Linear Regression: How to Make What's Old New Again

Kimberly Roye, Sara Jardine, and Christian Smart

GALORATH

2022 ICEAA Workshop

We exist to empower informed decision making so that organizations can achieve their goals with greater confidence.

GALORATH

Agenda

Linear Regression

- Supervised Machine Learning
- Optimization
- Example Software Sustainment Program Dataset

Regularization

- Ridge Regression
- Lasso Regression
- Elastic Net Regression

Gradient Descent

- Steps to improve regression results
- Regression Model Comparison
- Conclusion

GALORATH

Why Machine Learning?

1

Gaining Popularity

As data is being captured every second, there is an abundance of data available to analyze

2

Predictive Accuracy

Algorithms are available to help increase the predictive accuracy of simpler methods

3

Linear Regression is ML

Linear regression is not a forgotten method that will easily be replaced with other more complicated methods

4

A Time and Place for Everything

Alternative techniques to linear regression using Ordinary Least Squares exists and can be useful. These methods are not always better though

"Data is the new oil" – Clive Humby, mathematician and entrepreneur

Linear Regression

Most understood method of **supervised** machine learning

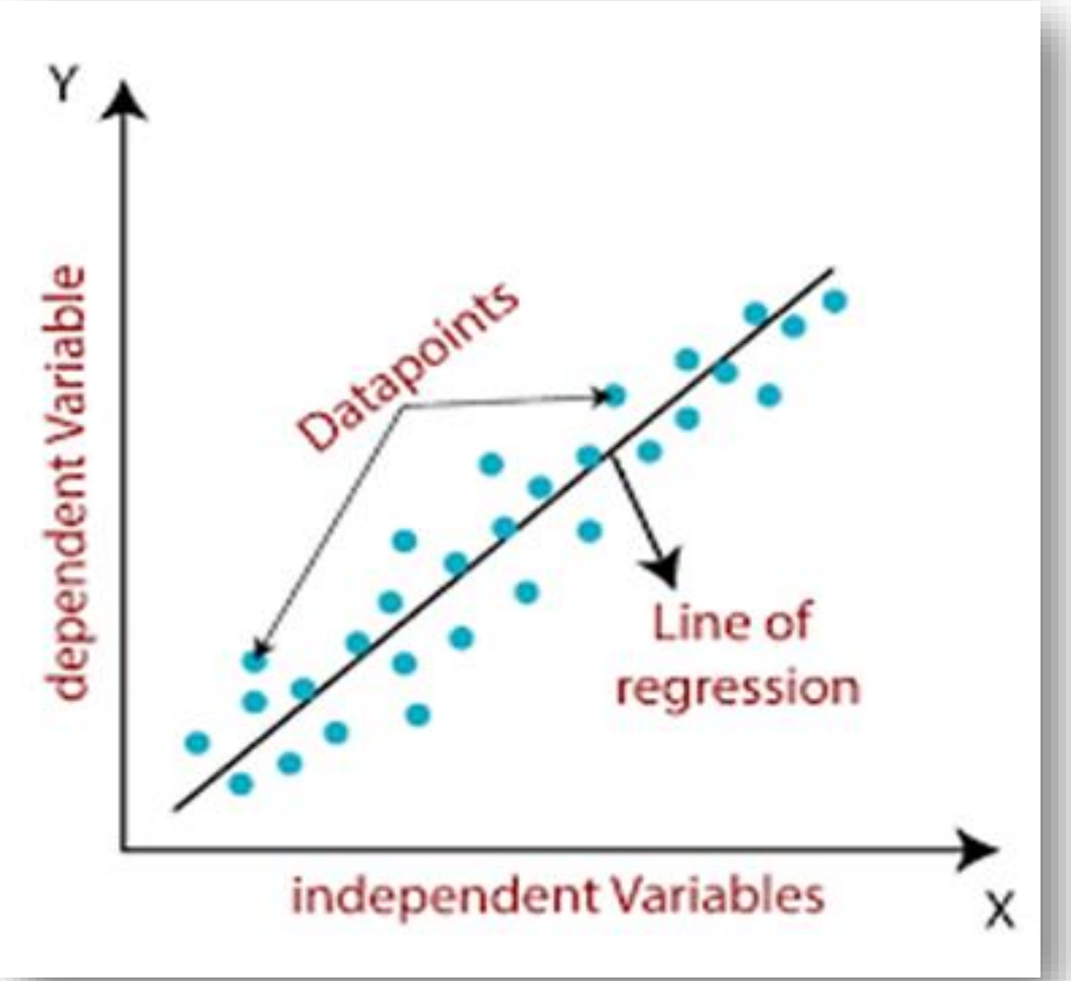
Linear Regression
Simplest form of regression, in which the predictor variables are assumed to have a linear relationship with the dependent variables

Assumptions: Input variables are assumed to be normally distributed and are not correlated with each other

Model Form: Y=ax + b
Y is the dependent variable and x is the independent variable; a is the slope and b is the y-intercept

Ordinary Least Squares (OLS) Method: *a* and *b* are selected through minimizing the sum of squares of residuals.

Residuals: the actual value minus the predicted value



Dataset Used for Regression Analysis

Software Sustainment dataset

Software Sustainment

The dataset includes variables collected for the analysis of **software sustainment data** for multiple DoD programs

Independent Variables

After analysis, the number of **Software Changes** and the **Duration** of the program are both influential in estimating effort

Dependent Variable

Effort is measured in total hours

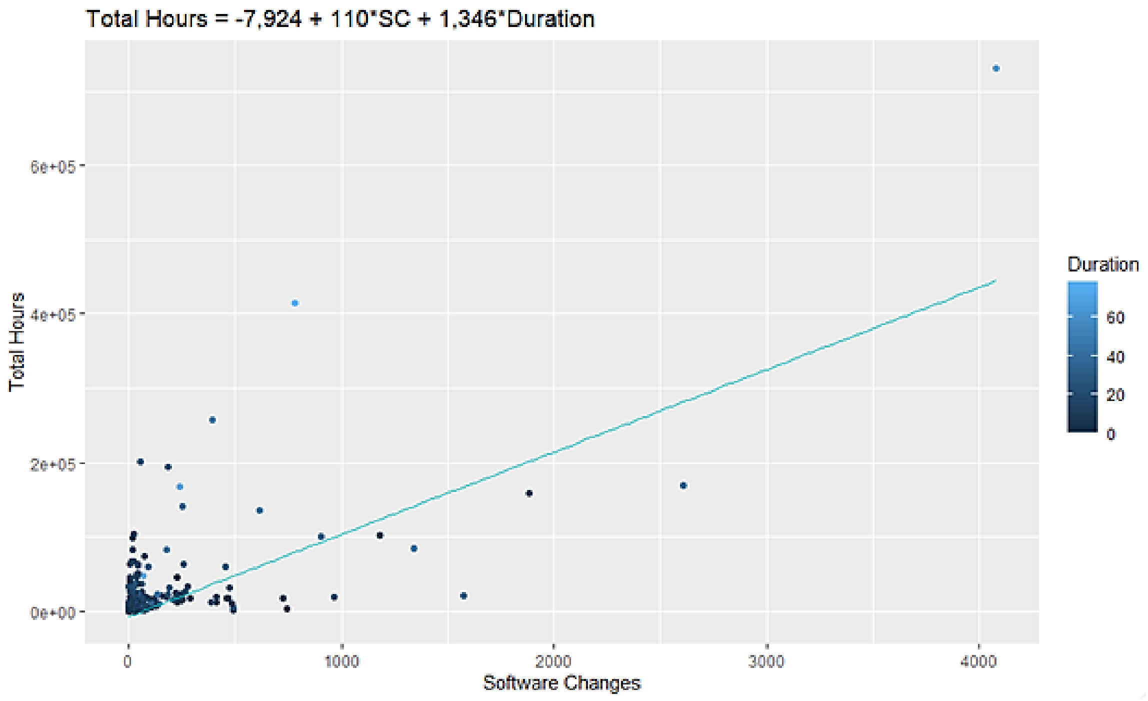
Multiple Linear Regression

Since two variables are included in the model that best estimates effort, the equation form for the model is

$$\hat{Y} = \beta_1 X_1 + \beta_2 X_2 + \beta_0 + \varepsilon$$

Software Sustainment

Multiple Linear Regression Model



Number of Datapoints

Original dataset was 316 datapoints. Model was trained with 221 datapoints while 95 datapoints were used the test the performance of the model

Purpose of Training and Testing

The partitioning of the dataset between training and test is done to determine how well the model predicts Total Hours based on new data that has not been included in the training or learning process of the linear regression algorithm

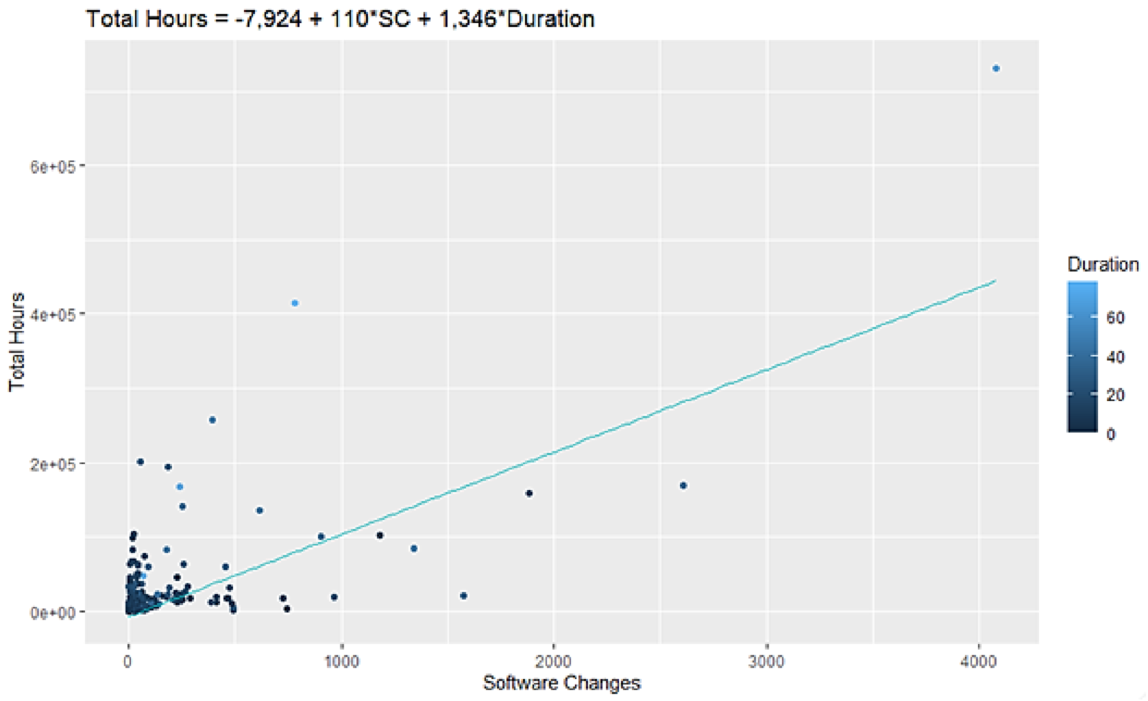
Dataset Trends

The majority of datapoints fit tightly to the line, but we observe several outliers on the plot

GALORATH

Software Sustainment

Multiple Linear Regression Model



Metric	Training	Test
R^2_{adj}	75%	75%
Root Mean Squared Error (RMSE)	26,928	48,089

Goodness-of-Fit Metrics

These metrics are calculated to be used to determine the statistical significance of regression models and compare multiple models

R^2_{Adjusted}

This metric tells us how much of the variability in Total Hours is explained by Software Changes and Duration

Root Mean Square Error (RMSE)

This metric is the standard deviation of the residuals and measures how spread out the residuals remain

Linear Regression

Improving linear models

01 **Regularization**

02 **Gradient Descent**

03 **Choose Best Model**

Linear regression models are known to be influenced by outliers. Regularization and gradient descent are two techniques that can be used to improve models. Regularization helps when coefficients are large, which can sometimes signify overfitting. Gradient descent can be used to optimize the coefficients, resulting in reduced error.

Regularization

Balancing bias and variance helps reduce overfitting

The What

Form of regression that constrains the coefficient estimates towards zero

The Why

Techniques reduce error by fitting a function on the given training set to avoid overfitting

The Goal

The goal is to create a simple model that reduces the risk of overfitting

Regularization: An Optimization Problem

$$Loss\ Function\ (SSE) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2$$

How is it done?

We want the estimated coefficients to generalize well on future data. This is achieved by regularizing the coefficients towards zero.

> **To find the smallest coefficients, you must minimize the loss function and shrink the coefficients towards zero**

GALORATH

Bias & Variance Tradeoff

Penalty

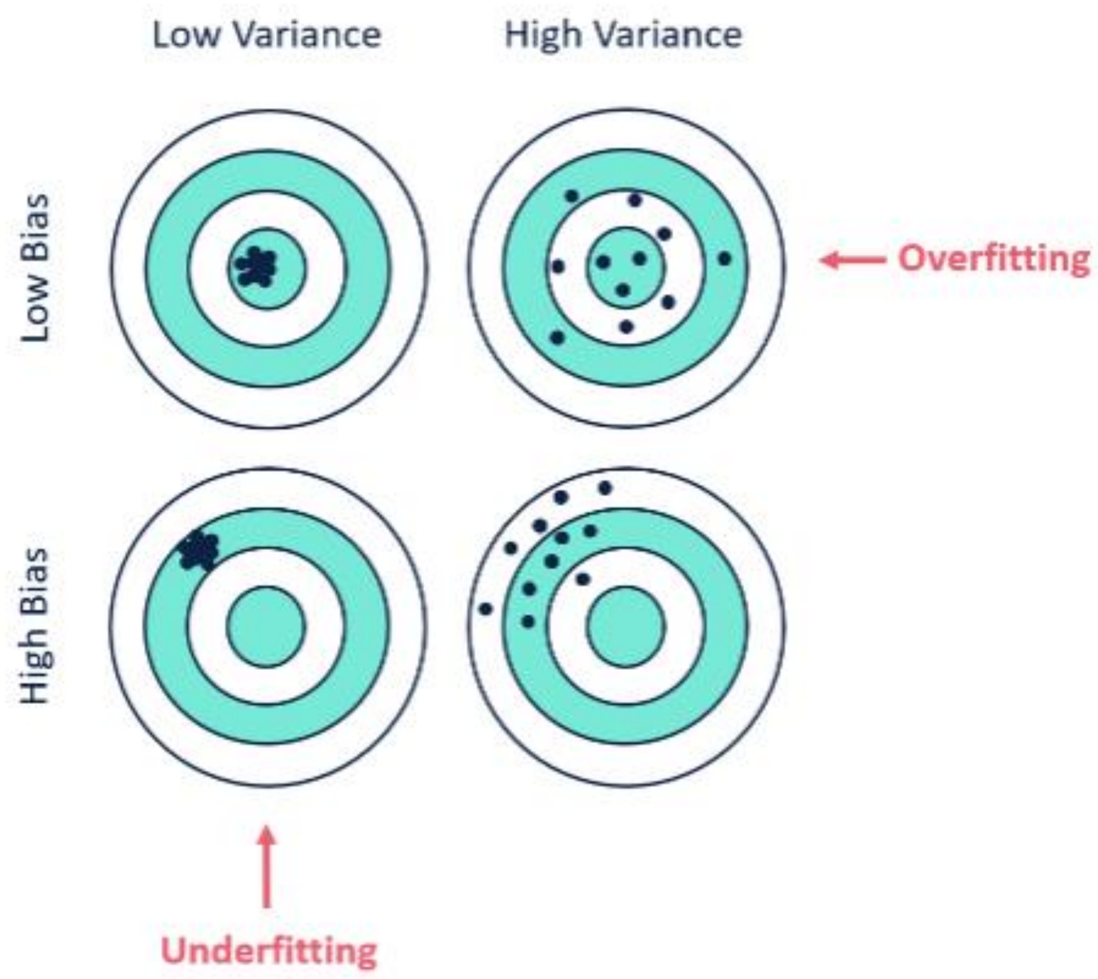
With the incorporation of a penalty, bias is introduced into the model buy reduces variance

Bias & Variance

Bias is the systematic tendency to overestimate or underestimate relative to the mean, while variance measures the dispersion of the estimate around the actual value

Irreducible Error/Noise

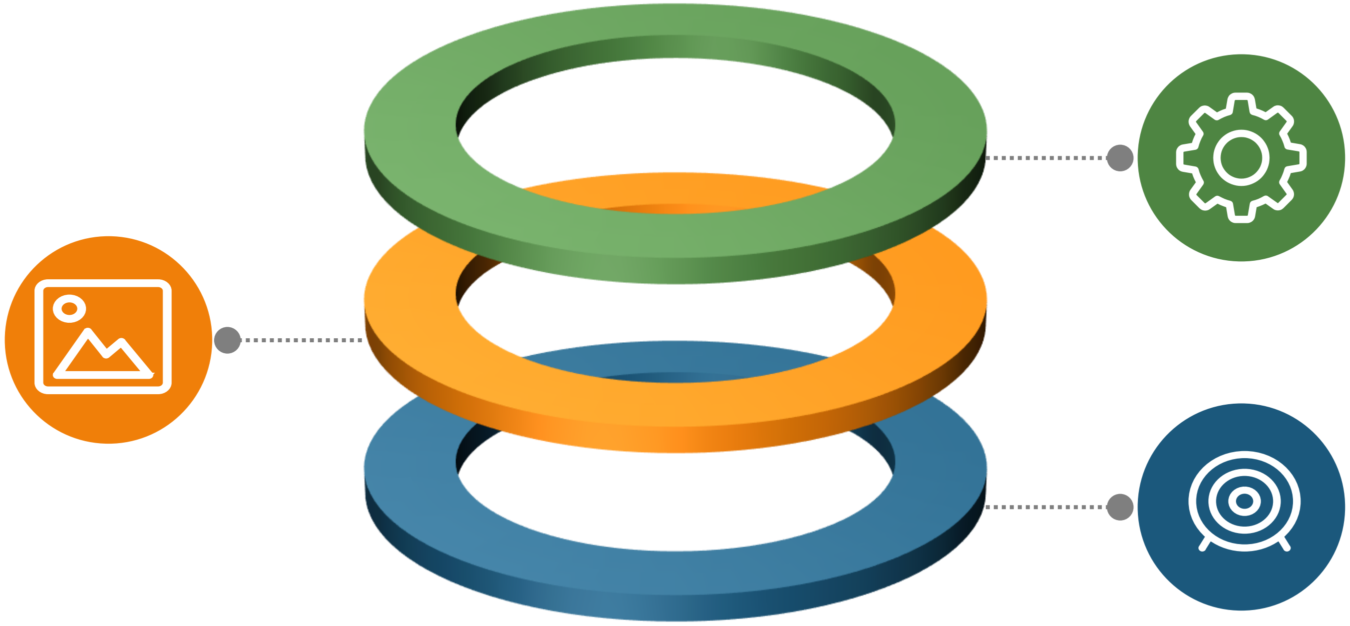
Models with high bias underfit the data, but with the addition of a minimal amount of bias, the variance is reduced. When there is high variance, the model tends to overfit the data



Regularization

Methods of Regularization

Regularization methods add a penalty term to constrain the slope parameters



Ridge Regression
Regression where the loss function is modified to minimize the complexity of the model

Lasso Regression
The loss function is modified to minimize the complexity of the model by limiting the sum of the absolute values of the model coefficients

Elastic Net Regression
Method that combines the properties of both ridge and lasso regression

Ridge Regression

Regularization Method #1

Bias

Ridge regression attempts to fit a new line to the data while introducing a small amount of bias

Minimize Loss Function

Ridge regression minimizes the loss function with an added function:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \sum \beta_1 x_i - \beta_0)^2 + \boldsymbol{\lambda} * \sum \boldsymbol{\beta_i^2}$$

Penalty

This method is an extension of OLS but adds a penalty to the method. Lambda, λ, determines the severity of this penalty. The value of λ can range from 0 to positive infinity.*

Choosing Penalty Value

The user must determine the value of λ that optimizes the regression results. This should be done by using cross-validation

* Note: When value of lambda is zero, the resulting model will be the same as the base case MLR OLS model

GALORATH

Lasso Regression

Regularization Method #2

Bias

Lasso also adds bias to the loss function

Minimize Loss Function

Lasso regression minimizes the loss function with an added function:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \sum \beta_1 x_i - \beta_0)^2 + \boldsymbol{\lambda} * \sum |\beta_i|$$

Penalty

An important difference between Lasso and Ridge is that as we increase the value of λ, the slope can shrink to zero. The value of λ can range from 0 to positive infinity.*

Choosing Penalty Value

Lasso seeks to discard useless variables from equation, so the models produced by Lasso will at times be simpler and easier to interpret

* Note: When value of lambda is zero, the resulting model will be the same as the base case MLR OLS model

What is Elastic-Net Regression

Regularization Method #3

Add Bias

Elastic-Net regression is a hybrid approach that combines the components of Ridge and Lasso regression

Minimize Loss Function

Elastic-Net regression minimizes the loss function with an added function:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \sum \beta_1 x_i - \beta_0)^2 + \lambda_1 * \sum \beta_i^2 + \lambda_2 * \sum |\beta_i|$$

Penalty

Cross-validation is used on different combinations of λ_1 and λ_2 to find the best values. The value of λ can range from 0 to positive infinity.*

Choosing Penalty Value

This hybrid approach groups and shrinks the parameters associated with the correlated variables or removes them if they are highly correlated. Elastic-Net tends to favor a more simplified model

* Note: When value of both lambdas are zero, the resulting model will be the same as the base case MLR OLS model

Gradient Descent

Optimization algorithm that approaches the least squared regression line using iterations

01

02

03

04

STEP 1

The algorithm begins the process by choosing a random line.

STEP 2

The parameters of the line are changed little by little to arrive at the best fit. Values are changed according to the gradient descent formula:

$$new\ intercept = old\ intercept\ - \alpha * \left(\frac{1}{m}\right) * \sum \left(h\left(x^i\right) - y^i\right)$$

$$new\ slope = old\ slope\ - \alpha * \left(\frac{1}{m}\right) * \sum \left(h\left(x^i\right) - y^i\right) * x^i$$

α is the learning rate and determines how large the rate of change should be on each iteration

STEP 3

The goal is to minimize the sum of squares of the error of cost

$$Cost = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)^2$$

STEP 4

The line with the smallest error is the line with the best fit

Regression Results Comparison

Software Sustainment Dataset

Method	Training		Test		Equation
	R^2_{adj}	RMSE	R^2_{adj}	RMSE	
Linear Regression	75%	26,928	75%	48,089	$Total\ Hours = -7{,}924 + 110 * SC + 1{,}346 * Duration$
Ridge Regression	75%	26,928	36%	48,089	$Total\ Hours = -7{,}924.07 + 110.99 * SC + 1{,}345.49 * Duration$
Lasso Regression	75%	26,928	36%	48,089	$Total\ Hours = -7{,}924.07 + 111 * SC + 1{,}345.49 * Duration$
Elastic Net Regression	75%	26,930	36%	48,056	$Total\ Hours = -7{,}924.07 + 111 * SC + 1{,}345.49 * Duration$
Gradient Descent	70%	42,995	21%	59,239	$Total\ Hours = 0.04 + 46.07 * SC + 1.23 * Duration$

Linear Regression prevails as the Best Model!

Conclusion

Linear regression is still a powerful Machine Learning technique that is oftentimes the best model!

Characteristics of a linear dataset include a limited range in either the dependent and/or independent variable

A good rule of thumb is that when the dependent and the independent variable data points being modeled are all within an order of magnitude of one another, the relationship is likely to be linear

Seek ways to improve your linear model by using regularization and gradient decent

See paper for additional ML techniques such as Bayesian methods to potentially improve regression results

Questions?