
OXMEDICA
Driving Global Education

Part 1
Linear Regression

Overview

1. Recap and functional forms
2. The linear regression model
3. A worked example
4. Correlation and Causation
5. Debating a real-life use of regression

Recap and functional forms

Questions in pairs: Supervised learning types

1. What is the aim of supervised learning?

2. How is it different to unsupervised learning?

3. What is the mathematical goal of supervised learning?

4. What are the two types of questions in supervised learning?

5. Based on the content we've covered over the last few days, think of 5 other questions about supervised learning you could ask.

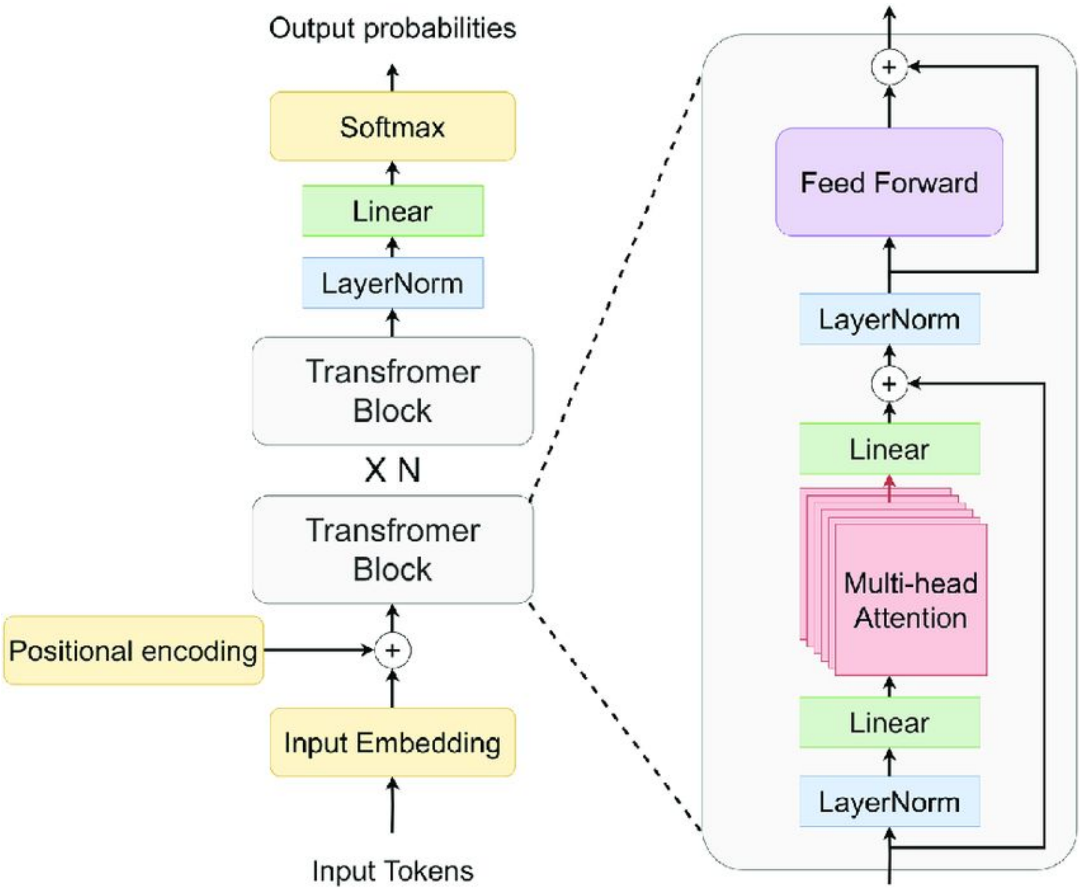
Notation refresher on the board

1. Parameters
2. Predictions
3. Models as functions
4. Loss functions

⚠️ Choosing the function ⚠️

- All supervised machine learning models can be interpreted as functions.
- We call different functional forms different ***models*** or ***architectures***
- What are the common different forms?
- How do we determine which function to use?

Even the most complex models today are still just functions…



Proof from the GPT 3.5 paper

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*

Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray

John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens

Amanda Askell† Peter Welinder Paul Christiano*†

Jan Leike* Ryan Lowe*

OpenAI

Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models *InstructGPT*. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

Model functions, parameters and loss functions

3.1 Unsupervised pre-training

Given an unsupervised corpus of tokens $\mathcal{U} = \{u_1, \ldots, u_n\}$, we use a standard language modeling objective to maximize the following likelihood:

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta) \tag{1}$$

where k is the size of the context window, and the conditional probability P is modeled using a neural network with parameters Θ. These parameters are trained using stochastic gradient descent [51].

In our experiments, we use a multi-layer *Transformer decoder* [34] for the language model, which is a variant of the transformer [62]. This model applies a multi-headed self-attention operation over the input context tokens followed by position-wise feedforward layers to produce an output distribution over target tokens:

$$\begin{aligned} h_0 &= UW_e + W_p \\ h_l &= \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n] \\ P(u) &= \texttt{softmax}(h_n W_e^T) \end{aligned} \tag{2}$$

where $U = (u_{-k}, \ldots, u_{-1})$ is the context vector of tokens, n is the number of layers, W_e is the token embedding matrix, and W_p is the position embedding matrix.

3.2 Supervised fine-tuning

After training the model with the objective in Eq. 1, we adapt the parameters to the supervised target task. We assume a labeled dataset $\mathcal{C}$, where each instance consists of a sequence of input tokens, $x^1, \ldots, x^m$, along with a label y. The inputs are passed through our pre-trained model to obtain the final transformer block's activation h_l^m, which is then fed into an added linear output layer with parameters W_y to predict y:

$$P(y|x^1, \ldots, x^m) = \texttt{softmax}(h_l^m W_y). \tag{3}$$

The function

Model functions, parameters and loss functions

Specifically, the loss function for the reward model is:

$$\operatorname{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_\theta\left(x, y_w\right) - r_\theta\left(x, y_l\right)\right)\right)\right] \tag{1}$$

where $r_\theta(x, y)$ is the scalar output of the reward model for prompt x and completion y with parameters θ, y_w is the preferred completion out of the pair of y_w and y_l, and D is the dataset of human comparisons.

The parameters

The loss function

Different choices for the function are incredibly important

Model	Functional Form	Model Complexity	Model Performance

Different choices for the function are incredibly important

Model	Functional Form	Model Complexity	Model Performance
Linear regression	$h_\theta(X)$	Low	Low

Different choices for the function are incredibly important

Model	Functional Form	Model Complexity	Model Performance
Linear regression	$h_\theta(X)$	Low	Low
Logistic regression	$h_\theta(X)$	Low	Low

Different choices for the function are incredibly important

Model	Functional Form	Model Complexity	Model Performance
Linear regression	$h_\theta(X)$	Low	Low
Logistic regression	$h_\theta(X)$	Low	Low
Classification and regression trees	$h_\theta(X)$	Medium	Low

Different choices for the function are incredibly important

Model	Functional Form	Model Complexity	Model Performance
Linear regression	$h_\theta(X)$	Low	Low
Logistic regression	$h_\theta(X)$	Low	Low
Classification and regression trees	$h_\theta(X)$	Medium	Low
Random forest and boosting	$h_\theta(X)$	Medium-High	High

Different choices for the function are incredibly important

Model	Functional Form	Model Complexity	Model Performance
Linear regression	$h_\theta(X)$	Low	Low
Logistic regression	$h_\theta(X)$	Low	Low
Classification and regression trees	$h_\theta(X)$	Medium	Low
Random forest and boosting	$h_\theta(X)$	Medium-High	High
Support vector machines	$h_\theta(X)$	High	High

Different choices for the function are incredibly important

Model	Functional Form	Model Complexity	Model Performance
Linear regression	$h_\theta(X)$	Low	Low
Logistic regression	$h_\theta(X)$	Low	Low
Classification and regression trees	$h_\theta(X)$	Medium	Low
Random forest and boosting	$h_\theta(X)$	Medium-High	High
Support vector machines	$h_\theta(X)$	High	High
Simple Neural Networks	$h_\theta(X)$	High	High

Different choices for the function are incredibly important

Model	Functional Form	Model Complexity	Model Performance
Linear regression	$h_\theta(X)$	Low	Low
Logistic regression	$h_\theta(X)$	Low	Low
Classification and regression trees	$h_\theta(X)$	Medium	Low
Random forest and boosting	$h_\theta(X)$	Medium-High	High
Support vector machines	$h_\theta(X)$	High	High
Simple Neural Networks	$h_\theta(X)$	High	High
Transformer Neural Networks	$h_\theta(X)$	Very Extreme	Very High

Linear regression

A motivating example: Education and future wages

See whiteboard

A simple functional form: World assumption

$$y^{(i)} = \alpha + \beta x_1^{(i)}$$

A simple functional form: The Linear regression model

$$\hat{y}^{(i)} = \hat{\alpha} + \hat{\beta} x_1^{(i)} + \varepsilon^{(i)}$$

Final Model

$$\hat{y}^{(i)} = \hat{\alpha} + \hat{\beta} x_1^{(i)}$$

Individual Task: Worksheet

https://www.harrymayne.com/oxmedica

Extending this to more features

$$\hat{y}^{(i)} = \hat{\alpha} + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \hat{\beta}_3 x_3^{(i)} + \hat{\beta}_4 x_4^{(i)}$$

Discussion

- How good is this model?
- Why might it fail?
- …lots of other possible questions…!

5 Min Break

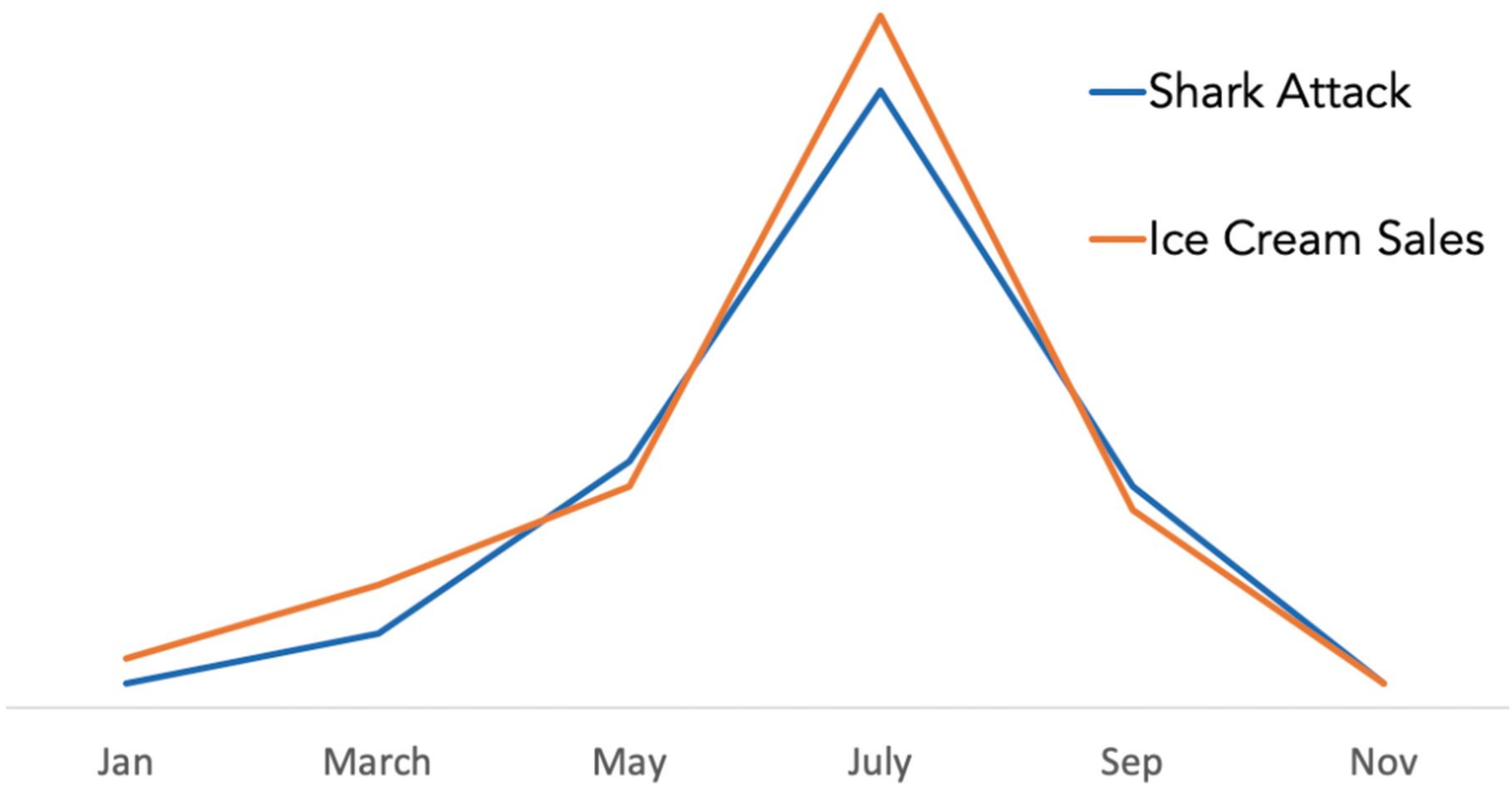
Part 2
Correlation vs Causation

Correlation and causation

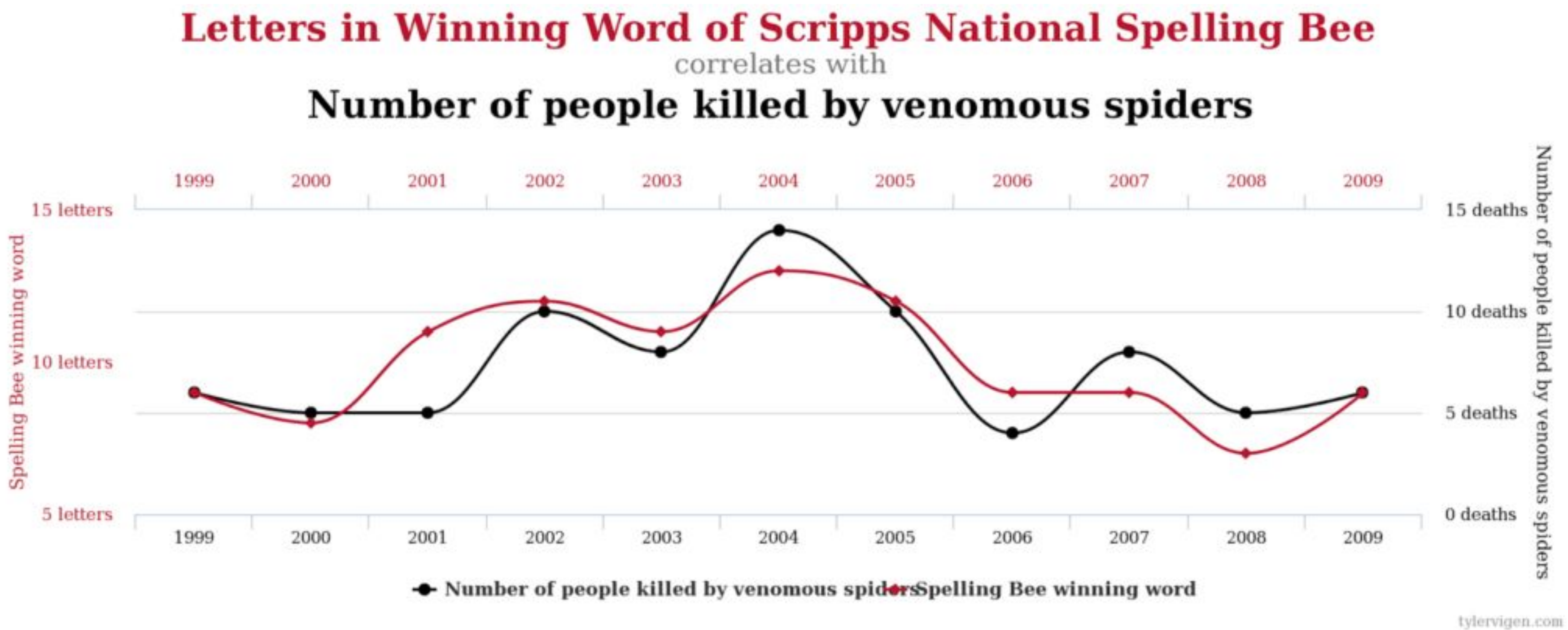
Correlation: A relationship between two variables.

Causality: A change in one variable causes a change in another variable.

- How are they different?
- Is this important?
- Examples of things which correlate well but are not causal?



[Source](#)


Letters in Winning Word of Scripps National Spelling Bee
correlates with
Number of people killed by venomous spiders
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
15 letters
10 letters
5 letters
Spelling Bee winning word
15 deaths
10 deaths
5 deaths
0 deaths
Number of people killed by venomous spiders
Number of people killed by venomous spiders
Spelling Bee winning word
tylervigen.com

[Source](#)

What does this really tell us?

$$\hat{y}^{(i)} = \hat{\alpha} + \hat{\beta} x_1^{(i)} + \varepsilon^{(i)}$$

EXTENSION: The Gauss-Markov Theorem

Gauss–Markov theorem

文A 17 languages

Article Talk

Read Edit View history Tools

From Wikipedia, the free encyclopedia

Not to be confused with Gauss–Markov process.
"BLUE" redirects here. For queue management algorithm, see Blue (queue management algorithm). For the color, see Blue.

In statistics, the **Gauss–Markov theorem** (or simply **Gauss theorem** for some authors) [1] states that the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero. [2] The errors do not need to be normal for the theorem to apply, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance).

The requirement for unbiasedness cannot be dropped, since biased estimators exist with lower variance and mean squared error. For example, the James–Stein estimator (which also drops linearity) and ridge regression typically outperform ordinary least squares. In fact, ordinary least squares is rarely even an admissible estimator, as Stein's phenomenon shows--when estimating more than two unknown variables, ordinary least squares will always perform worse (in mean squared error) than Stein's estimator.

Part of a series on
Regression analysis
Models
Linear regression · Simple regression · Polynomial regression · General linear model
Generalized linear model · Vector generalized linear model · Discrete choice · Binomial regression · Binary regression · Logistic regression · Multinomial logistic regression · Mixed logit · Probit · Multinomial probit · Ordered logit · Ordered probit · Poisson
Multilevel model · Fixed effects · Random effects · Linear mixed-effects model · Nonlinear mixed-effects model
Nonlinear regression · Nonparametric · Semiparametric · Robust · Quantile · Isotonic · Principal components · Least angle · Local ·

Reverse Causality

Reverse Causality: When a change in the y causes a change in the x in a causal way. I.e. we assumed the relationship was the other way round.

Reverse Causality

Mental Health and Social Media Use:

- What is the **perceived causality**
- What might the **reverse causality** be?

- **Perceived Causality:** Increased social media use leads to poor mental health.
- **Reverse Causality:** Individuals with poor mental health are more likely to spend more time on social media as a form of escapism or social connection.

A Debate

Task in groups of 4: Debate

In your group of 4 you are in teams of 2, which take different sides of the debate. You are going to debate the following clause

“Crime rates are generally higher in areas with lower socioeconomic status. Given this, should law enforcement use socioeconomic data to predict future crime rates and allocate resources for crime prevention?”

One pair will argue that you should use regression models for this, the other will argue that you should not. You and your pair have **15 minutes** to prepare 2 arguments for your side. You are welcome to do some research and use evidence in your arguments

After 15 minutes you will take it in turn to make one argument (uninterrupted by the other pair). After all arguments have been made, you have another **5 minutes** to prepare a response to the other team’s arguments.

EXTRA
A real example

Paper comprehension

Only if time…