# Regression, Explained Visually

## An Interactive, Educational Introduction to Regression

Sam Lau
UC Berkeley
EECS
samlau95@berkeley.edu

## ABSTRACT

Regression, the estimation of relationships between variables, is one of the primary use cases of machine learning. However, regression is often taught using a statistics-first approach. While important, this teaching method often fails to help students develop a deep understanding of the topic beyond the equations. In this educational project, I take an intuition-first approach to explaining regression. The resulting web application uses a combination of interaction and animation to engage the learner and build their intuition for linear regression, polynomial regression, and cross-validation.

## General Terms

Machine Learning

## Keywords

Regression, Education, Visual, Interactive

## 1. INTRODUCTION

How do individuals learn about regression? I consider three major types of learners studying regression for the first time:

1. Individuals primarily using textbooks.
2. Individuals primarily using online resources.
3. Individuals in a traditional classroom setting, eg. at a university.

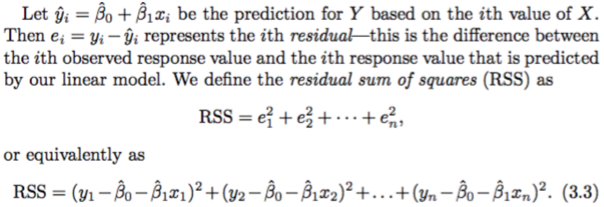
I argue that while each of these three major methods can be effective, all of them at times permit the learner to pass through without a deep understanding of regression. I will then argue for the value of this project in helping the learner develop the intuition needed to understand regression.

### 1.1 Learning through textbooks

One standard approach for learning regression is through studying a textbook. This has a relatively low barrier to entry because the two canonical textbooks for machine learning are available online free of charge: Introduction to Statistical Learning (ISL) and The Elements of Statistical Learning (ESL) [7] [6].

However, this method poses a number of challenges for the first-time learner. Although textbooks are thorough, they are often very theoretical and assume a strong background in mathematical notation. For example, ISL contains this excerpt within the first couple pages after mentioning linear regression:

**Figure 1: Excerpt from ISL**



Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for $Y$ based on the $i$th value of $X$. Then $e_i = y_i - \hat{y}_i$ represents the $i$th *residual*—this is the difference between the $i$th observed response value and the $i$th response value that is predicted by our linear model. We define the ***residual sum of squares*** (RSS) as

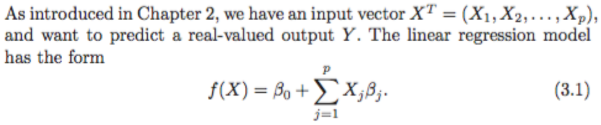
$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \quad \textbf{(3.3)}$$

ESL contains this excerpt on the second page of its linear regression section:

**Figure 2: Excerpt from ESL**



As introduced in Chapter 2, we have an input vector $X^T = (X_1, X_2, \ldots, X_p)$, and want to predict a real-valued output $Y$. The linear regression model has the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j. \quad (3.1)$$

While equations are necessary for describing the exact formulation of regression, these textbooks place a significant emphasis on the theoretical underpinnings of regression. Both ISL and ESL describe linear regression mathematically for over 50 pages with over 30 equations (in the case of ESL, 84 equations) before giving their first exercise.

Simply based on way textbooks present the material, the learner can be misled to believe that focusing on the mathematics of regression are the best way to understand it when recent research in education suggests otherwise [9].

### 1.2 Learning through online resources

Another common approach to learn regression is through online resources. There are a number of freely available articles and courses for the learner. However, online content has high variability in its quality. For example, a Google search of ”intro to regression machine learning” yields several articles, lecture slides, and links to Python libraries for machine learning.

This type of online content has a number of issues. In constrast to textbooks, online articles have no promises of quality or completeness. Lecture slides are usually posted without the accompanying lecture, making them difficult to follow.

More glaringly, online articles typically contain very similar content to textbooks. Most are primarily composed of text and images. While many online articles are often easier for the layperson to understand, most do not help the learner any more than textbooks do because the presentation of the content is the same. A learner might as well go to a textbook to read about regression since the content is presented more thoroughly.

### 1.3 Learning in a traditional classroom setting

Learning through a traditional classroom setting has a number of benefits over textbook and online resource-based learning. For example, students are encouraged to ask questions because they can receive immediate feedback. The material is usually presented in a thorough way and asking the student to complete assignments allows the student to actively engage with the material, in stark contrast to the previous two learning methods.

However, the way students are evaluated in the classroom often creates incentives that work against developing intuition for regression. Namely, the majority of exams in these courses again place a large emphasis on the mathematics of regression. In Stanford’s CS229, all of their past exam problems on regression involve heavy use of matrix algebra [8]. Most of their problems can be solved without prior knowledge of regression! A student seeking to succeed on an exam in such a class would be incentivized to understand calculus and linear algebra instead of regression.

## 2. APPROACH

All methods of learning have pros and cons. In this project, I choose to forgo mathematical completeness and instead emphasize active learning. Whereas textbooks, online articles, and lectures encourage passive learning through absoption of information, I aim to encourage the learner to actively ask and answer their own questions.

The basic premise of this project is to encourage the user to engage with the concepts of regression through interaction and animation. For example, for the introduction to linear regression the user can drag points around and see both regression line and data immediately update. This naturally encourages the follow questions:

1. Does the slope become negative if the points curve downward?
2. What if all the points lie on the same line except one point?
3. What happens if all the points have the same x-value?
4. What happens if all the points have the same y-value?

The user can immediately test out these questions and get answers. Because of this the user is encouraged to ask more questions, developing an understanding of the behavior of regression through implicitly posing hypotheses. Then, the user is equipped to connect their understanding with the mathematics. The mathematics is a supplement to their intuition, not the other way around.

The rest of this section discusses the topics covered by this project and typical questions that a user can pose and answer.

### 2.1 Linear Regression

In this section the project introduces the idea of linear regression as a method of prediction. As mentioned earlier, the user is presented with a scatterplot with the linear regression line overlain on top. The user can move the points around and see the data, regression line, and regression equation update in real-time. I then proceed to explain the least squares error formulation. In addition to the questions mentioned above, this section allows the user to answer the following questions:

1. What is the squared error of a point when the line passes through the point?
2. What is the error of a point when the line misses?
3. Is the error larger when the line is further away from the point?
4. Can the error have a negative value?

### 2.2 Polynomial Regression

This section introduces polynomial regression as a way of dealing with nonlinear patterns in the data. This section contains the same dataset with multiple polynomials of increasing degrees overlain on top. When a point is moved, all the curves update. This encourages the following questions:

1. Can the degree 2 polynomial also fit a linear pattern?
2. Can the degree 5 polynomial also fit a linear pattern?
3. If the fitted degree 2 polynomial is concave down, is the degree 5 polynomial also concave down?

### 2.3 Training Error

This section introduces training error as a potenial method of evaluating which regression model is the best fit for the data. This section allows the user to move the data points, redrawing the regression curves and recalculating the training error each time. This encourages the following questions:

1. When does linear regression have lower training error than a degree 2 polynomial?
2. When does degree 2 polynomial have lower training error than a degree 5 polynomial?
3. Does a degree 5 polynomial always have a lower training error than a degree 2? Than a linear fit?
4. Does a degree 10 polynomial always have a training error of 0 on this dataset?

## 2.4 Cross Validation

This section explains why the training error is not an accurate way to select a model and proceeds to explain why the cross validation error is more appropriate. This section has little interactive parts because dragging training set points makes the validation data invalid. However, it answers the following questions:

1. Does the validation error give a more accurate measure of model fit?
2. If a model fits the training data perfectly, will it have a small validation error?
3. If a model underfits the training data, will it also have a large validation error?
4. Will increasing the degree of the polynomial always decrease the training error? The validation error?

# 3. IMPLEMENTATION

This project is implemented as a web application. The code is open-source and is available at `https://github.com/SamLau95/regression-explained`. The majority of the code is written in ES6 Javascript, using the React [3] and Redux [4] libraries to render the user interface and keep track of the application state. I used the Highcharts [2] library to provide plotting functionality.

I used the publicly available dataset from Larry Winner (University of Florida) on study on ice cream conducted in 1997 [1] [5].

# 4. RESULTS AND DISCUSSION

The resulting web application is available at `http://www.samlau.me/regression-explained/`. The page provides an introduction to regression with interactive charts. Figures 3 and 4 show the same chart before and after user interaction.

I'm pleased with the result and the project has high user engagement among a randomly selected beta test group (read: my friends). I hope that building this project will help others develop an intuitive understanding for machine learning through regression and I look forward to working on more educational projects in the future.

**Figure 3: Default dataset with fitted degree 5 polynomial**

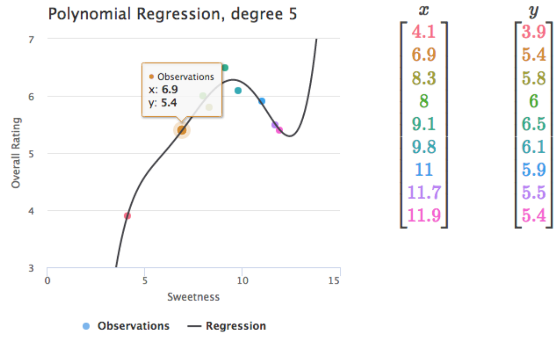
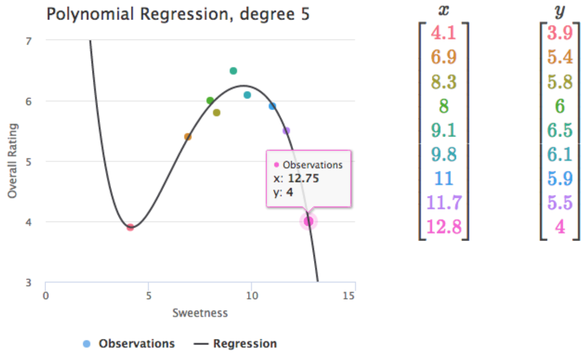


**Figure 4: User-changed dataset with fitted degree 5 polynomial**



# 5. REFERENCES

[1] datasets. `http://www.stat.ufl.edu/~winner/datasets.html`. (Accessed on 05/05/2017).

[2] Interactive javascript charts for your webpage | highcharts. `https://www.highcharts.com/`. (Accessed on 05/05/2017).

[3] React - a javascript library for building user interfaces. `https://facebook.github.io/react/`. (Accessed on 05/05/2017).

[4] reactjs/redux: Predictable state container for javascript apps. `https://github.com/reactjs/redux`. (Accessed on 05/05/2017).

[5] J.-X. GUINARD, C. ZOUMAS-MORSE, L. Mori, B. Uatoni, D. Panyam, and A. Kilara. Sugar and fat effects on sensory properties of ice cream. *Journal of food science*, 62(5):1087–1094, 1997.

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

[8] A. Ng. Cs 229: Machine learning (course handouts). `http://cs229.stanford.edu/materials.html`. (Accessed on 05/04/2017).

[9] T. Van Gog and N. Rummel. Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review*, 22(2):155–174, 2010.