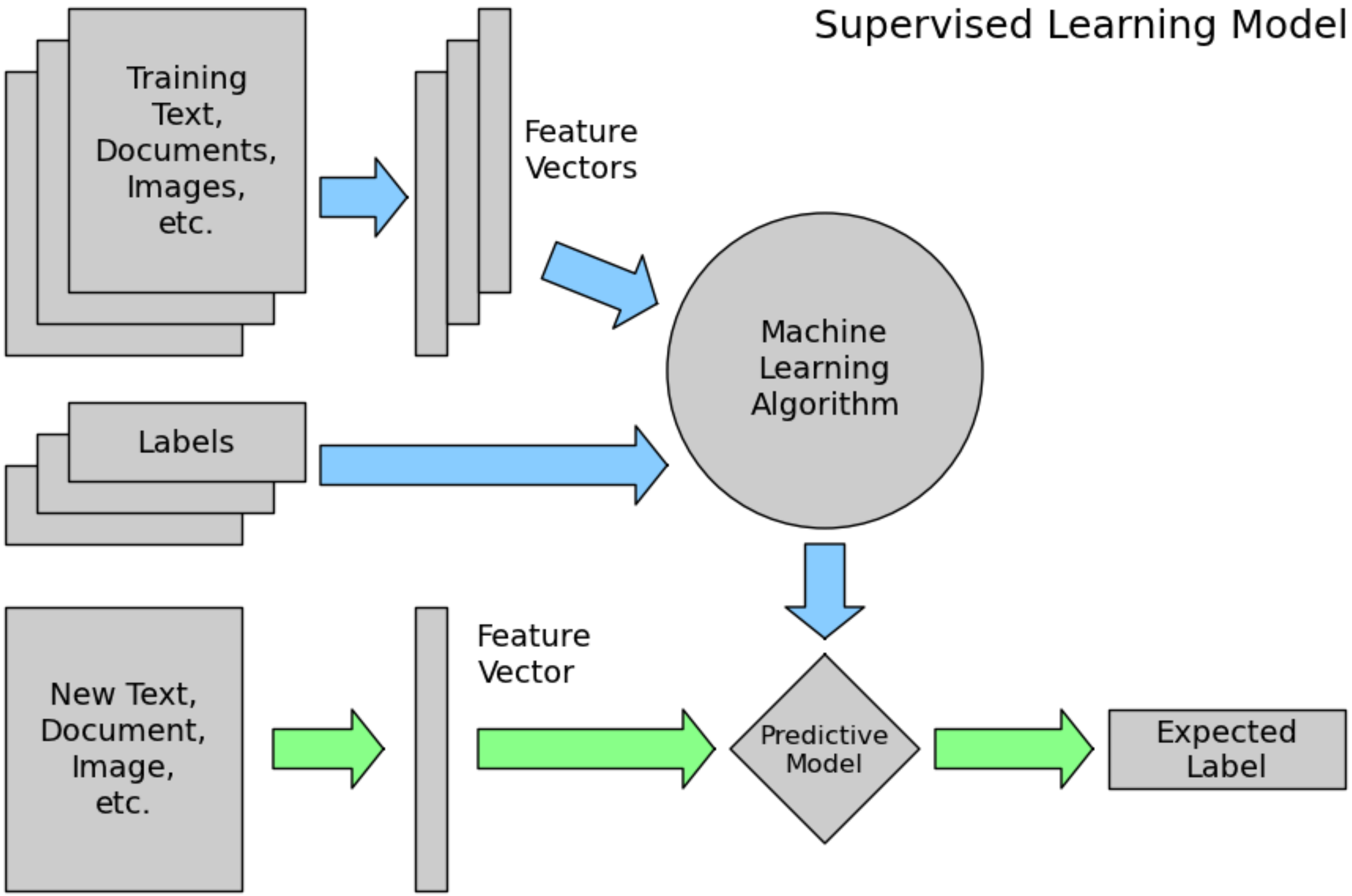
GENERAL ASSEMBLY

INTRODUCTION TO REGRESSION ANALYSIS

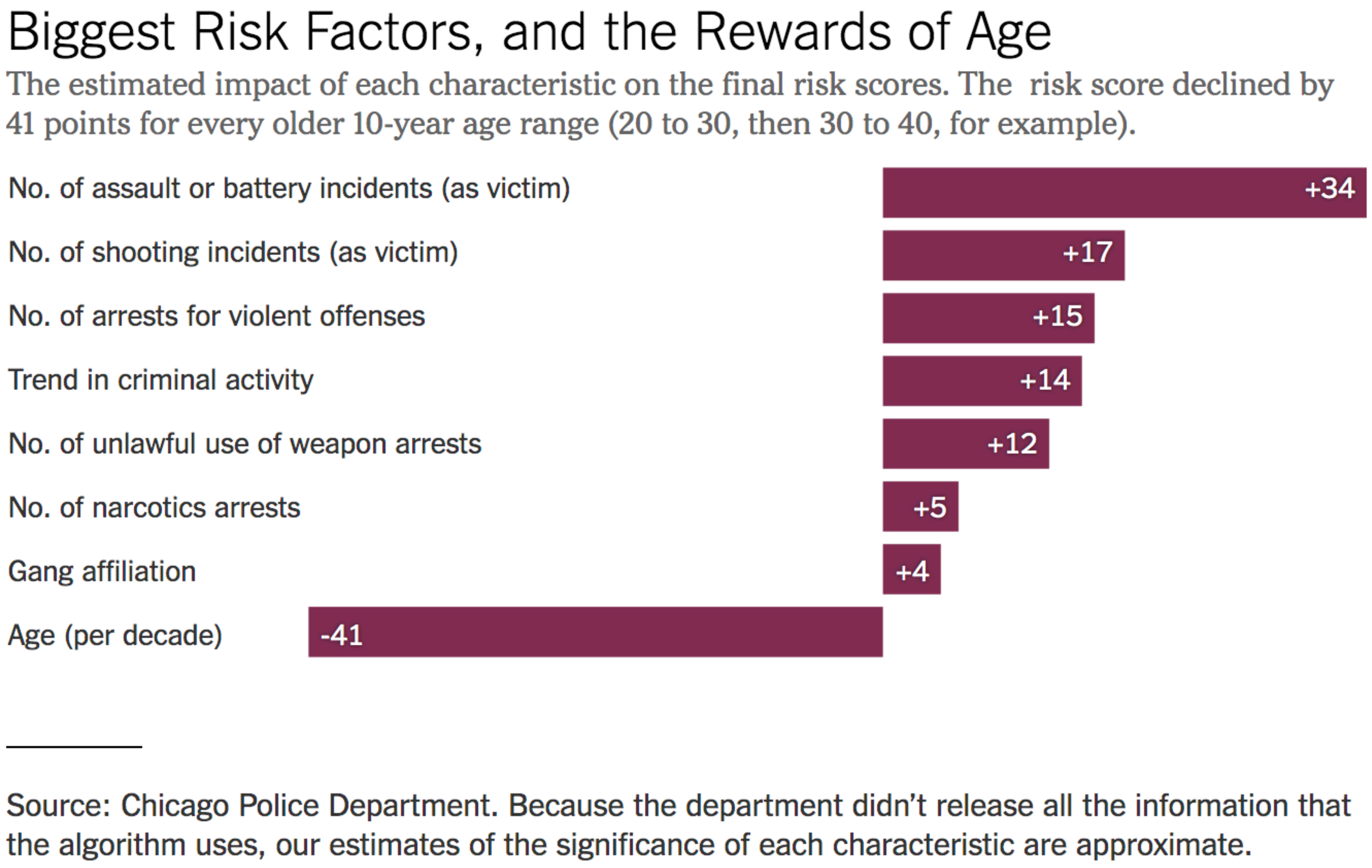
INTRODUCTION

SIMPLE LINEAR REGRESSION

MACHINE LEARNING WORKFLOW


Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

LINEAR REGRESSION EXAMPLE



Biggest Risk Factors, and the Rewards of Age

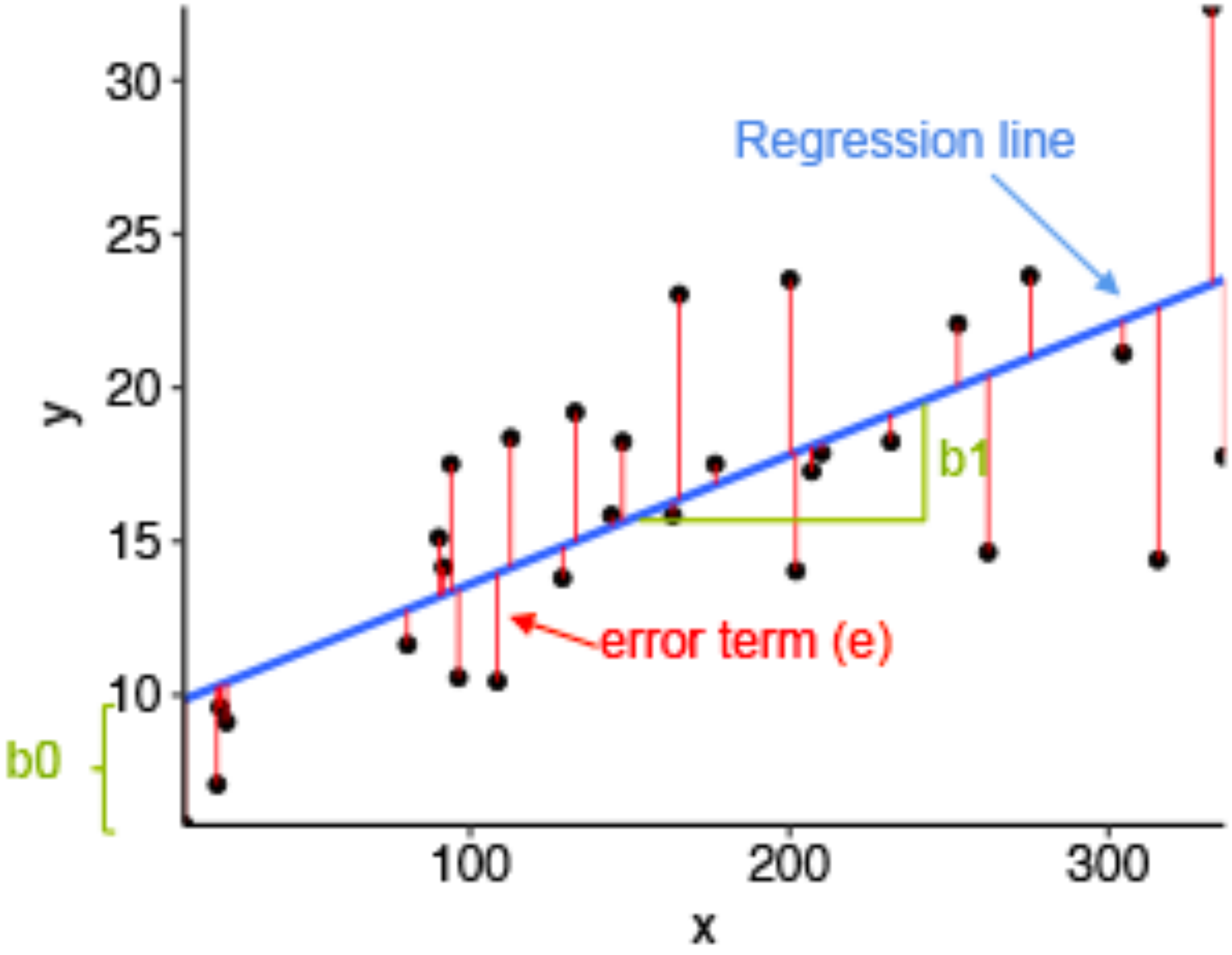
The estimated impact of each characteristic on the final risk scores. The risk score declined by 41 points for every older 10-year age range (20 to 30, then 30 to 40, for example).

Source: Chicago Police Department. Because the department didn’t release all the information that the algorithm uses, our estimates of the significance of each characteristic are approximate.

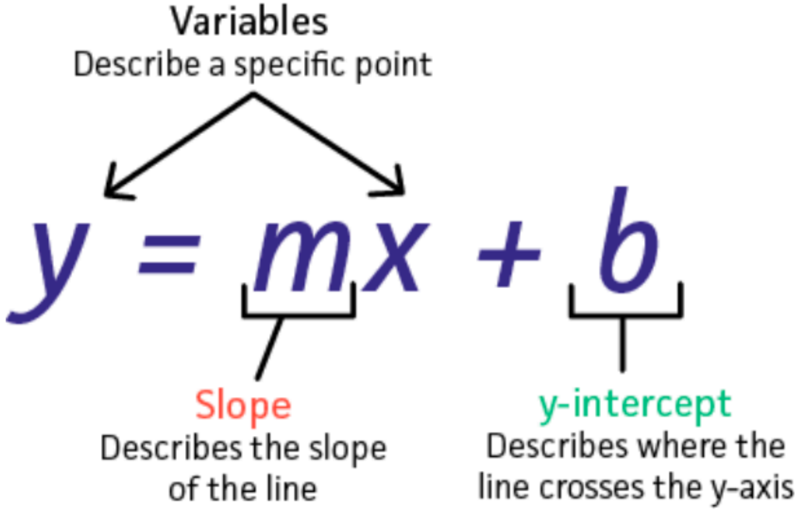
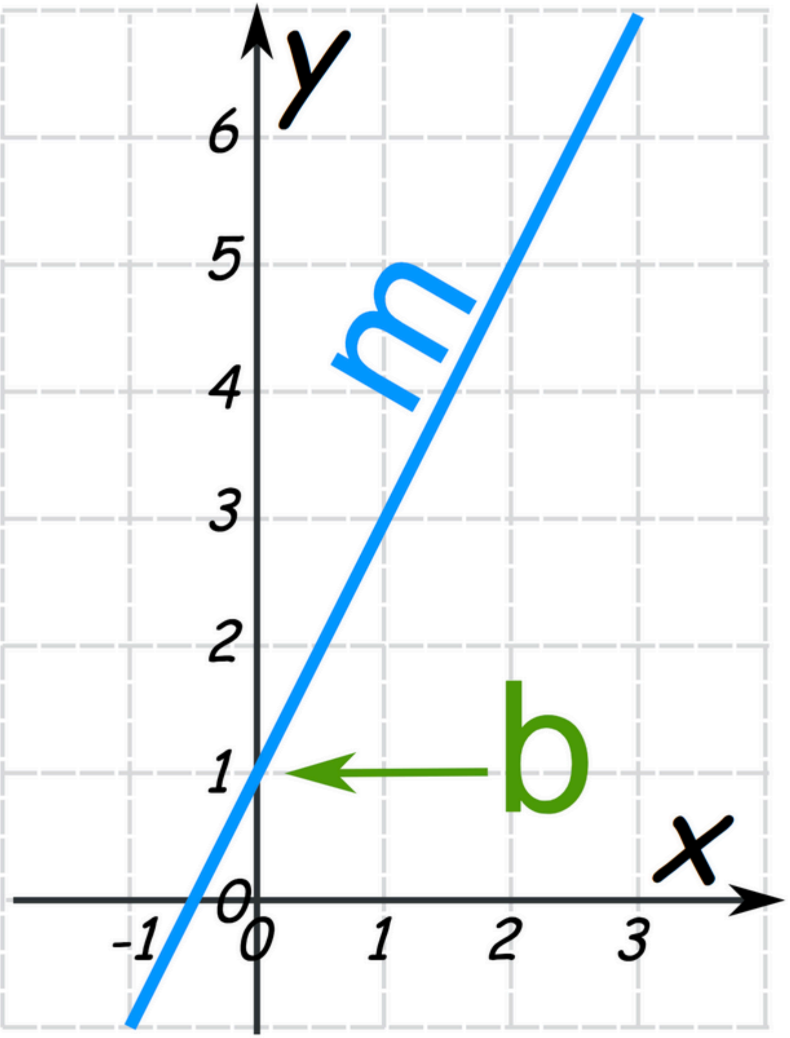
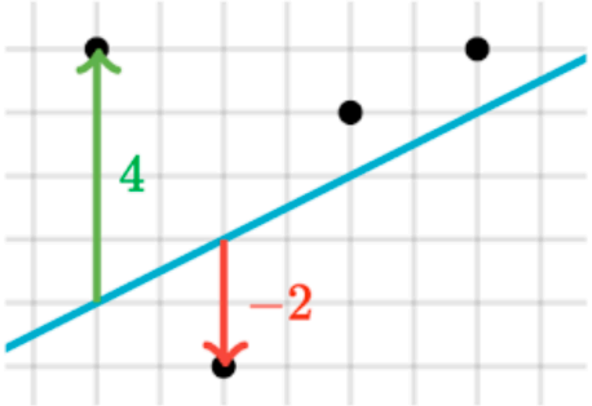
source: https://www.nytimes.com/2017/06/13/upshot/what-an-algorithm-reveals-about-life-on-chicagos-high-risk-list.html

LINEAR REGRESSION COMPONENTS

1. Predictor (x)
2. Outcome (y)
3. Line of best fit
4. Error term (e)
5. Coefficient (b1)
6. Intercept (b0)

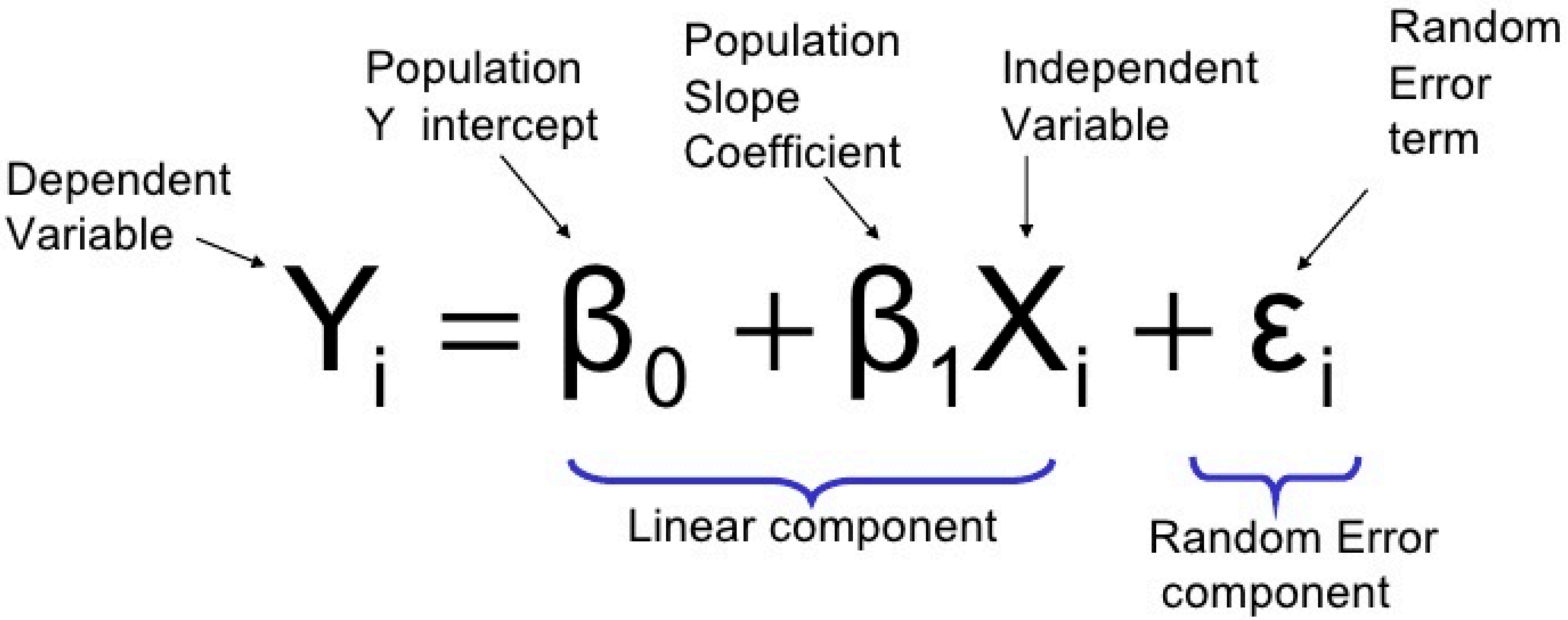


LINEAR REGRESSION

Equation	Slope of line	Residuals
 Variables Describe a specific point $y = mx + b$ Slope Describes the slope of the line y-intercept Describes where the line crosses the y-axis		

LINEAR REGRESSION

- For any one point in the dataset, its y-value is predicted as follows:



Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

MULTIPLE LINEAR REGRESSION

Simple Linear Regression

$$y = b_0 + b_1{}^*x_1$$

Multiple Linear Regression

Dependent variable (DV) Independent variables (IVs)

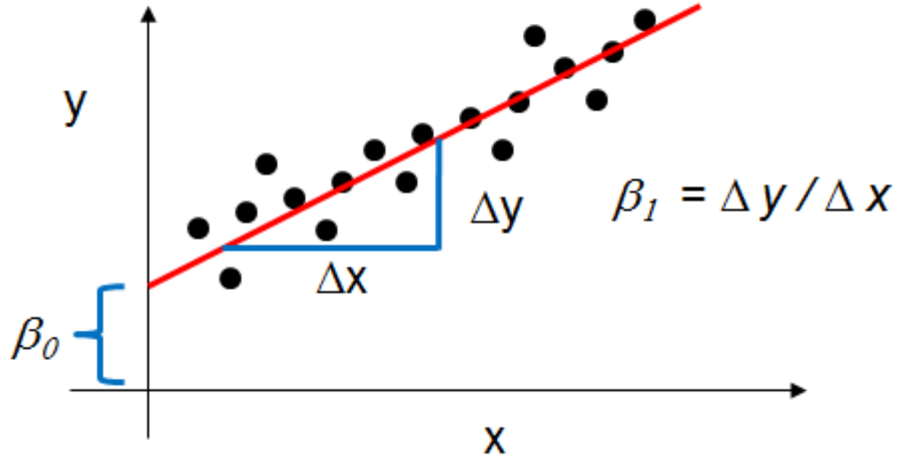
$$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2 + \dots + b_n{}^*x_n$$

SIMPLE LINEAR REGRESSION

- Def: Explanation of a continuous variable given a series of independent variables

- The simplest version is just a line of best fit: y = mx + b

- Explain the relationship between **x** and **y** the starting point **b** and the power in



SIMPLE LINEAR REGRESSION

- However, linear regression uses linear algebra to explain the relationship between *multiple* x's and y.

- The more sophisticated version: y = beta * X + alpha (+ error)

- Explain the relationship between the matrix **X** and a dependent vector **y** using a y-intercept **alpha** and the relative coefficients **beta**.

MULTIPLE REGRESSION ANALYSIS

- Simple linear regression with one variable can explain some variance, but using multiple variables can be much more powerful.
- We want our multiple variables to be mostly independent to avoid multicollinearity.
- Multicollinearity, when two or more variables in a regression are highly correlated, can cause problems with the model.

RESIDUALS: MULTIPLE LINEAR REGRESSION

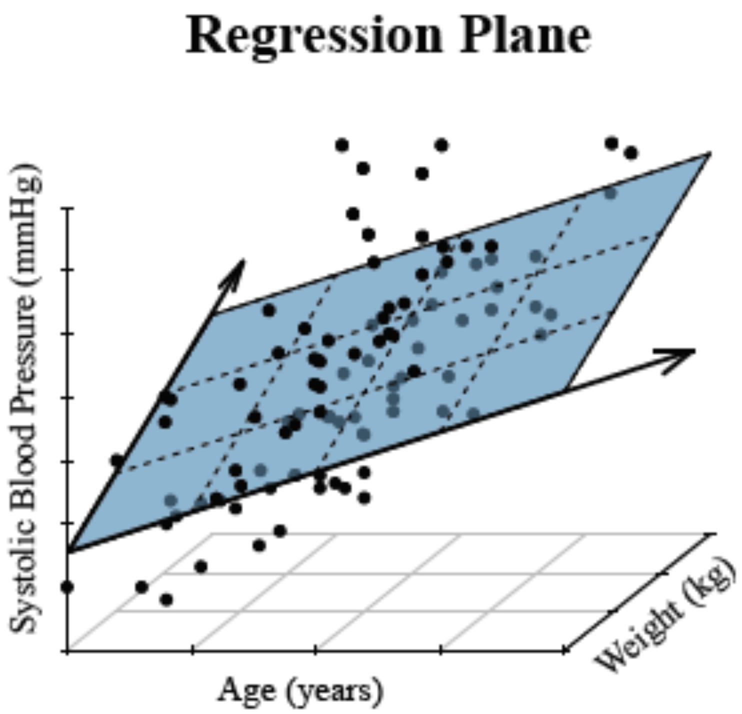

Regression Plane

Figure 2.25: Systolic blood pressure linearly increases with age, but also with bodyweight. A line in two directions forms a plane.

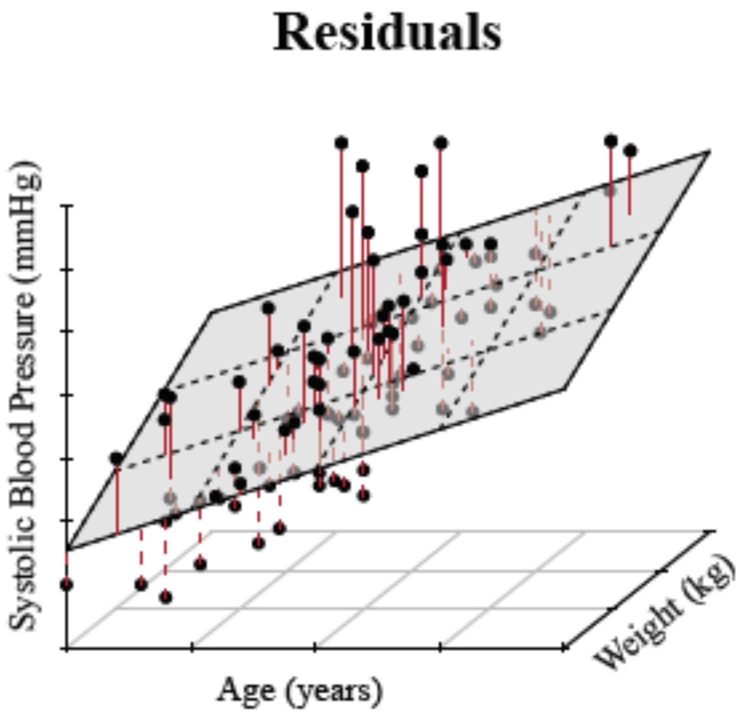

Residuals

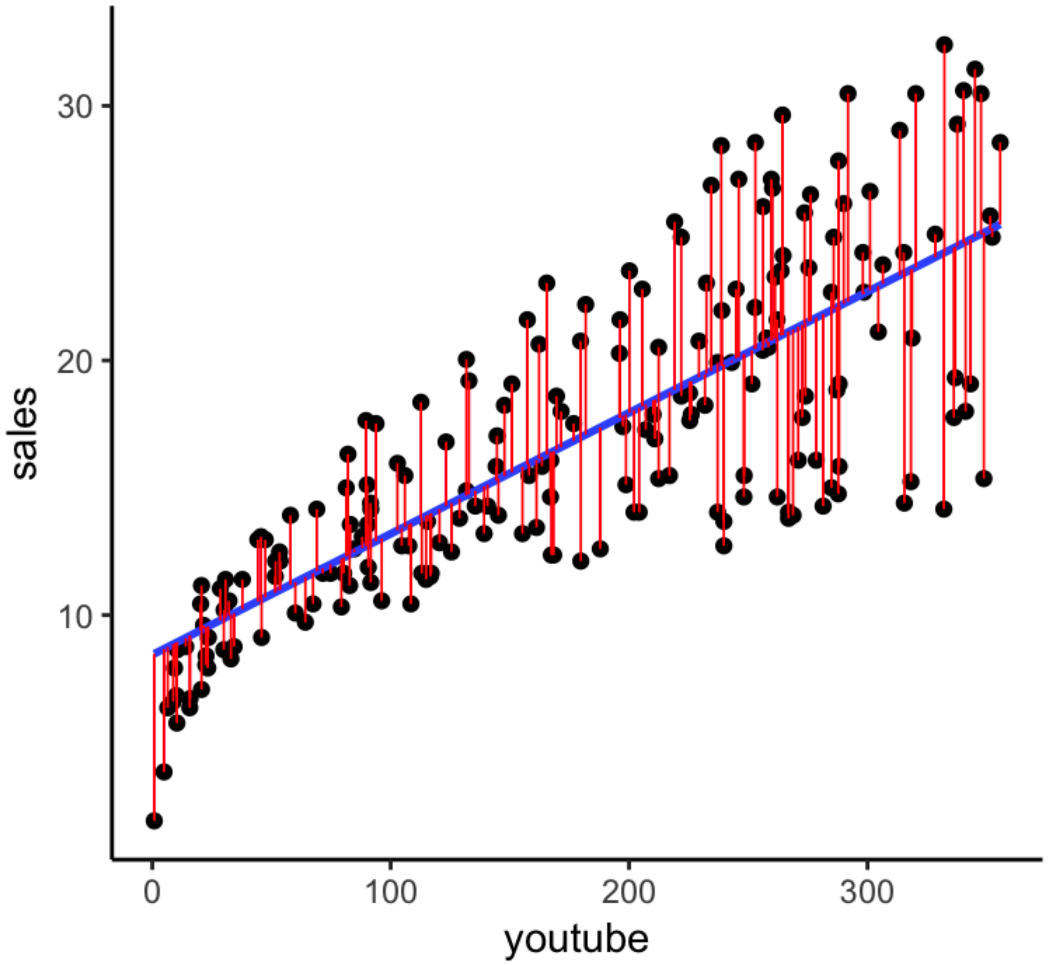
Figure 2.26: The residuals of figure 2.25 are the vertical distances to the plane. Negative residuals are indicated by dashed linepieces.

LINEAR REGRESSION

RESIDUALS

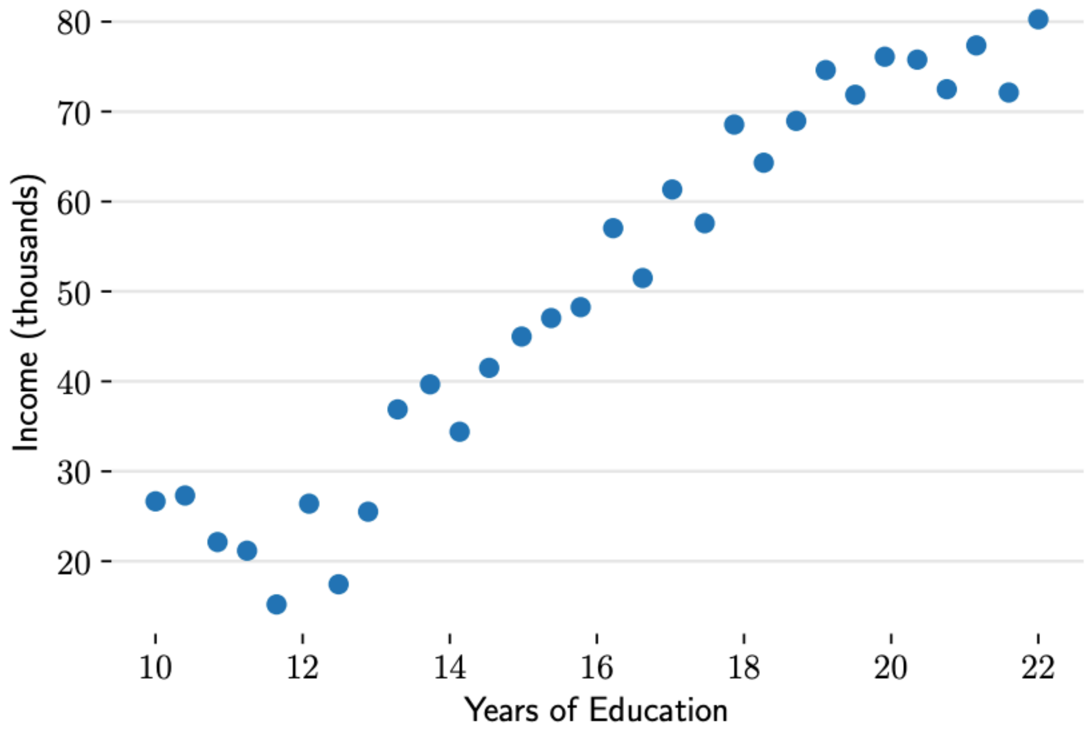
RESIDUALS

- A residual is the difference between the observed y-value (from scatter plot) and the predicted y-value (from regression equation line).
- It is the vertical distance from the actual plotted point to the point on the regression line.
- You can think of a residual as how far the data "fall" from the regression line.
- Also known as “error term”

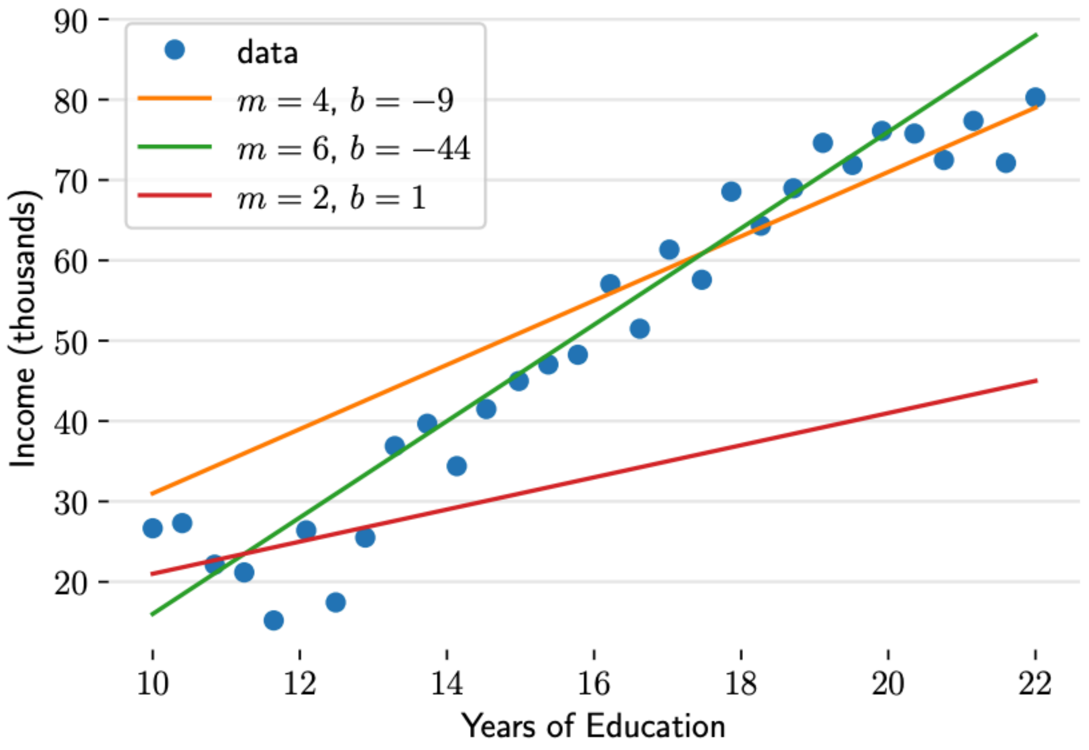


LOSS FUNCTION

How does scikitlearn choose the “best” line from all possible lines? By means of a **loss function**. Loss functions formalize how bad it is to produce output y^ when the truth was y. The “loss” is the cost of error.



(a) Raw Income data relating years of education to annual income.

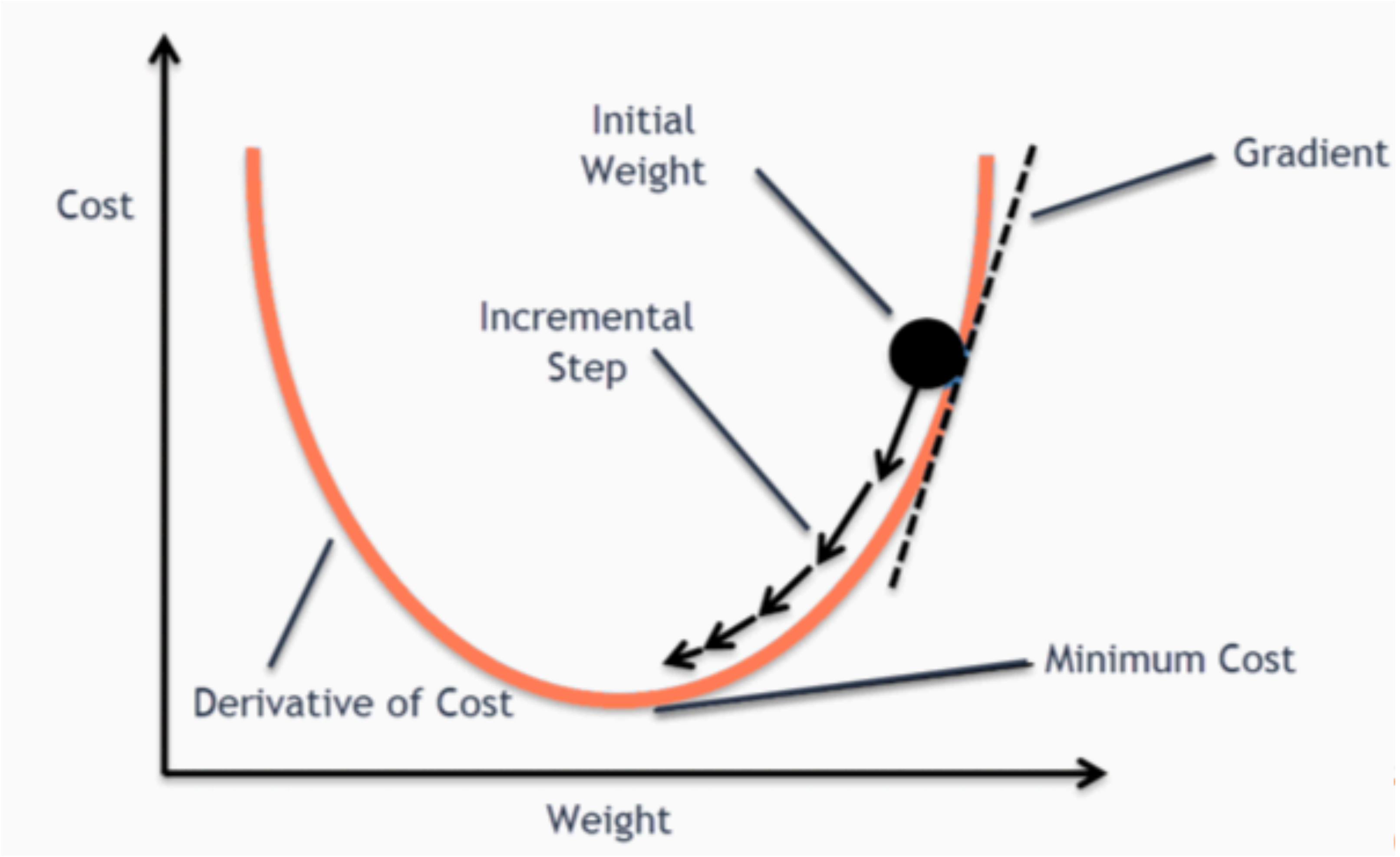


(b) Some potential linear fits to the Income data with the parameterization $y = mx + b$.

GRADIENT DESCENT

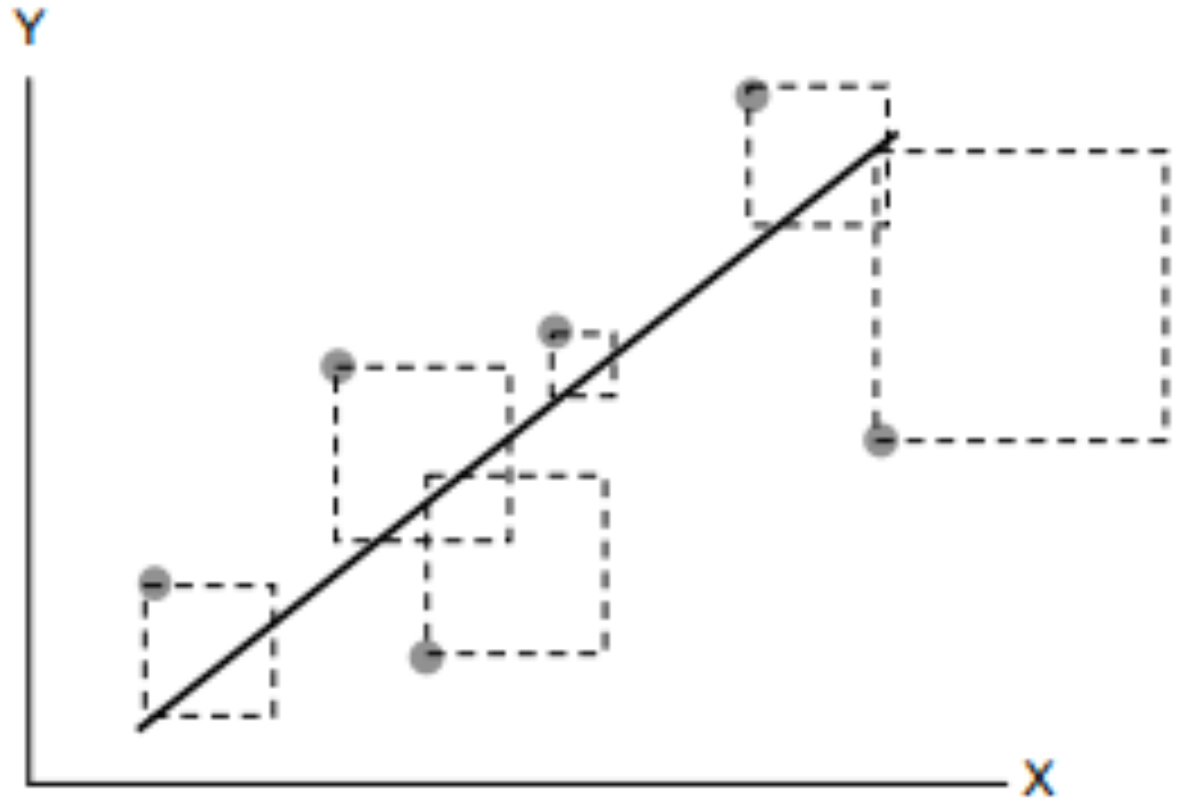
- How does scikitlearn determine the line of best fit?

The **gradient** of H at a point is a plane vector pointing in the direction of the steepest slope or grade at that point. The steepness of the slope at that point is given by the magnitude of the **gradient** vector.

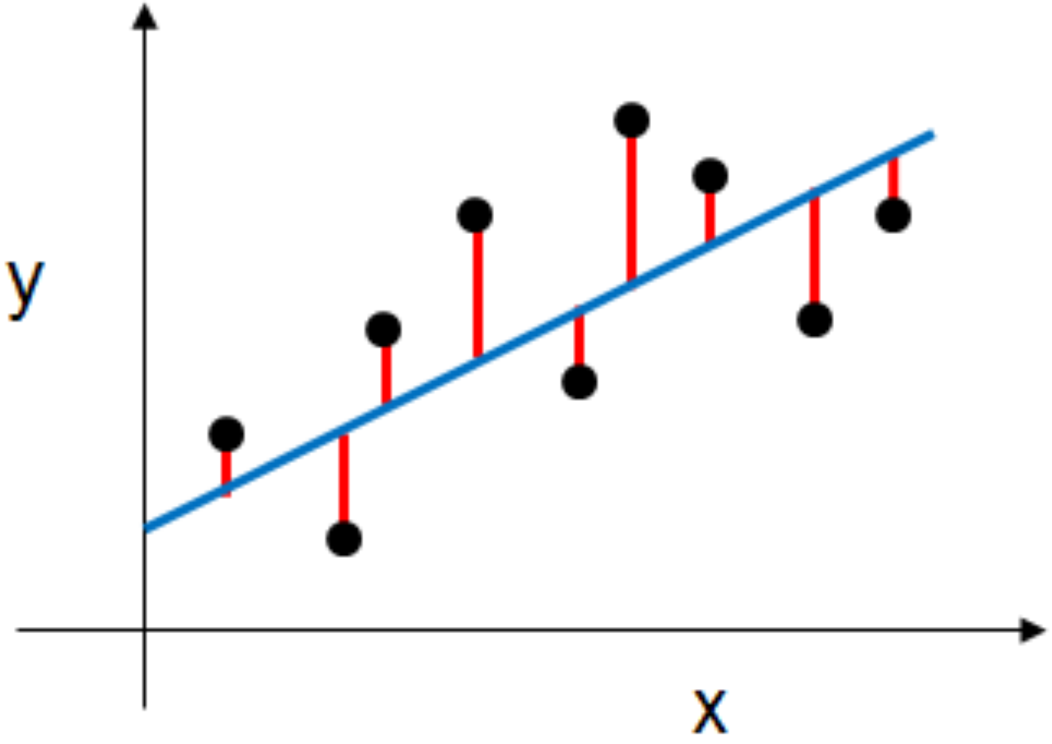


LOSS FUNCTION

residual sum of squares (RSS): the sum of the squares of residuals. It is a meas of the discrepancy between the data and an estimation model, such as a linear regression. A small RSS indicates a tight fit of the model to the data.



Least Squares IS THE LOSS FUNCTION FOR LINEAR REGRESSION



Model Prediction

$$SS_{residuals} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Observed Result

LOSS FUNCTION

In a model with a single explanatory variable, RSS is given by:

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

where:

- yi is the ith value of the variable to be predicted,
- xi is the ith value of the explanatory variable, and
- f(x) is the predicted value of yi

LINEAR REGRESSION

ASSUMPTIONS

ASSUMPTIONS OF LINEAR REGRESSION

- Linear regression works **best** when:
    - The data is normally distributed (but doesn’t have to be)
    - X’s significantly explain y (have low p-values)
    - X’s are independent of each other (low multicollinearity)
    - Resulting values pass linear assumption (depends upon problem)
- If data is not normally distributed, we could introduce *bias*.

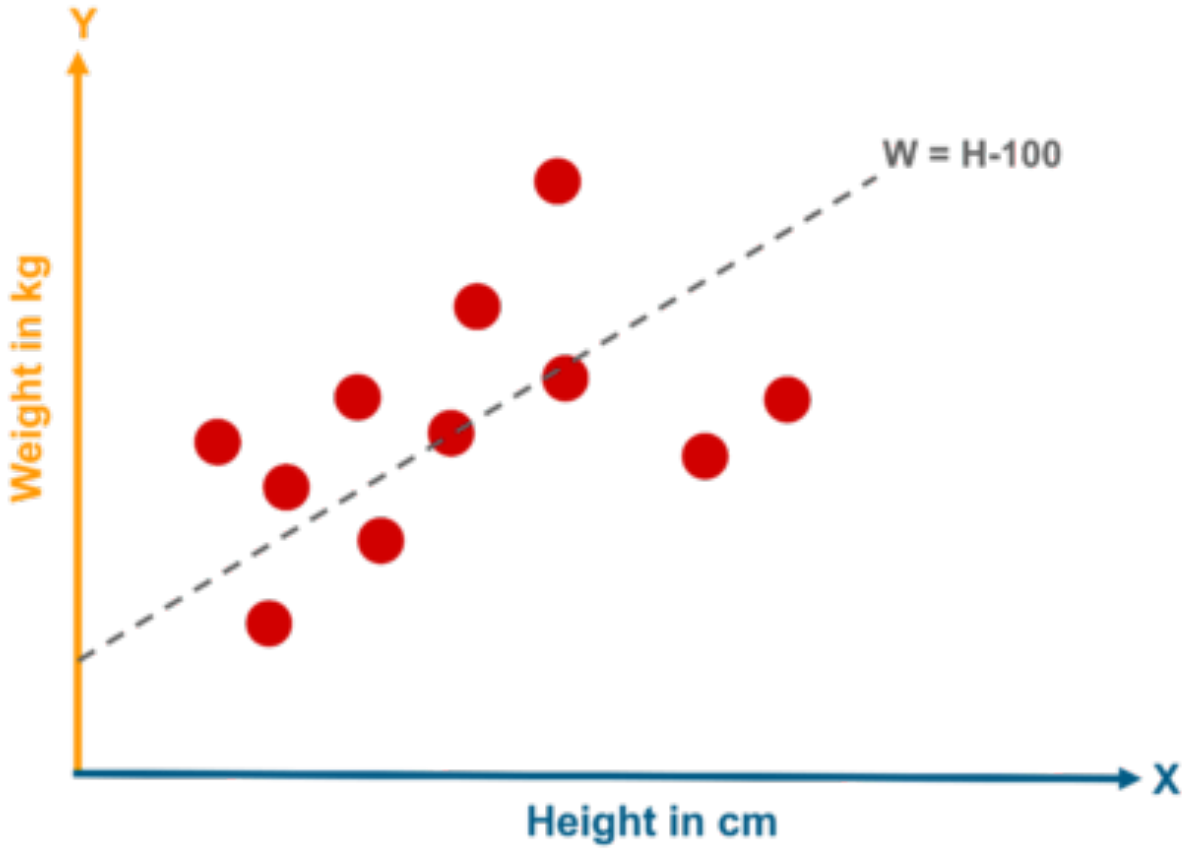
LINEAR REGRESSION ASSUMPTIONS

1. continuous variables
2. linear relationship
3. no significant outliers
4. independence of observations
5. homoscedasticity
6. residuals are normally distributed

ASSUMPTION #1. CONTINUOUS VARIABLES

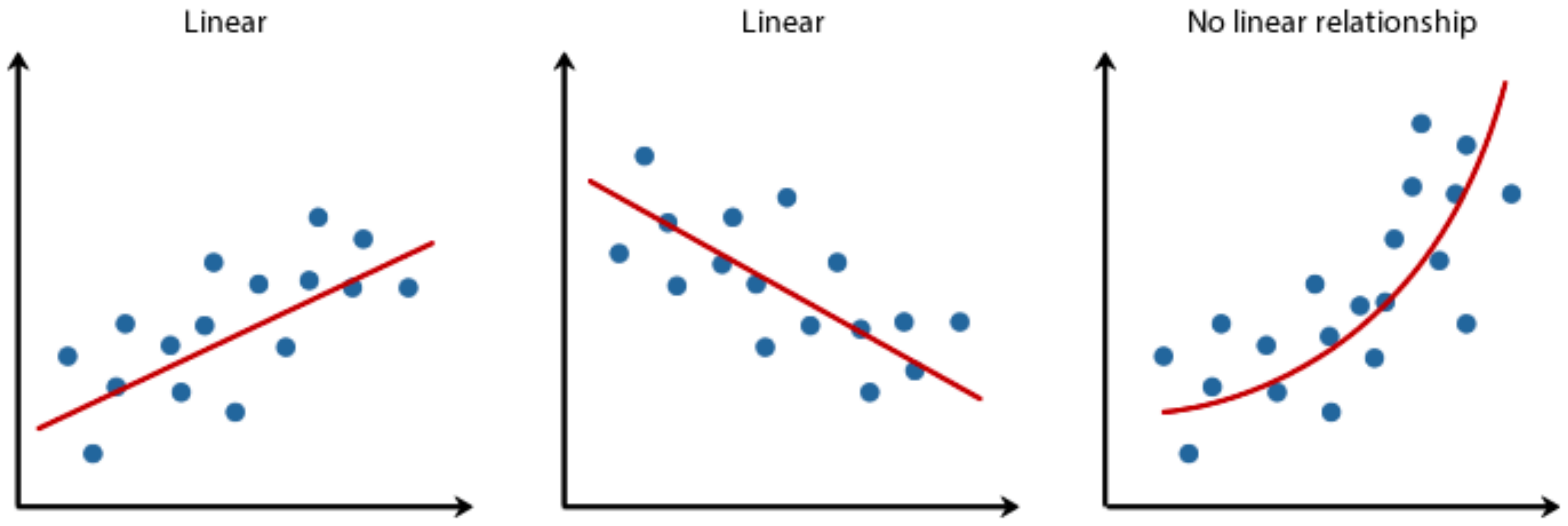
Outcome variable should be measured at the continuous level

1. CONTINUOUS:
   - Weight
   - Exam score (0-100)
   - Square Footage
2. CATEGORICAL:
   - Large, medium, small
   - Pass/Fail
   - Type of home



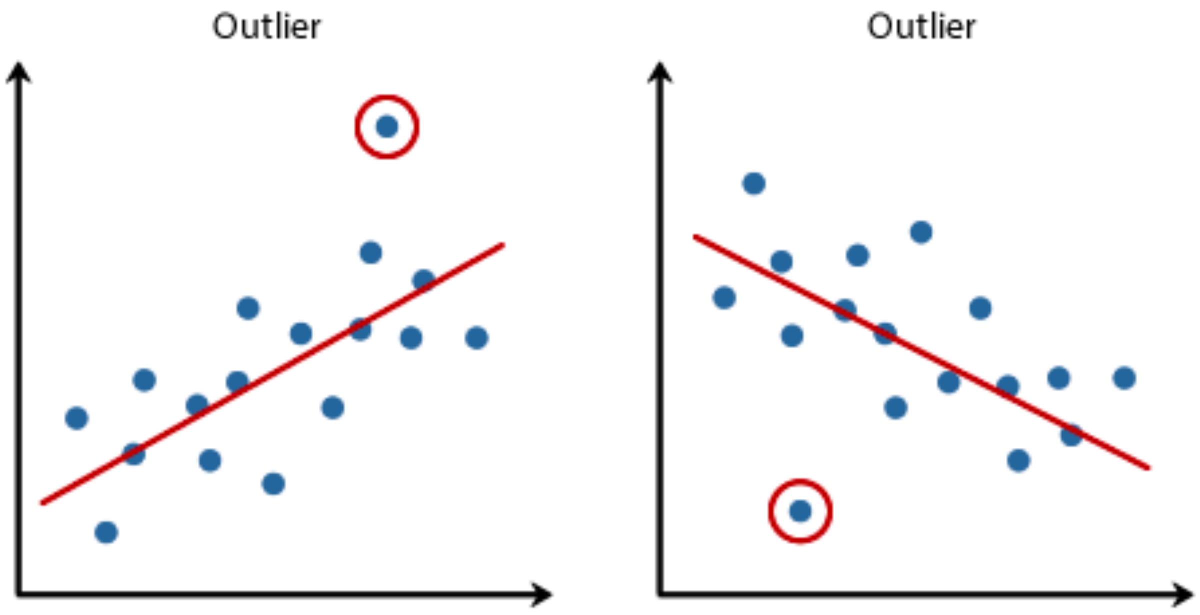
ASSUMPTION #2. LINEAR RELATIONSHIP

1. There needs to be a **linear relationship** between the target variable and its predictors



ASSUMPTION #3. no significant outliers

An outlier is an observed data point that has a dependent variable value that is very different to the value predicted by the regression equation. As such, an outlier will be a point on a scatterplot that is (vertically) far away from the regression line indicating that it has a large residual

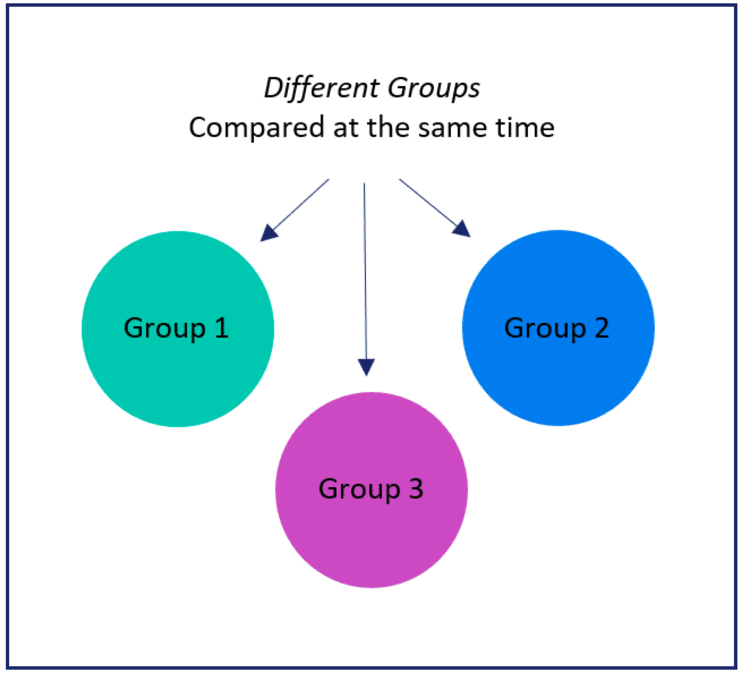


ASSUMPTION #4. independence of observations

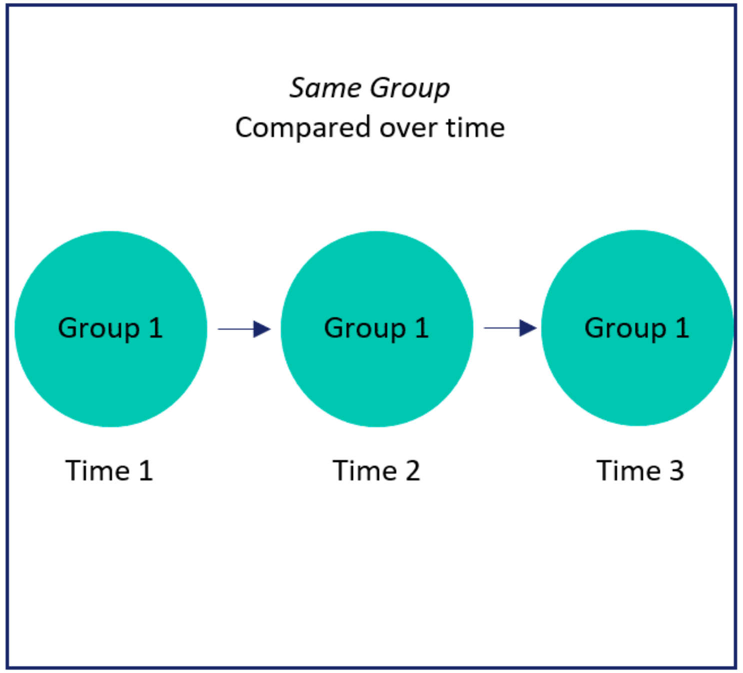
Every observation in the dataset must be independent of every other observation.

We generally have two types of data: cross sectional and longitudinal. Cross -sectional datasets are those where we collect data on entities only once. For example we collect IQ and GPA information from the students at any one given time (think: camera snap shot). Longitudinal data set is one where we collect GPA information from the same student over time (think: video). In cross sectional datasets we do not need to worry about Independence assumption. It is “assumed” to be met.

Cross-Sectional Study



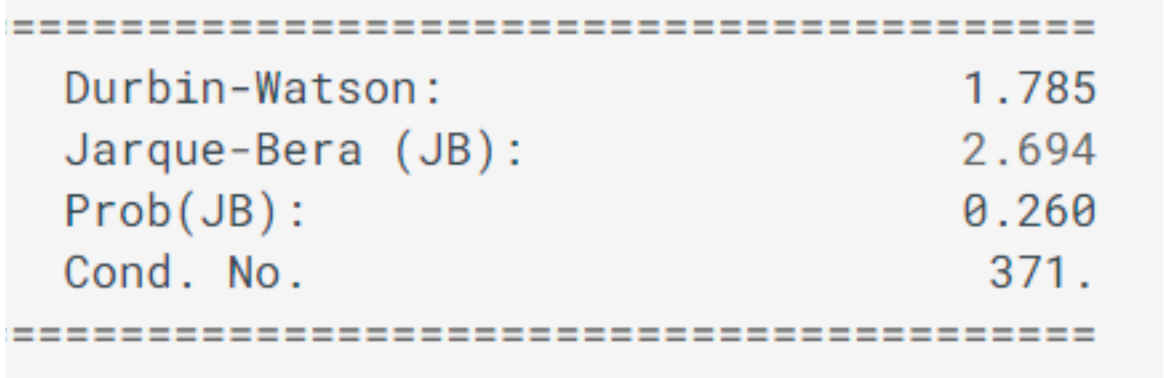
Longitudinal Study



DURBIN-WATSON TEST

One of the assumptions of regression is that the observations are independent.

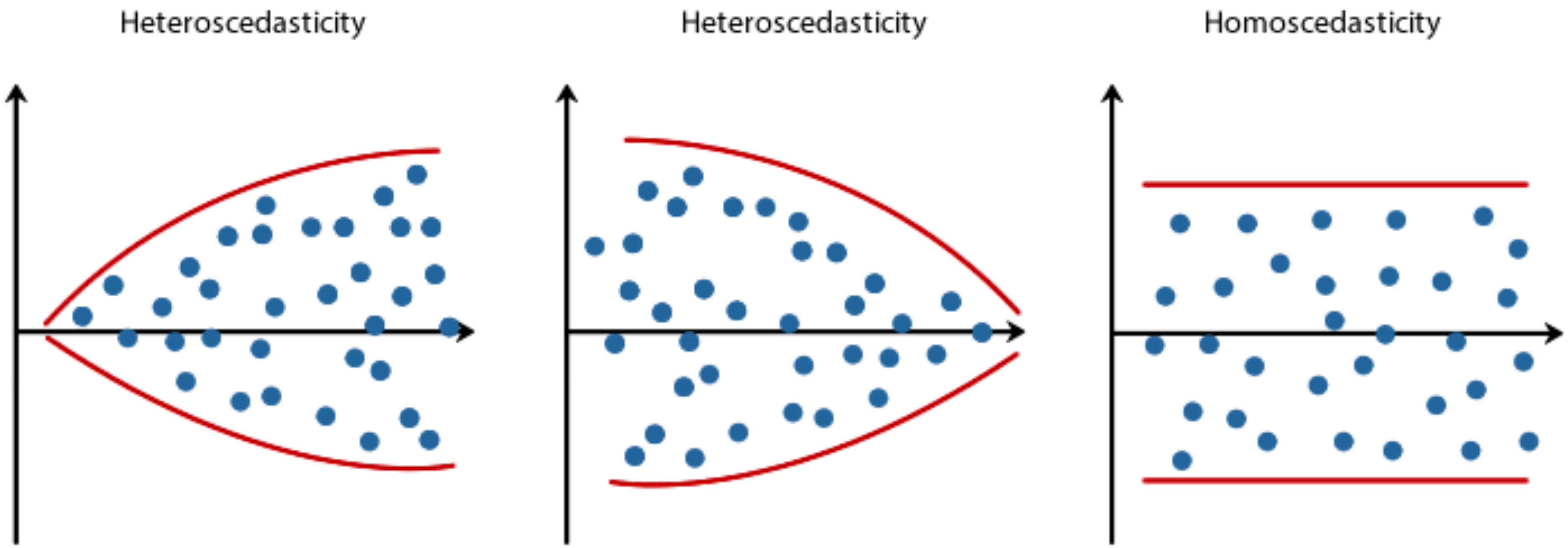
If observations are made over time, it is likely that successive observations are related.
If there is no autocorrelation (where subsequent observations are related), the Durbin-Watson statistic should be between 1.5 and 2.5.



```
==============================================
  Durbin-Watson:                        1.785
  Jarque-Bera (JB):                     2.694
  Prob(JB):                             0.260
  Cond. No.                              371.
==============================================
```

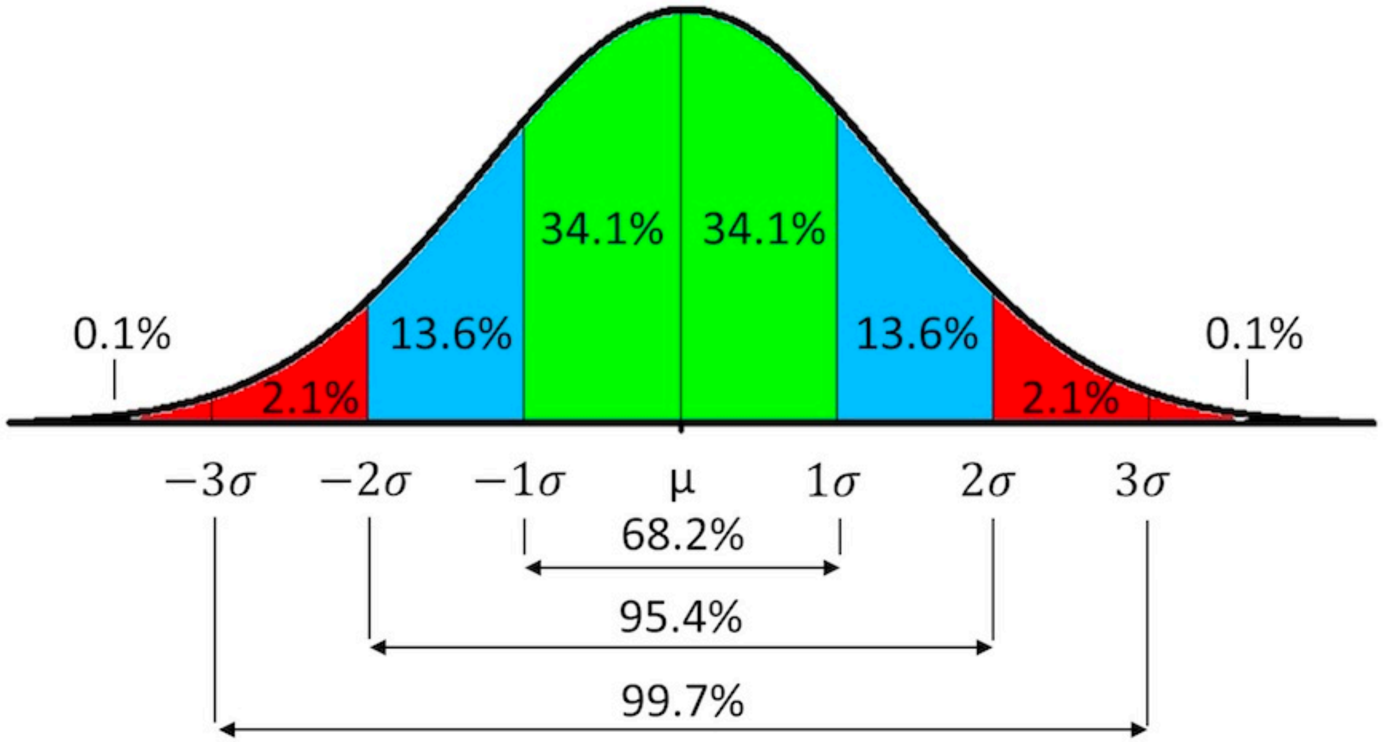
ASSUMPTION #5. homoscedasticity

Your data needs to show homoscedasticity, which is where the variances along the line of best fit remain similar as you move along the line.



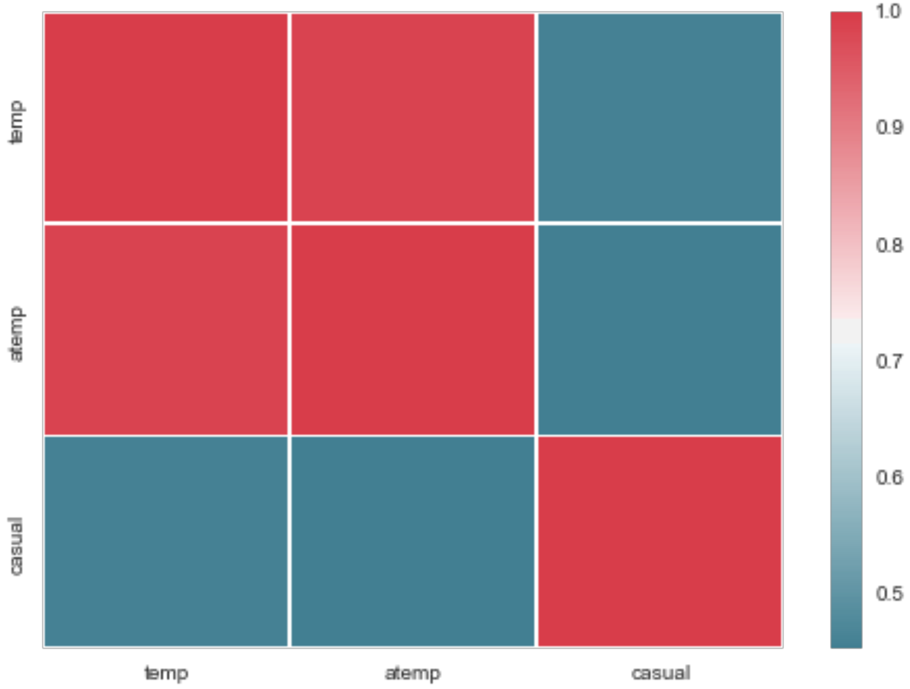
ASSUMPTION #6. residuals are normally distributed

The residuals (errors) of the regression line should be approximately normally distributed



MULTICOLLINEARITY: BIKE DATA EXAMPLE

- We can look at a correlation matrix of our bike data.
- Even if adding correlated variables to the model improves overall variance, it can introduce problems when explaining the output of your model.
- What happens if we use a second variable that isn't highly correlated with temperature?

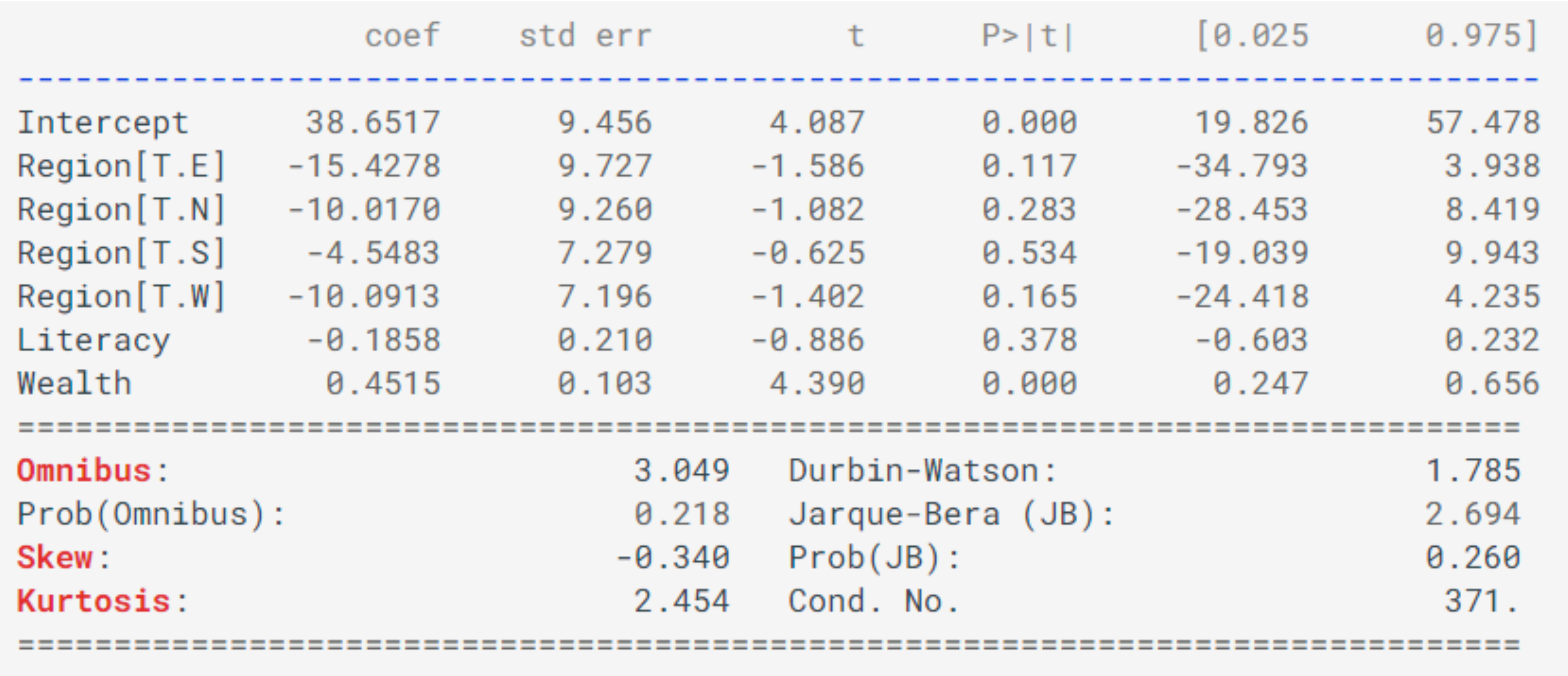


LINEAR REGRESSION

INTERPRET AND EVALUATE A MODEL

INTERPRET LINEAR REGRESSION RESULTS

statsmodels

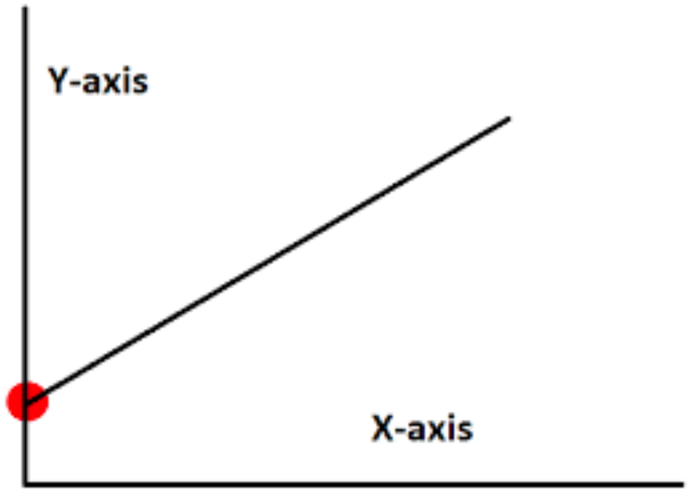


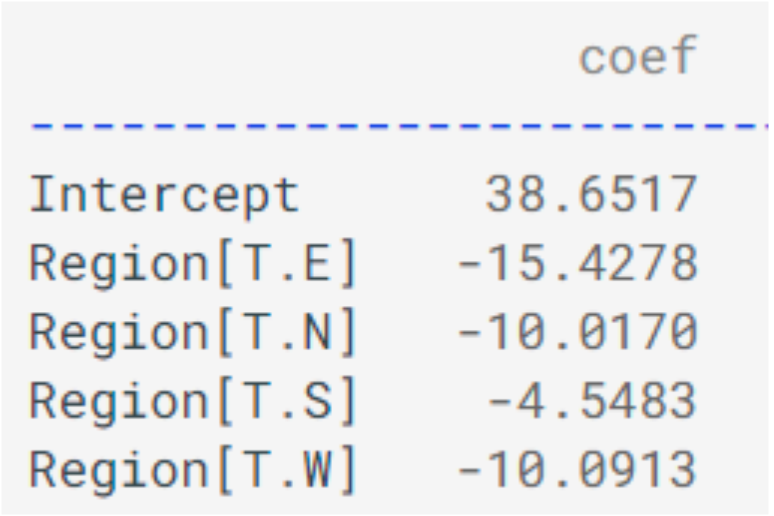
```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     38.6517      9.456      4.087      0.000      19.826      57.478
Region[T.E]  -15.4278      9.727     -1.586      0.117     -34.793       3.938
Region[T.N]  -10.0170      9.260     -1.082      0.283     -28.453       8.419
Region[T.S]   -4.5483      7.279     -0.625      0.534     -19.039       9.943
Region[T.W]  -10.0913      7.196     -1.402      0.165     -24.418       4.235
Literacy      -0.1858      0.210     -0.886      0.378      -0.603       0.232
Wealth         0.4515      0.103      4.390      0.000       0.247       0.656
==============================================================================
Omnibus:                        3.049   Durbin-Watson:                   1.785
Prob(Omnibus):                  0.218   Jarque-Bera (JB):                2.694
Skew:                          -0.340   Prob(JB):                        0.260
Kurtosis:                       2.454   Cond. No.                         371.
==============================================================================
```

https://medium.com/swlh/interpreting-linear-regression-through-statsmodels-summary-4796d359035a

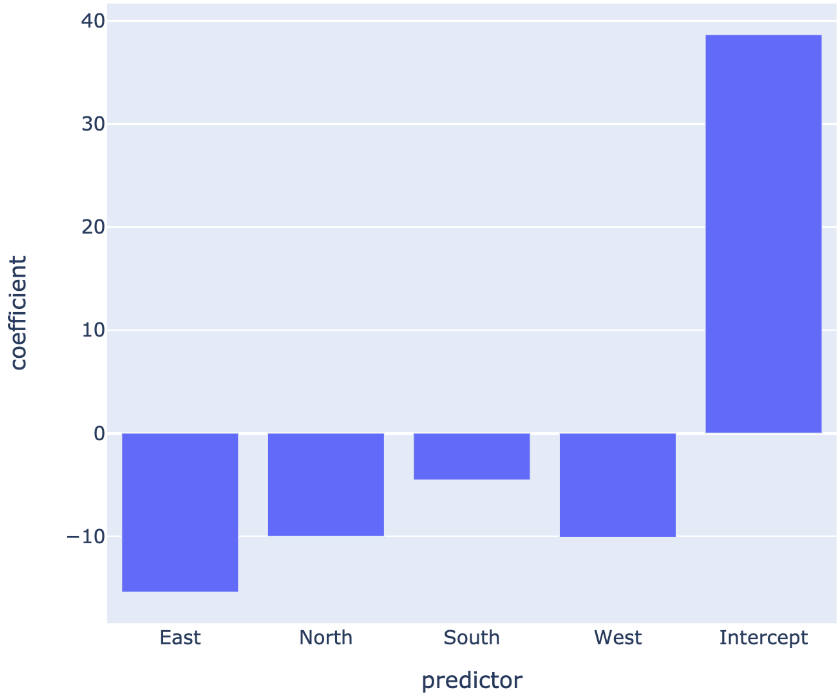
INTERPRET LINEAR REGRESSION RESULTS

statsmodels





```
                     coef
---------------------------
Intercept         38.6517
Region[T.E]      -15.4278
Region[T.N]      -10.0170
Region[T.S]       -4.5483
Region[T.W]      -10.0913
```



EVALUATE LINEAR REGRESSION RESULTS

- 1. Mean Absolute Error (MAE) is the mean of the absolute value of the errors.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - y_j|$$

- 2. Mean Squared Error (MSE) is the mean of the squared errors

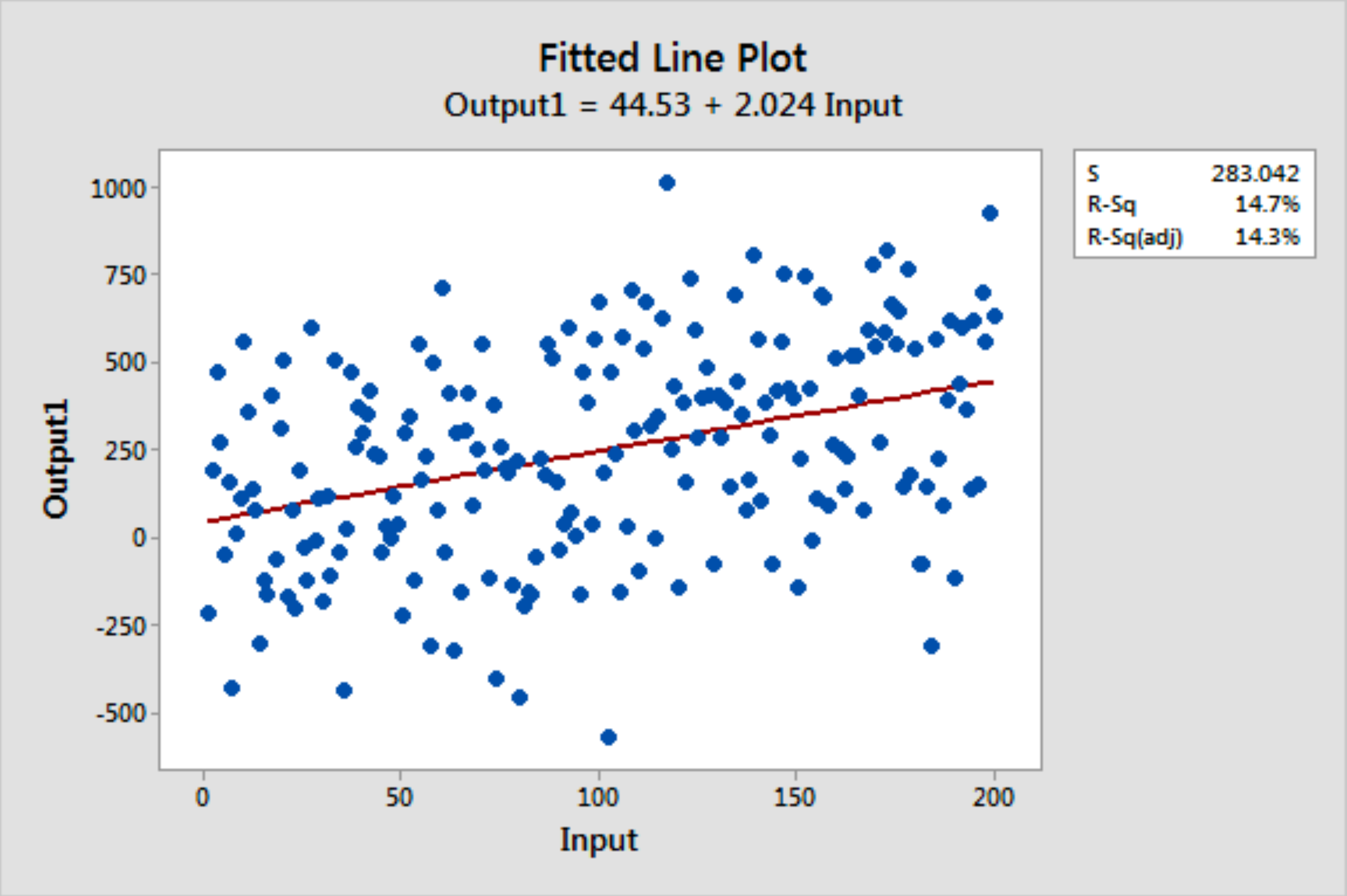
$$MSE = \frac{1}{N}\sum_{i}^{n}(Y_i - y_i)^2$$

- 3. Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

EVALUATE LINEAR REGRESSION RESULTS

- R-squared is a statistical measure of how close the data are to the fitted regression line.
- It ranges from 0-100%.
- 100% indicates that the model explains all the variability of the response data around its mean.
- Image of low R-squared



WHAT IS R-SQUARED?

- R-squared, the central metric introduced for linear regression
- Also known as “Coefficient of determination”
- R-squared measures explained variance.
- Which model performed better, one with an r-squared of 0.79 or 0.81?