Regressions

A theoretical introduction to regressions

Do you use regressions every day, but feel unsure about the math behind them? Maybe you plug ‘n chug numbers into the lm function in R, or the LinearRegression function from the Sklearn library in Python to extract the relevant values without fully understanding what’s happening beneath the surface. If that sounds familiar, you are not alone. In this three-part series, I will introduce you to regressions, move on to simple linear regressions, and conclude with multiple linear regressions.

Regressions and regression analysis are essential tools in fields like economics, actuarial science, market analytics, data science, and machine learning. For statisticians and data analysts, they’re the bread and butter of understanding relationships between variables. To me, regressions are to algebra what calculus is to machine learning—they’re foundational and incredibly versatile.

The goal is to help you understand what’s really happening when you use regression functions in Python or R—what’s going on inside the so-called *‘black box’*. In machine learning, a black box refers to a model whose internal process is hidden from us. While we can observe the inputs and outputs, the math and decision-making processes of the model are not transparent. In this series, I will walk you through the math, and manually break down the calculations in Python, so you can see how everything works step by step.

Quick History

The term “regression” has its roots in an interesting context. Although regression analyses were popularized by Economists in the 50s and 60s, Sir Francis Galton first used it in the late 1800s while studying the relationship between the heights of fathers and their sons. Intuitively, tall fathers should have tall(er) sons, but his calculations found that sons of tall

fathers regressed to the average height. The trend of height to "regress" toward the mean inspired the name, not the calculations themselves. It is important to note here that Sir Francis Galton was also the father of eugenics (*side eye*). Over time, regression analysis has evolved into a powerful tool for understanding and modeling relationships between variables.

At its core, regression is a statistical method that helps us understand the relationship between two or more variables. Understanding the relationship between variables has a lot of real-world applications; retailers know how much sugar to keep in stock based on customer behavior, real estate agencies know how much to list a house on the market for, and policy advisors can be informed about the potential consequences of a policy even before they are put into place. Regressions help us to not only make predictions based on historical data but also identify patterns in real-time data.

Terminology

Now that we understand what regression is and how it can be used, let's break down some common terms you'll encounter when building, evaluating, and interpreting a model. Knowing these terms will give you a solid foundation to tackle regression equations and calculations in upcoming posts.

1. **Dependent variable**

Also known as: target variable, regressand, outcome variable; commonly detonated as ***Y***

The dependent variable is what we are trying to estimate or understand. How does it change based on the values of other variables? In Galton's height study, the dependent variable was the height of sons.

2. **Independent variable**

Also known as: predictors, regressors, and input variables; commonly detonated as ***X***

The independent variable is used to predict the dependent variable. In Galton's case, the independent variable was the height of the fathers.

3. Intercept

Commonly denoted as $\boldsymbol{\beta 0}$ (the first character is Beta in the Greek alphabet)

The intercept is the point where the regression line crosses the y-axis. It can also be thought of as the value of the dependent variable (*Y*) when all the independent variables (*X*) are 0.

4. Coefficient

Also known as: slope, gradient; commonly detonated as $\boldsymbol{\beta 1}$

The coefficient measures how much the dependent variable changes when an independent variable increases by one unit. Graphically, it's the slope of the line. If a model has *n* predictors, it will have *n* coefficients.

5. Residuals

Also known as: fitting deviation

Residuals are the differences between the actual values and the predicted values from a model. For example, if your model predicts a house will sell for $780,000, but it actually sells for $790,000, the residual is $10,000. Residuals tell us how well the model fits the data.

6. Error term

Denoted as **ε** (Epsilon in the Greek alphabet)

The error term is the difference between the true value and the prediction of a perfect model. In the real world, we rarely build a 'perfect model' whose 'perfect' predictions are identical or very close to the actual values. Instead, we settle for a model that estimates the true relationship between variables. Because we can't know the exact difference between an ideal prediction and the actual value, we use residuals as approximations of error terms. So, though there is no practical difference between residuals and errors, there is a *theoretical* difference between them.

7. Correlation

Correlation measures the degree to which two variables are related and how they change together. A *positive correlation* implies that if one variable increases the other increases too. A *negative correlation* suggests that if one variable increases the other decreases. Finally, a *0*

correlation indicates no relationship between two variables exists. Dependent and independent variables can be correlated with each other. It's important to note that correlation does not imply causation.

For example, when temperature increases, the sales of ice cream will also increase (positive correlation). However, on hot days, most customers want either ice cream or soda, indicating that if ice cream sales increase, soda sales will decrease (negative correlation).

Types of Regressions

Great, you've built a strong foundation in regression terminology! Next, we'll put this knowledge to use by exploring the regression family tree. Depending on your data and the question you're trying to answer, different types of regression models may be better suited. Let's map out the regression family and highlight some of the most common types.

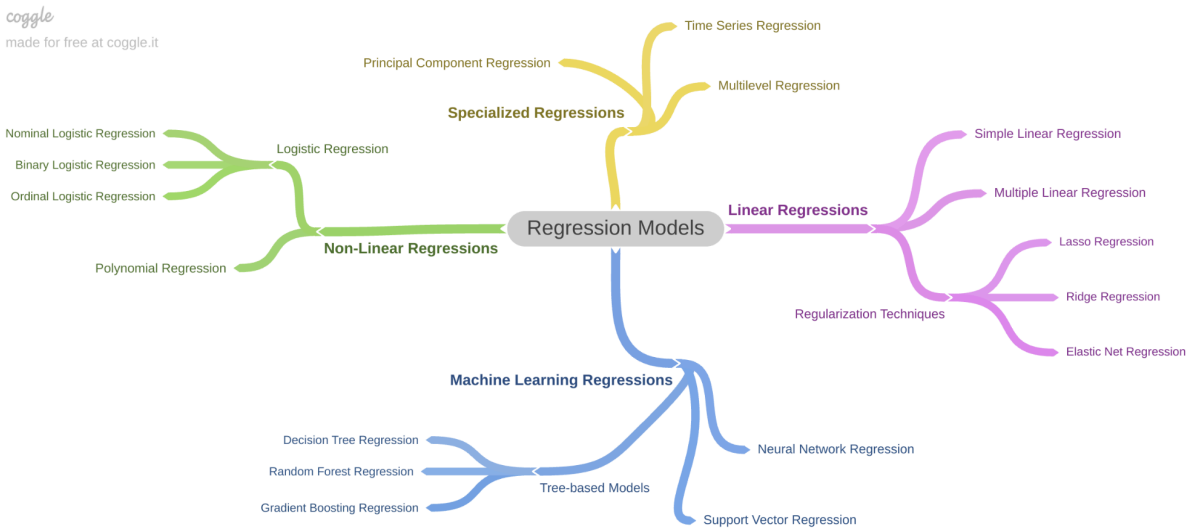


Figure 1

I've broken down the most popular regression models into four categories:

1. **Linear regressions**: Linear regressions are used when the relationship between the regressand and regressor(s) is linear. These types of regressions are ideal for predicting continuous variables (e.g. sales revenue).

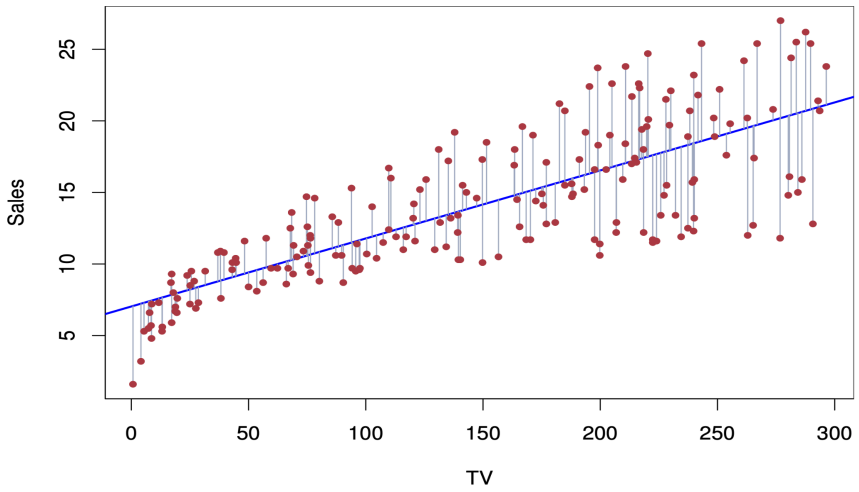


Figure 2[1]

Sometimes, linear regressions are applied to a model to improve its ability to generalize data and reduce its complexity when needed, thereby increasing its accuracy. This technique is known as *regularization.* If you're just getting started with regressions, you can learn more about this later. For now, this foundational understanding is all you need.

2. **Non-Linear regressions:** Non-linear regressions model the relationship between a regressand and regressor(s) when it is non-linear. These relationships may follow curves, exponential growth/decay, or other patterns that are not lines (e.g. the height of a basketball after it's bounced). What might this look like? Here are some examples:

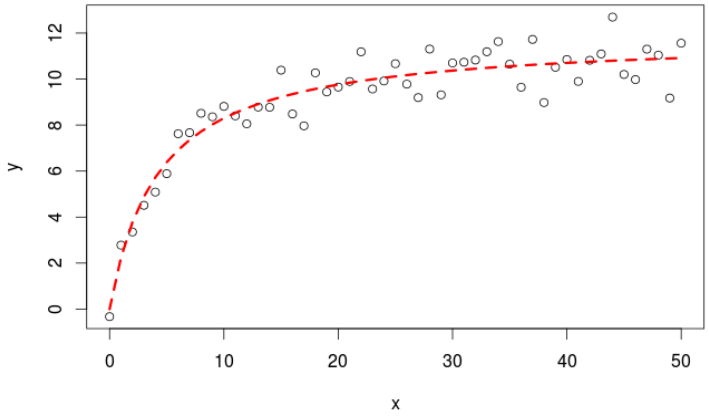
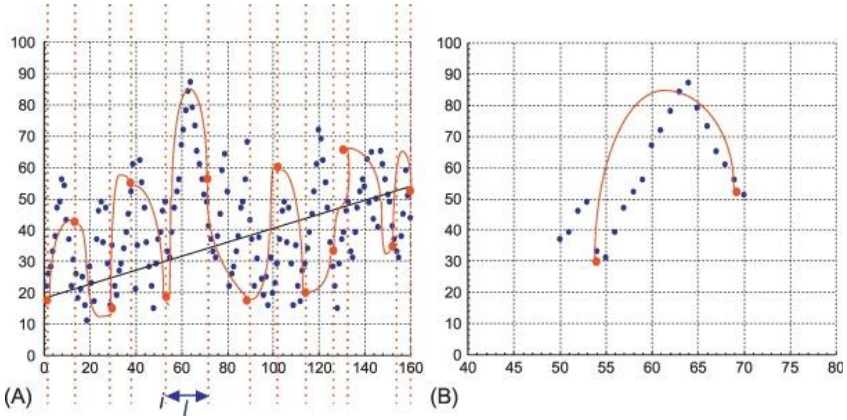


Figure 3[2]

[1] *Bacallado, Sergio, and Jonathan Taylor. "Simple Linear Regression — STATS 202." Web.stanford.edu, 2022, web.stanford.edu/class/stats202/notes/Linear-regression/Simple-linear-regression.html.*

[2] *Hertzog, Lionel. "First Steps with Non-Linear Regression in R | DataScience+." Datascienceplus.com, 2016, datascienceplus.com/first-steps-with-non-linear-regression-in-r/. Accessed 14 Mar. 2025.*



Figures 4 & 5[3]

One example of non-linear regressions is *Polynomial Regression*, which can model curvilinear patterns in the data. Another key category is *Logistic Regression*, which predicts the probability of an event occurring. Logistic regression is widely used for classification problems and can be broken down into three types: nominal (unordered categories), binary (two outcomes), or ordinal (ordered categories).

3. **Machine learning regressions:** While traditional regressions focus on understanding relationships between variables, machine learning regressions are designed for making predictions. Unlike traditional methods, ML regressions can efficiently handle larger and more complex datasets while automating the ranking of features based on their importance to the predictors—a task that is performed manually in linear and non-linear regression models. These regressions mark an evolution into more sophisticated, black-box methods.

   The first species to evolve from fins to feet were the tree-based models. *Decision tree regressions* split a dataset by the values in features - ranked most to least important - until it creates groups of like data points or is asked to stop at n number of splits. This flexibility allows decision trees to handle various types of input, such as binary, continuous, and categorical variables. For example, a bank could use decision trees when

[3] *Pal, Ranadip. "Overview of Predictive Modeling Based on Genomic Characterizations." Predictive Modeling of Drug Sensitivity, 2017, pp. 121–148, https://doi.org/10.1016/b978-0-12-805274-7.00006-3. Accessed 23 Dec. 2021.*

they are predicting whether to approve a loan for a customer based on their financial profile. A graphical representation of that tree would look like this:

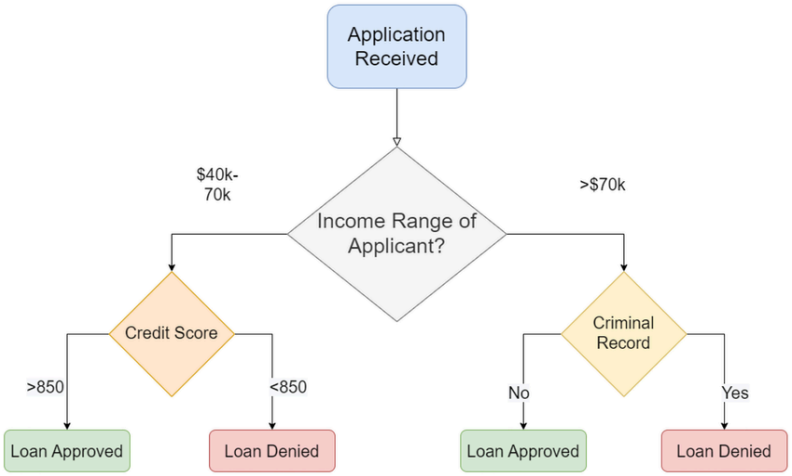


Figure 6[4]

Amplifying the power of decision trees gave rise to models like *random forest regression* and *gradient-boosting machines*. Random forests average the results of many independent trees, creating a more robust model. On the other hand, gradient-boosting machines build decision trees sequentially, with each tree learning from the errors of its predecessor to improve predictions. Tree regressions have widespread applications, like fraudulent transaction detection and disease diagnoses.

Other advanced methods, like *neural network regressions* and *support vector regressions (SVRs)*, are adaptations of machine learning algorithms specifically tailored for regression tasks. For example, though neural networks are more commonly associated with classification problems, they can be fine-tuned for regression prediction problems. These models excel at handling complex, non-linear datasets and uncovering patterns that traditional methods might miss. While these topics deserve their own deep dive, know that they represent some of the most powerful tools in the machine learning toolbox.

4. **Specialized regressions**: Though there is no formal, universally accepted "specialized" regression class, I included this category to highlight models built for specific use cases

[4]*Hassija, Vikas , et al. "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence." Cognitive Computation, vol. 16, no. 1, 24 Aug. 2023, https://doi.org/10.1007/s12559-023-10179-8.*

outside the traditional linear, non-linear, and machine learning families. These models address unique challenges and often cater to niche problems. *Principal component regression (PCR)* addresses multicollinearity issues in datasets with many correlated variables. Alternatively, *time series regressions* are models built on historical, time-ordered data, and are used to analyze patterns and relationships to predict future outcomes. These regressions are particularly skilled at identifying trends, seasonality, and other time-dependent patterns. Lastly, *multilevel regressions*, also known as hierarchical linear modeling, are used when data has a nested structure. For instance, student performance can be analyzed at multiple levels: within classrooms, schools, and even districts.

These specialized regressions extend the versatility of regression modeling, offering solutions for problems that traditional models may struggle to address.

Wrapping Up: From Complex Models to Simple Beginnings

Regression is a versatile tool that adapts to various challenges, from understanding relationships between variables to solving complex prediction problems with specialized models. But while we've explored a wide array of advanced and specialized regressions, they all trace their roots back to one rudimentary model: simple linear regression. This foundational approach, often the first step in learning regression analysis, provides the building blocks for understanding how variables relate to one another.

In my next post, I'll take you inside the **black box of simple linear regression**. We'll break down the math behind the model, manually implement it in Python, and explore key metrics used to evaluate its performance. As we progress through the series, we'll also dive into more advanced topics - some of which are on the mind map above.

If you've ever wondered how simple linear regression actually works under the hood, stay tuned—next time, we're coding it from scratch!