Linear Regression

Created Date: 2022-08-15

> Metadata 🗃️
>
> - Title: Linear Regression
> - Author: Andrew Jones
> - Reference: Data Science Infinity

> Links & Tags 🔗
>
> - Index: Course Note Index
> - Atomic Tag: #datascience
> - Subatomic Tags: #machinelearning #linearregression

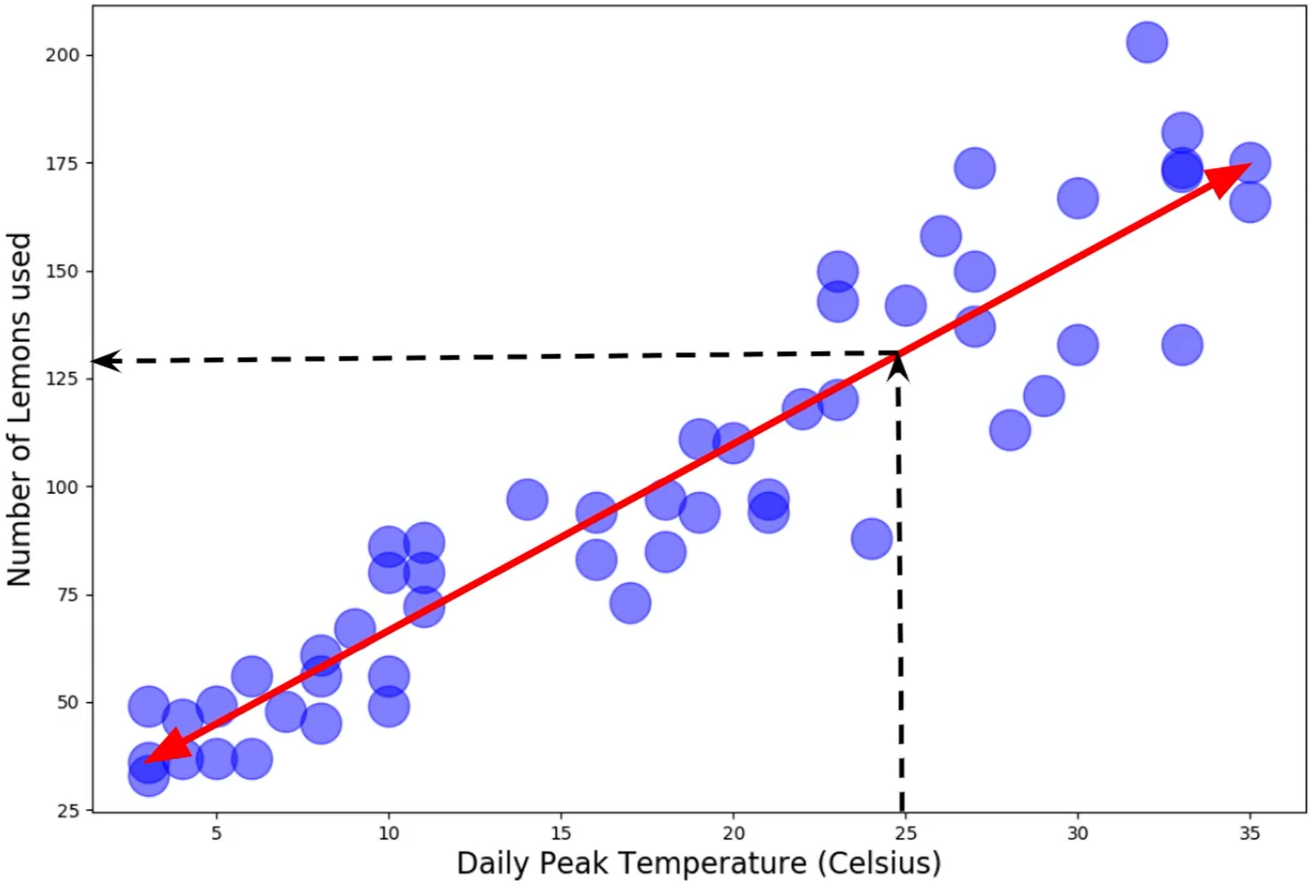
High-Level Overview

Jupyter Notebook: Basic Linear Regression Template

> Linear Regressions is a model that fits the best possible straight line through the data as a way to generalize the relationship between the output variable and either a single input variable (simple linear regression) or multiple input variables (multiple linear regression).

- The regression equation can be used to plug in input variables, and predict the associated output variable
- The R^2 measure (Coefficient of Determination) assesses the accuracy, or goodness of fit, of the regression line
  - 0 = No Relationship

- 1 = Perfect Relationship
- Explains how much of the variation in the output variable (y) is explained by the input data (x)



Advanced Theory

Jupyter Notebook: Advanced Linear Regression Template

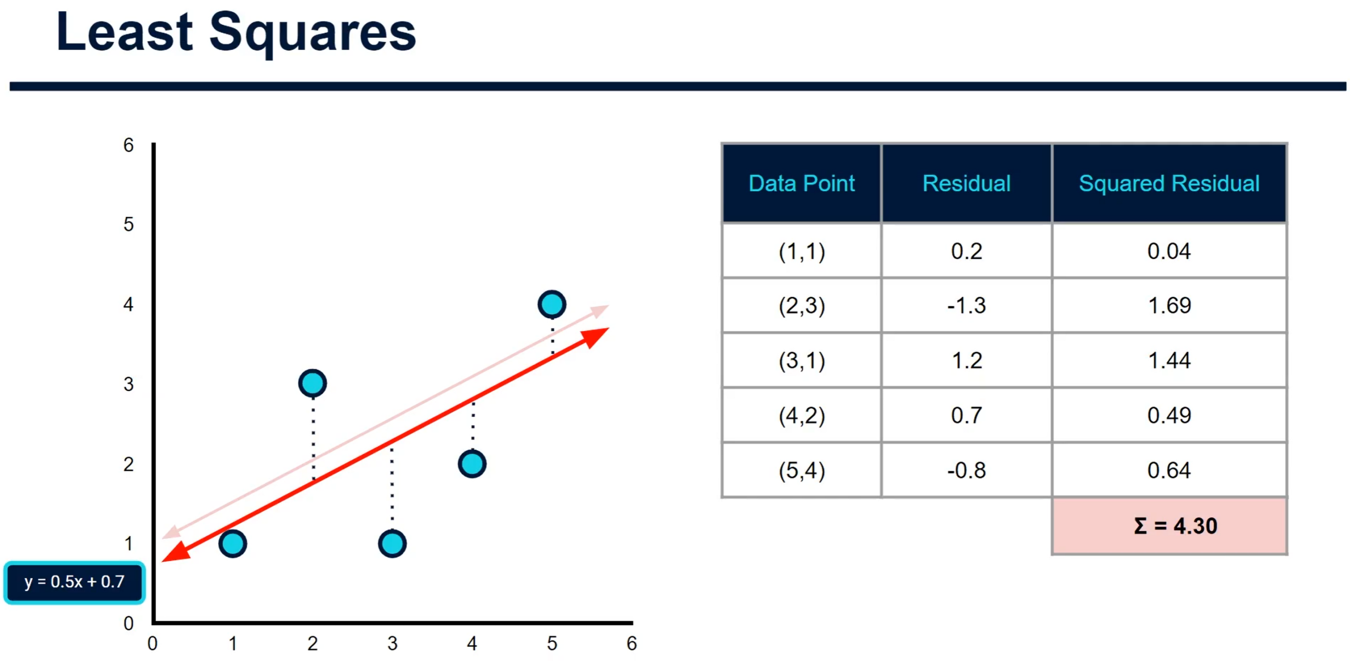
Regression Line Formula

- Line of best fit equation
  - $y = bx + c$
  - y Output Value (Predicted)
  - b Slope of the Line
  - x Input Value
  - c Y-Intercept

Finding the Best Line Using Least Squares

> Least Squares is an approach used in regression problems to approximate a "line of best fit" by minimizing the sum of the squared residuals.

- Residual is the difference between the actual data point and the regression line (y)
- Residuals are squared in the least squares method to convert negative numbers to positive numbers
- The goodness of fit score is the sum of the squared residuals, and we want to find a line with lowest goodness of fit score possible
  - Finding a line that minimizes the sum of the squared residuals is the "least squares"
- If we plotted the sum of squared residuals for all the possible regression lines, we would get a u-shaped curve
  - We can then use gradient descent (differentiation) to determine the line with the least squares (realistically this is found with a program)

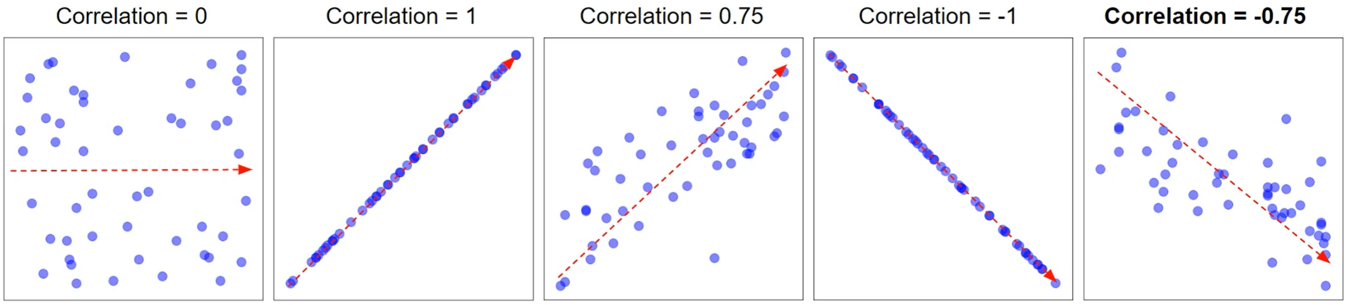


Least Squares

Data Point	Residual	Squared Residual
(1,1)	0.2	0.04
(2,3)	-1.3	1.69
(3,1)	1.2	1.44
(4,2)	0.7	0.49
(5,4)	-0.8	0.64
		Σ = 4.30

Evaluating Model Fit Using R-Squared

> R = Correlation (Expresses the strength and direction of a relationship between two variables).
>
> - *AKA Coefficient of Correlation*
> - *Scale between -1 and 1*



Correlation = 0

Correlation = 1

Correlation = 0.75

Correlation = -1

Correlation = -0.75

> R^2 Shows the percentage of variance in our output variable y, that is being explained by our input variables (x)
>
> - *AKA Coefficient of Determination*
> - *Scale between 0 and 1*

- Squaring R converts negative numbers into positive numbers
  - Since a 0.75 and a -0.75 R score has the same correlation, just in a different direction, we can safely square them to have the same R^2 score for both
    - $0.75^2 = 0.56$
    - $-0.75^2 = 0.56$
  - This allows us to assess the model fit equally with negative and positive correlations
  - R^2 should be calculated on the training and test data to assess over-fitting
    - Results should be similar for each set of data
    - A higher score on the training data indicates over-fitting

$R^2 = \frac{SSR[MeanLine]-SSR[RegressionLine]}{SSE[MeanLine]}$

- Mean Line is a horizontal line where $y = mean$
- Regression Line is the "line of best fit" determined by the least squares method

$R^2 = 1 - \frac{RSS}{TSS}$

- RSS Sum of Squared Residuals
- TSS Total Sum of Squares

Where;

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

- y_i Sample Value
- $f(x_i)$ Predicted Value
- n Number of Observations

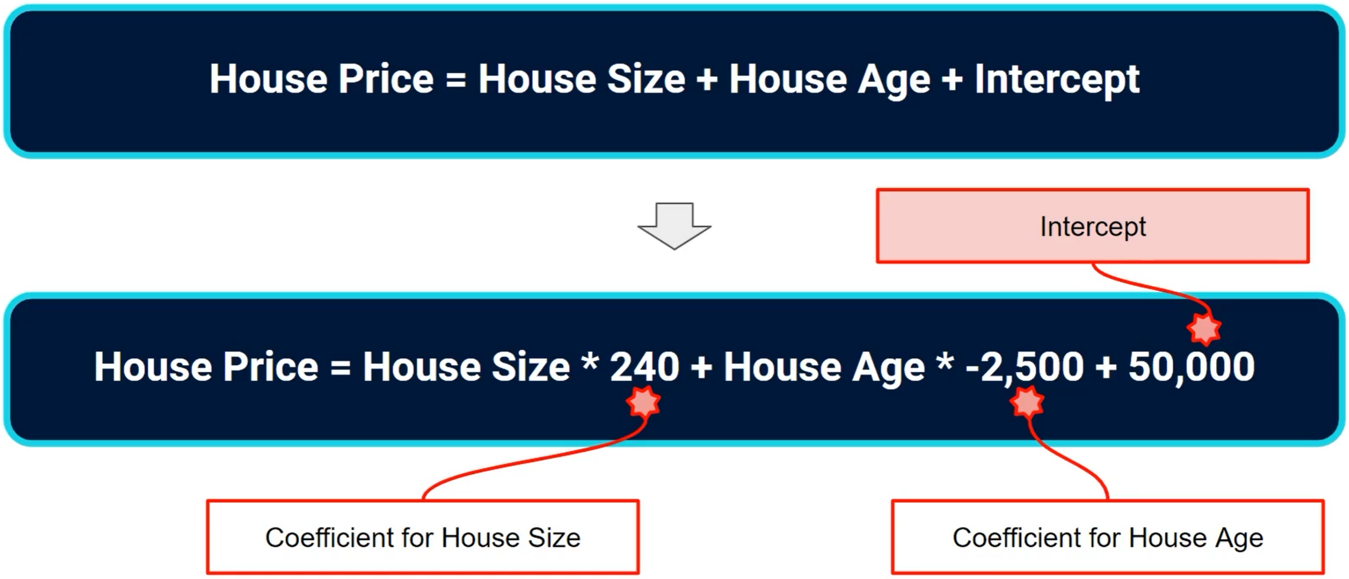
$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- y_i Sample Value
- $\bar{y}$ Mean Sample Value
- n Number of Observations

Multiple Regression

- Instead of a line of best fit, we use a plane of best fit
  - $y = b_1x_1 + b_2x_2 + b_nx_n + c$
  - y Output Value (Predicted)
  - b Coefficient of x (How much will y change for a one-unit increase in x)
  - x Input Value

- c Y-Intercept



House Price = House Size + House Age + Intercept

House Price = House Size * 240 + House Age * -2,500 + 50,000

> *The house size coefficient is the expected increase of house price for every additional square foot of house size*

Adjusted R-Squared

- Calculating R-Squared for multiple input variables is roughly done the same way, we just sum the R^2 for each input variables together
- This can inflate the R-Squared value as every input variable added to a model increases the R^2 value and never decreases it, even if the relationships is by chance

> Adjusted R-Squared compensates for the addition of input variables and only increases if the variable improves the model above what would be obtained by probability

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(n-1)}{(n-p-1)}$$

- n Number of Data Points

- p Number of Input Variables

Understanding P-Values

> A p-value helps us assess whether the results of some finding or test we have conducted are either, likely to be common, or likely to be strange. This helps us decide how we should proceed (such as should we consider this input variable to have a statically significant effect on the output variable).

- Summary statistics for a multiple linear regression will include the input variable, coefficient, and p-value for each coefficient
- The p-value tests the coefficient for each input variable based on the null hypothesis that the coefficient is actually equal to 0 (no relationship)
  - $H_o : Coefficient = 0$ *There is no relationship between the input and output variables*
  - $H_a : Coefficient > 0$ *There is a relationship between the input and output variables*
    - A low p-value indicates we can reject the null hypothesis, meaning there is confidence in a relationship
    - A high p-value indicates we accept the null hypothesis, meaning there is not enough evidence to suggest the input and output variable have a relationship
      - The common threshold is a p-value of 0.05 ($\alpha = 0.05$) which is considered the acceptance criteria
        - Reject null when $\alpha < 0.05$ (there is a relationships)
        - Accept null when $\alpha > 0.05$ (there is no relationship)
        - The confidence level is $1 - \alpha$ so in the example above, $CL = 0.95$ (or 95%)

	Coefficient	p-value
Intercept	50,000	0.00
House Size	240	0.01
House Age	-2,500	0.02
House Number on Street	14	0.19