML Model?

By Christian Albrecht Jensen - Public Domain

By Julien-Léopold Boilly- Public Domain



By Charles Wellington Furse - Public Domain



Comprehension Check:
Which of the following is *not* an assumption of OLS?

1. Errors are distributed normally.

2. The relationship between Y and X is linear in the parameters.

3. The data set is representative of the population.

4. The errors have a mean of zero.

5. The errors have constant variance.

Comprehension Check:
Which of the following is *not* an assumption of OLS?

1. Errors are distributed normally.

2. The relationship between Y and X is linear in the parameters.

3. **The data set is representative of the population.**

4. The errors have a mean of zero.

5. The errors have constant variance.

Linear Regression
AKA Ordinary Least Squares (OLS)

For linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

1. The relationship between *Y* and the *X*'s are linear in its parameters.
2. $\epsilon \sim N(0, \sigma)$*
3. σ is constant (homoskedasticity).*
4. Errors are i.i.d (implicitly, x_i and x_j are independent $\forall\ i, j$)*

If you ***only*** want to ***predict***, you only need this.

If you also want to ***explain***, you need these.

Prediction vs. Explanation

If prediction is your goal:

Only “linear in parameters” assumption really matters.
You can (mostly) go wild with transformations.
Still watch out for structural breaks (data drift).

If explanation is your goal:

Statistical inference really matters → all assumptions apply.
Be careful with transformations that obscure explanation.
Sometimes it’s ok to trade good predictions for simplicity.

You can capture non linear relationships with OLS.

"Linear *in the parameters*" is key. Consider…

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

There is non-linear absolute relationship between *Y* and *X*.

But by including the quadratic of *X*, the relationship is *linear in the parameters*.

Marginal Effect

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

"A one unit change in X_1 leads to a β_1 unit change in Y."

Units of X and Y matter in interpretation, not mathematically.

e.g., if Y is in millions of dollars and X is in cents…Eek.

Working with transformed variables means you have to reverse transform to get marginal effects in human-interpretable units.

Marginal Effect → Take the Derivative

Things are more complicated with non-linear relationships.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon \qquad \frac{dY}{dX} = \beta_1 + 2\beta_2 X$$

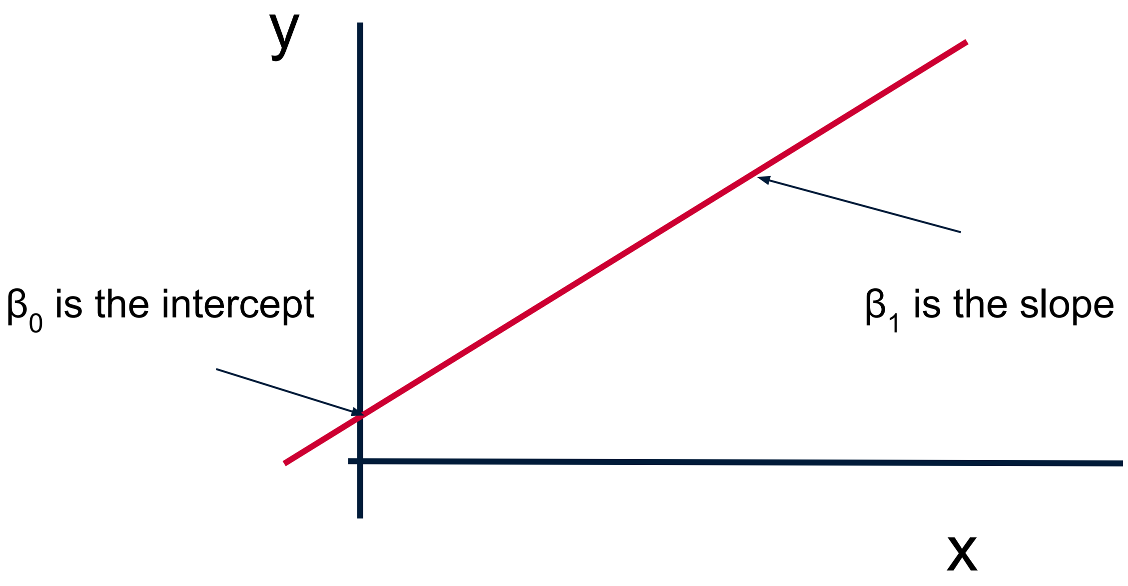
In this case, you can report the marginal effect different ways:

1. At a specific point (e.g, at the mean or median of *X*).
2. Averaged across the range of *X* in the training set.

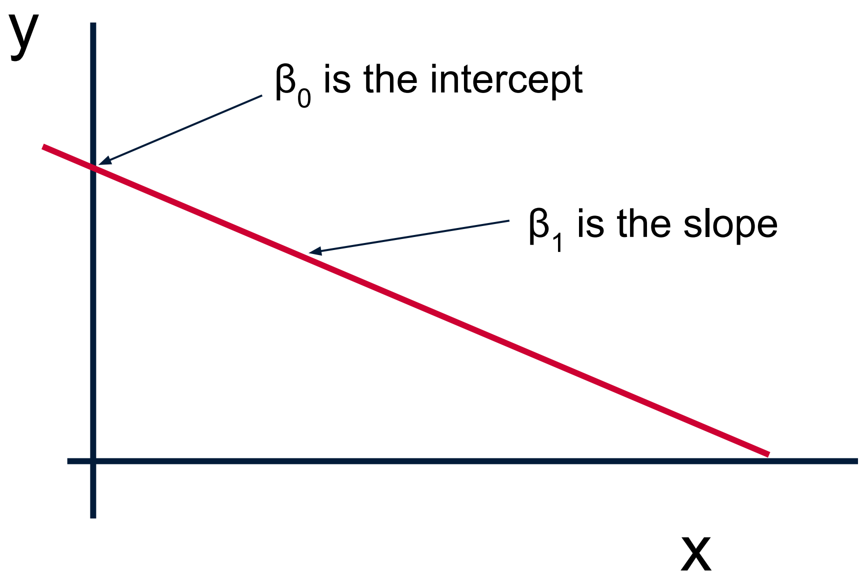
We have to do similar for logistic regression (week 3).

Marginal Effects: Linear Parameters & "Constant Returns"

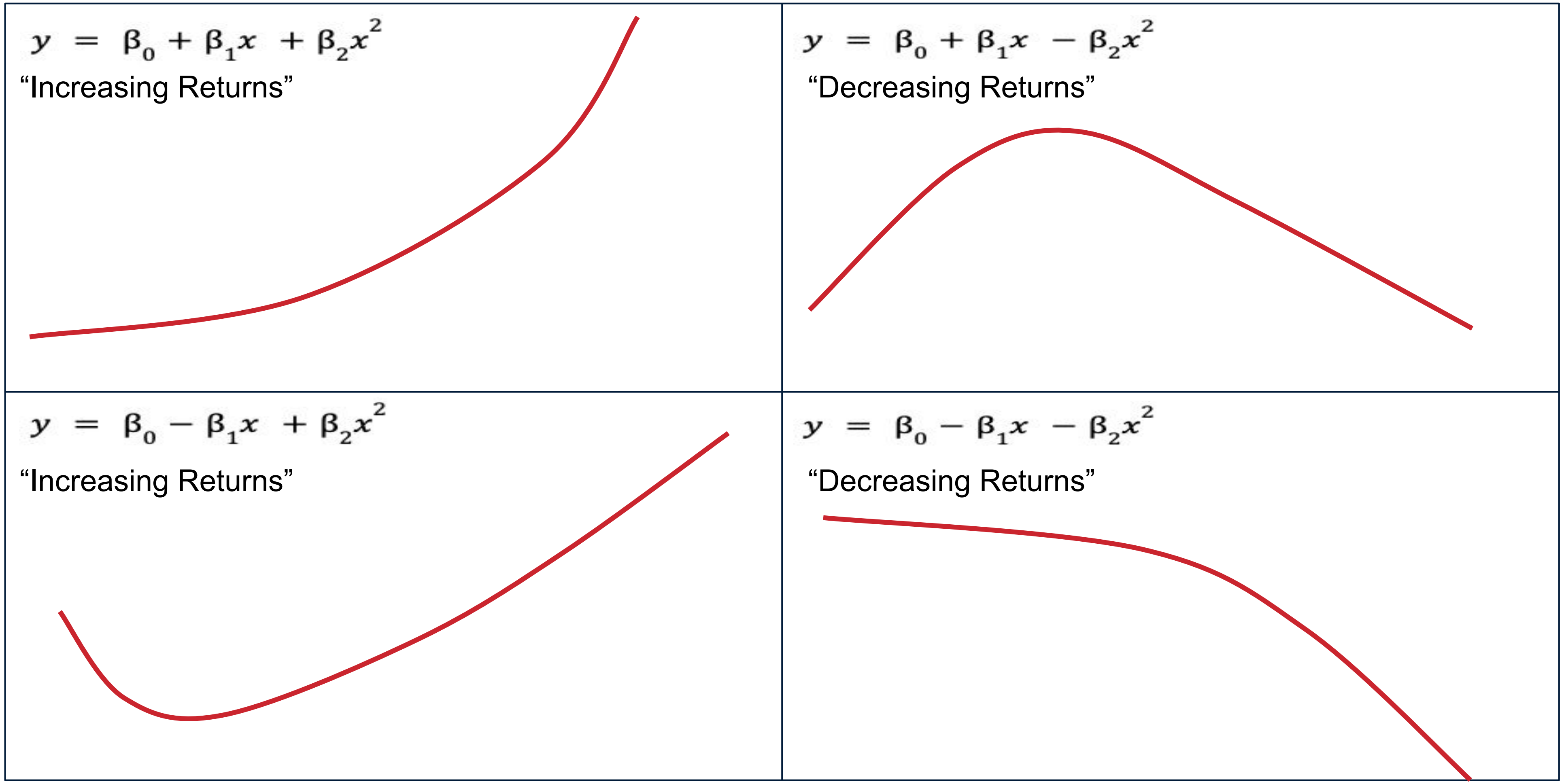
$y = \beta_0 + \beta_1 x$



$y = \beta_0 - \beta_1 x$



Marginal Effects: Using Quadratics



Coefficients in same direction → Acceleration (hockey stick)

Coefficients in opposite direction → Deceleration and Critical points

Critical points are where slope changes direction

A Marginal Effects Helper Table

Transform	Model	Marginal Effect of x on y
None	$y = \beta_0 \pm \beta_1 x$	$\pm \beta_1$
Square/quadratic	$y = \beta_0 \pm \beta_1 \pm \beta_2 x^2$	$\pm \beta_1 \pm 2\beta_2 x$
Square root	$y = \beta_0 \pm \beta_1 x \pm \beta_2 \sqrt{x}$	$\pm\beta_1 \pm \frac{1}{2}\beta_2 \frac{1}{\sqrt{x}}$
Cube root	$y = \beta_0 \pm \beta_1 x \pm \beta_2 x^{1/3}$	$\pm\beta_1 \pm \frac{1}{3}\beta_2 \frac{1}{x^{1/3}}$
Standardization	$y = \beta_0 \pm \beta_1 x \pm \beta_2 \frac{x - \mu_x}{\sigma_x}$	$\pm\beta_1 \pm \frac{\beta_2}{\sigma_x}$
Normalization/MinMax	$y = \beta_0 \pm \beta_1 x \pm \beta_2 \frac{x - min_x}{max_x - min_x}$	$\pm\beta_1 \pm \frac{\beta_2}{max_x - min_x}$

Focus just on these for now.

Comprehension Check:
Which of the following is true?

1. Backward elimination begins with no X variables—an empty model.

2. "Stepwise selection" is a combination of both forward and backward selection techniques.

3. P-values are not typically used to test the significance at each step.

4. Forward selection starts with all variables in the model and deletes the worst variables one at a time.

Comprehension Check:
Which of the following is true?

1. Backward elimination begins with no X variables—an empty model.

2. **"Stepwise selection" is a combination of both forward and backward selection techniques.**

3. P-values are not typically used to test the significance at each step.

4. Forward selection starts with all variables in the model and deletes the worst variables one at a time.

Automated Model Selection

Forward Selection, Backward Selection, Stepwise Selection

Be careful if explanation is your goal. In all cases:
- Biases in coefficients, predictions, standard errors
- Biases degrees of freedom of model (multiple testing)
- Biases p-values towards significance

If only prediction is your goal, you can mostly go wild..

Goodness of Fit Metrics: Numeric Prediction

MSE (minimize) - squared units of Y -

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X) - Y)^2$$

RMSE (minimize) - units of Y -

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(X) - Y)^2}$$

MAE (minimize) - units of Y -

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{f}(X) - Y|$$

R-squared (maximize) - prop. of var. -

$$1 - \frac{\sum_{i=1}^{n}(Y - \hat{f}(X))^2}{\sum_{i=1}^{n}(Y - E[Y])^2}$$

Adjusted R^2

All else constant, R^2 increases by adding predictor variables.

This can lead to overly-complex models and over fit.

Adjusted R^2 adds a penalizing factor for the number of predictor variables.

$$R^2_{adj.} = 1 - \frac{(1 - R^2)(n-1)}{n-p-1}$$

Modeling Process

Generalizes to all supervised learning methods.

1. Data pre-processing
2. Partition: train/test
3. Data exploration
4. Feature engineering
5. Feature & model selection
6. Model evaluation
7. Predictions & conclusions

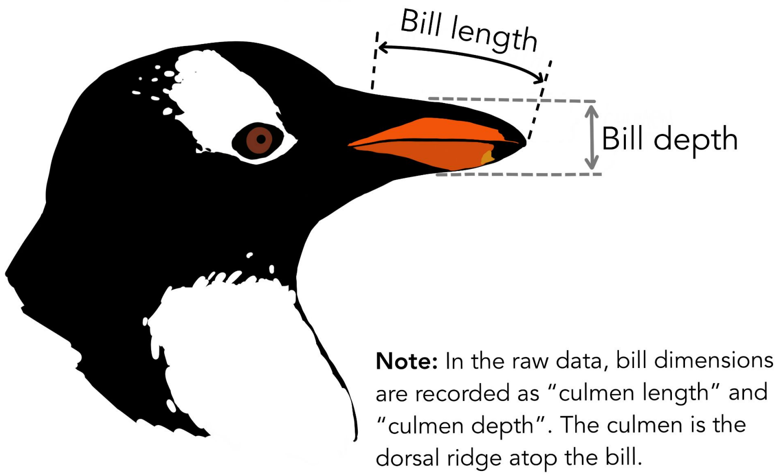
Week 1 Live Session Agenda

1. Course Introduction
2. Machine Learning Overview
3. Linear Regression Review
4. **Coding Examples**

Code: Introducing the Palmer Penguins Data Set

https://allisonhorst.github.io/palmerpenguins/

Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.



Note on Python Packages: statsmodels & sklearn

scikit-learn (sklearn)

- **Standard ML library in Python**; works seamlessly with tools like mlxtend for model selection.
- Geared toward **prediction tasks**: focus on training/test workflows, cross-validation, pipelines.
- Outputs coefficients and intercept, but no rich inference (no p-values, confidence intervals by default).
- Default choice for most machine learning models in this course.

statsmodels

- Built for **statistical inference and explanation**.
- Provides rich output: standard errors, p-values, confidence intervals, R^2, adjusted R^2, etc.
- Preferred **when the goal is to explain relationships or test hypotheses**, not just predict.
- Common in econometrics and social sciences, where interpretability is critical.

Demo: Regression on Penguins Data

Exercise: Regression on Penguins Data

End