Machine Learning for social sciences

Machine Learning: Session 2

William Aboucaya

Introduction

In this course, we will focus on regressions. This class of methods allows:

Introduction

In this course, we will focus on regressions. This class of methods allows:

- To **model relationships** between variables.

Introduction

In this course, we will focus on regressions. This class of methods allows:

- To **model relationships** between variables.
- To **predict outcomes** based on input data.

Introduction

In this course, we will focus on regressions. This class of methods allows:

- To **model relationships** between variables.
- To **predict outcomes** based on input data.
- To **understand how changes in input affect output**.

Introduction

In this course, we will focus on regressions. This class of methods allows:

- To **model relationships** between variables.
- To **predict outcomes** based on input data.
- To **understand how changes in input affect output**.

Regression is one of the most widely used techniques in data science, economics and social sciences.

Motivating Example

Scenario:

- You're analyzing how students perform based on different factors.
- Data collected:
  - Time spent studying in hours
  - Time spent sleeping in hours
  - Attendance at the classes
  - Exam score

Motivating Example

Scenario:

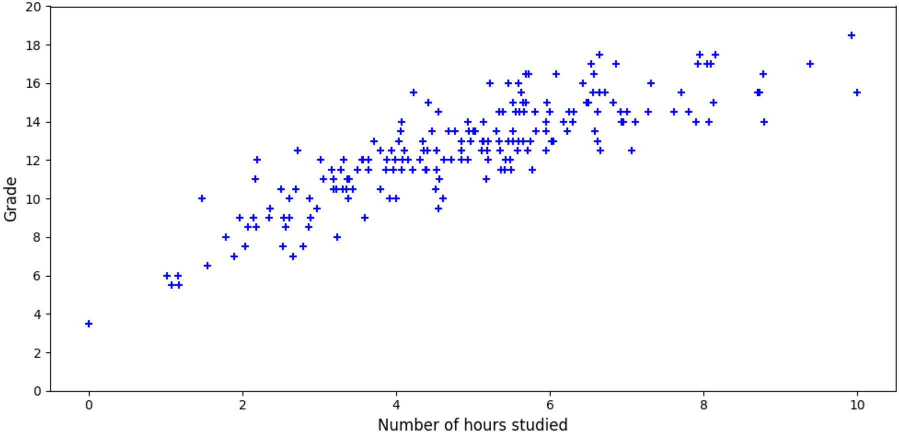
- You're analyzing how students perform based on different factors.
- Data collected:
    - Time spent studying in hours
    - Time spent sleeping in hours
    - Attendance at the classes
    - Exam score

Key Question:

If a student studies for 5 hours, attends all classes but has had a bad night's sleep before the exam, what score can we expect?

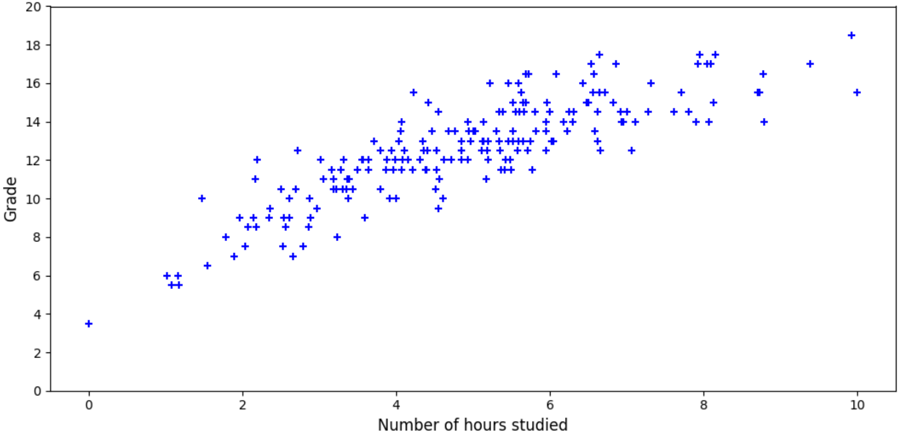
Visualizing the data

Partial representation of the data we will work with:



Visualizing the data

Partial representation of the data we will work with:



Is the relationship linear? Nonlinear? Binary?

What's Coming Up

We'll examine 3 types of regression using our example:

1. **Linear Regression** – Predicting exact score.
2. **Polynomial Regression** – Modeling nonlinear trends.
3. **Logistic Regression** – Predicting pass/fail outcomes.

Same data, different modeling goals.

Linear regression

A linear regression is a basic ML method to **predict a value** based on **one or multiple factors.**

Linear regression

A linear regression is a basic ML method to **predict a value** based on **one or multiple factors.**

The method tries to model the relationship between your features $[x_1; x_2; ...; x_n]$ and your desired output value y as a linear equation:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

With βs the weights given to each factor by **fitting** your equation to the training data, and $\hat{y}$ your predicted output.

Side-note: Model training / fitting

Goal: Find the best-fitting line that minimizes the difference between the predicted and actual values.

Model:

$$\hat{y} = \beta_0 + \beta_1 x_1$$

Side-note: Model training / fitting

Goal: Find the best-fitting line that minimizes the difference between the predicted and actual values.

Model:

$$\hat{y} = \beta_0 + \beta_1 x_1$$

Loss Function (Least Squares method):

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

Side-note: Model training / fitting

Goal: Find the best-fitting line that minimizes the difference between the predicted and actual values.

Model:

$$\hat{y} = \beta_0 + \beta_1 x_1$$

Loss Function (Least Squares method):

$$\text{Minimize} \quad \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Side-note: Model training / fitting

Goal: Find the best-fitting line that minimizes the difference between the predicted and actual values.

Model:

$$\hat{y} = \beta_0 + \beta_1 x_1$$

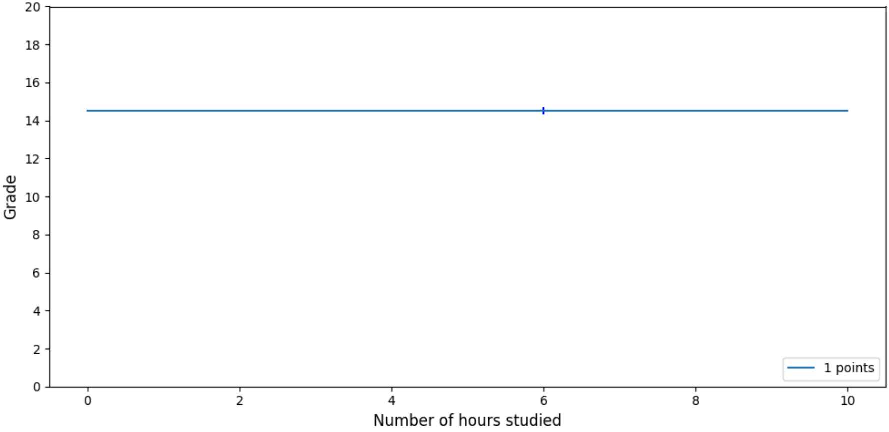
Loss Function (Least Squares method):

$$\text{Minimize} \quad \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

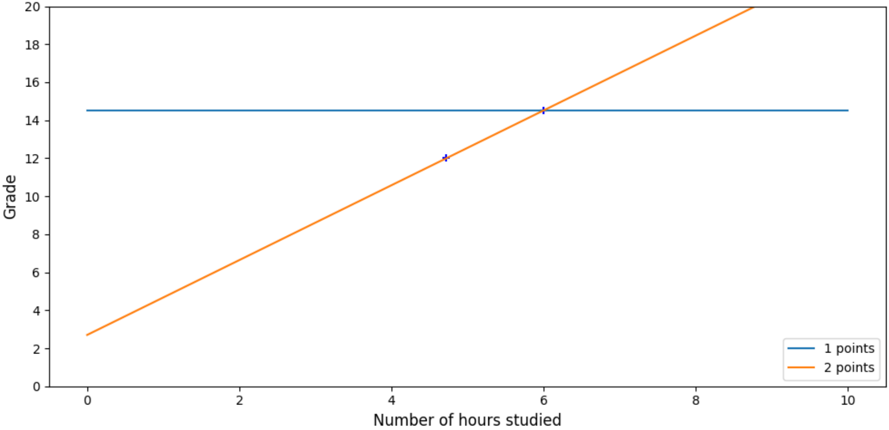
Solution: Choose β_0 and β_1 that minimize the squared error.

This class does not go into the mathematical details of how to compute your βs, we will have a Python library do it for us. But if you are interested, searching for "gradient descent" is a good first step.

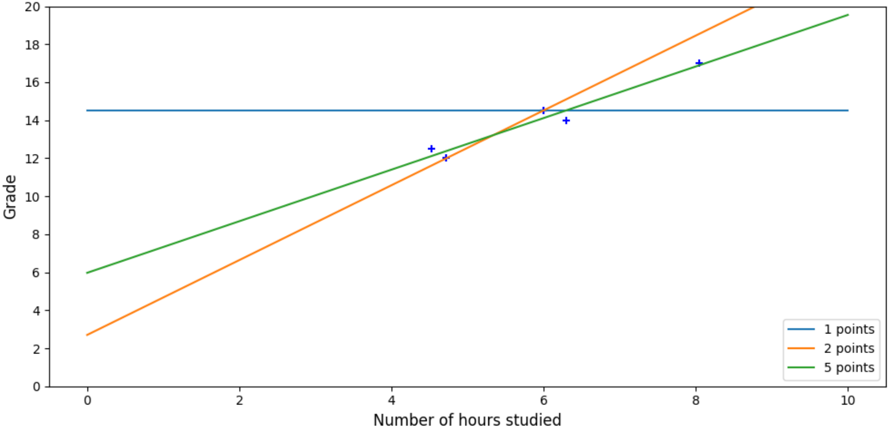
Linear regression in practice



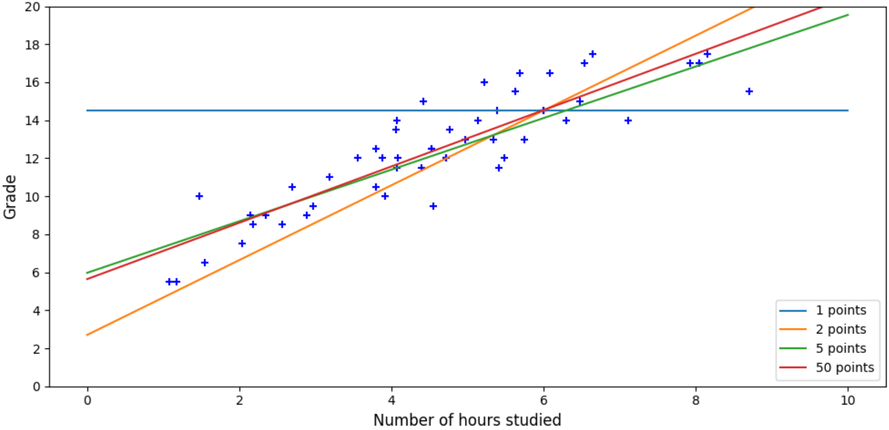
Linear regression in practice



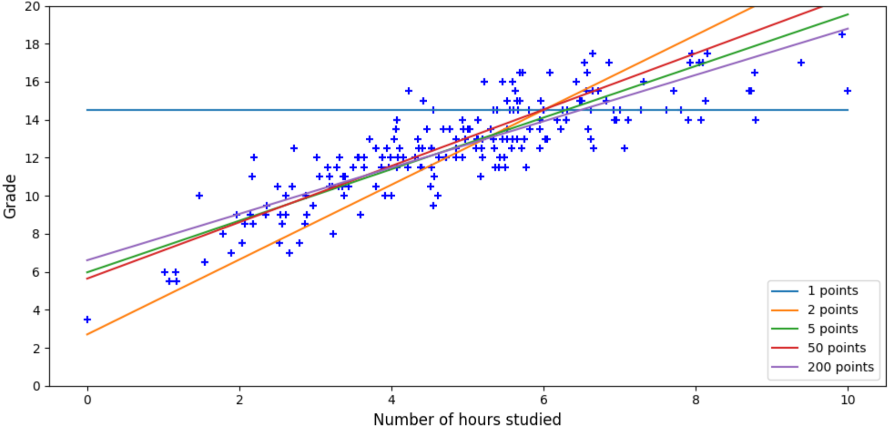
Linear regression in practice



Linear regression in practice



Linear regression in practice



Limits of linear regression

Problem: Using linear regressions is perfectly fine... to model linear relationships between independent variables and an output. But how can we model:

- Non-linear relationships?
- Relationships between the interaction of two features and an output?

Limits of linear regression

Problem: Using linear regressions is perfectly fine... to model linear relationships between independent variables and an output. But how can we model:

- Non-linear relationships?
- Relationships between the interaction of two features and an output?

Polynomial regressions and addition of **interaction terms** allow us to model these more complex relationships.

Polynomial regression

Idea: Extend linear regression by adding powers of the predictor variable(s) to model nonlinear relationships.

General Form:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n$$

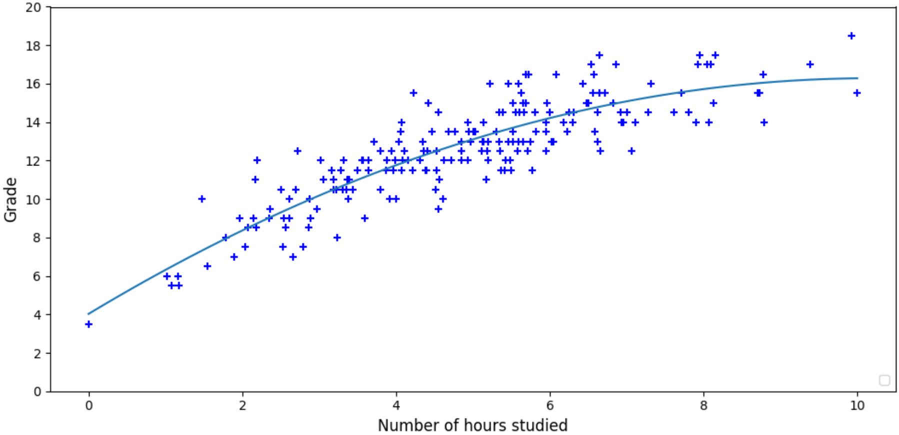
Key Points:

- Still a *linear model* in terms of the parameters β.
- Captures curves and nonlinear patterns.
- Risk of *overfitting* if degree is too high.

Polynomial regression

Example: Predict exam score from hours studied:

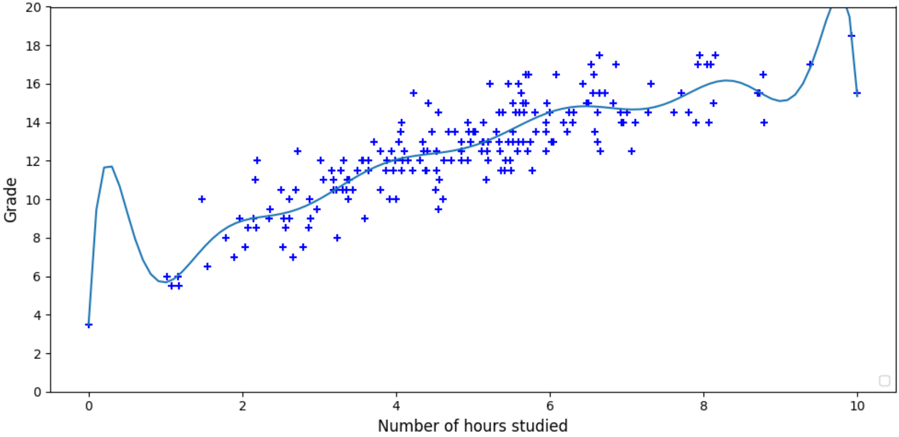
$$\text{Score} = \beta_0 + \beta_1(\text{Hours}) + \beta_2(\text{Hours})^2$$



Polynomial regression

Overfitting example by increasing the degree of the regression:

$$\text{Score} = \beta_0 + \beta_1(\text{Hours}) + \beta_2(\text{Hours})^2 + \cdots + \beta_{12}(\text{Hours})^{12}$$



Interaction terms

What are interaction terms?
Interaction terms capture the effect of two (or more) variables acting together by adding a product term. For example:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2)$$

Interpretation: The effect of x_1 depends on the level of x_2, and vice versa.

Interaction terms

Why are they interesting?

- Real-world relationships are rarely additive.
- Capture *conditional effects* between features.
- Reveal synergistic or moderating effects.

Interactions allow more flexible modeling:

- Example: Hours studied (x_1) and sleep quality (x_2) both affect exam score.
- Their combined effect may be larger (or smaller) than the sum of their individual effects.

Interaction terms

Shortcomings:

- **Interpretability**: Interaction coefficients can be difficult to explain.
- **Complexity**: The number of possible interactions grows quickly with the number of predictors.
- **Overfitting**: Including unnecessary interactions can harm generalization.
- **Multicollinearity**: Interaction terms often correlate with main effects.

From Regression to Classification

- So far, we have studied:
  - **Linear Regression —** predicts a continuous outcome.
  - **Polynomial Regression —** extends linear regression with nonlinear terms.

- But what if our target variable is **categorical**?
  - Example: Pass (1) or Fail (0)
  - Linear regression would produce invalid predictions (e.g., values < 0 or > 1)

- **Logistic Regression** solves this by modeling probabilities instead.

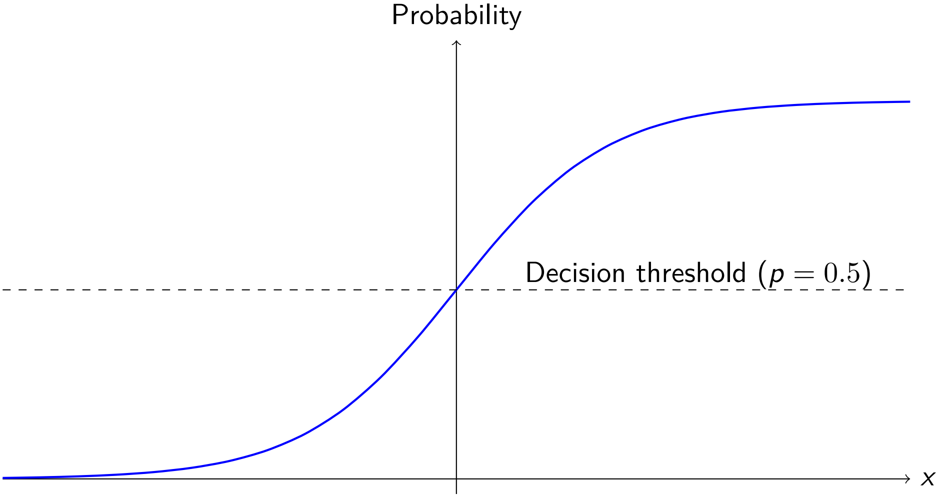
What is Logistic Regression?

- Logistic Regression models the probability that a given input belongs to a certain class.
- Instead of predicting y directly, we predict the probability:

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

- Output is constrained between 0 and 1.
- Commonly used for binary classification tasks, but can be generalized for multi-class.

Visualizing the Sigmoid Function



Example: Predicting Student Success

Goal: Predict whether a student passes or fails a course.

Features:

- x_1 = Hours of study
- x_2 = Hours of sleep per day
- x_3 = Number of classes attended

The logistic model:

$$P(\text{Pass}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}}$$

Decision Boundary

- The decision boundary occurs where $P(\text{Pass}) = 0.5$.
- That is, where:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = 0$$

- This separates students predicted to pass from those predicted to fail.

N.B. The 0.5 threshold is an example but the value can be modified depending on your needs. For example, a higher threshold is more suitable if you need very few false positives.

Interpretation of Coefficients

Interpretation:

- Coefficients relate to the log-odds of the outcome.
- Example: β_1 shows how the log-odds of passing change with one extra hour studied.
- Odds ratio $= e^{\beta_1}$ gives the multiplicative effect on odds.

In our example:

- Positive β_1: more study hours $\rightarrow$ higher chance to pass.
- Positive β_2: too little sleep might reduce the chance.
- Positive β_3: attending more classes helps.

Strengths and Limitations

Strengths

- Simple and interpretable.
- Works well for binary outcomes.
- Probabilistic predictions.

Limitations

- Assumes linear relationship between predictors and log-odds.
- Not ideal for complex nonlinear boundaries.
- Sensitive to outliers and imbalanced classes.

Summary

Which regression for a given task?

Example task	Model Type	Why?
Predict raw exam score	Linear / Polynomial	Continuous output
Model optimal study hours	Polynomial	Nonlinear relationship
Predict pass/fail	Logistic	Binary output