Data Science

Unit 3-01: Regression Evaluation

AKADEMI GA

GENERAL ASSEMBLY

COURSE CONTENT

Week 1 : Data Science Foundations

Installation and Github, Python fundamentals, Introduction to Pandas

Congratulations!

Week 2 : Working with Data

More pandas, basics of probability and statistics, Exploratory Data Analysis (EDA), working with data, use statistical analysis and visualisation

Congratulations!

Week 3 : Data Science Modeling

Linear regression Train/Test/Split, Classification, Logistic Regression

Week 4 : Data Science Applications

Using APIs, Natural Language Processing, Time Series Analysis

Week 5: Final Presentation

Present your capstone project

Week 3: Data Science Modeling

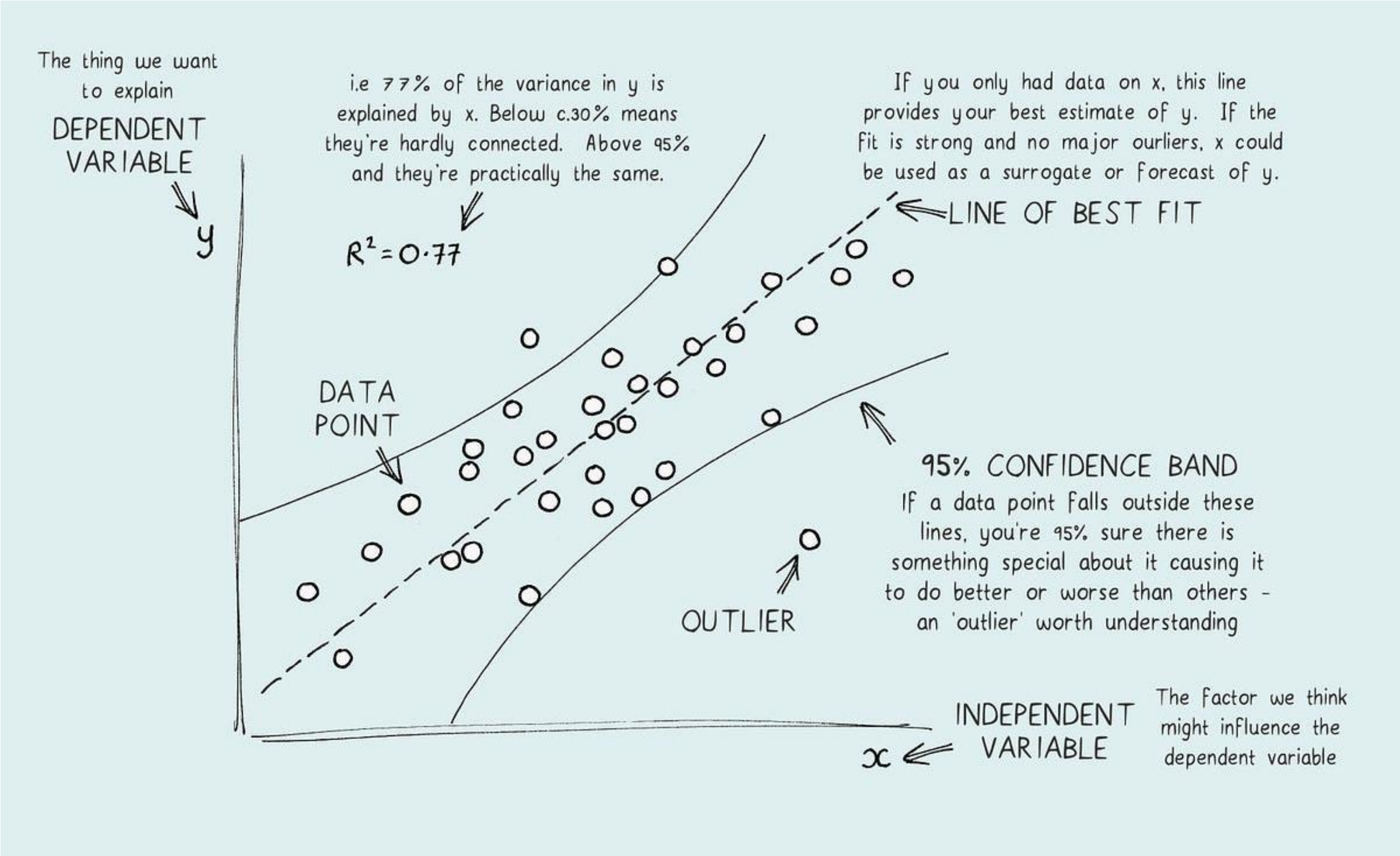
- *In Unit 3, we will use machine learning Python modules.*
- *We will review the theory of machine learning and hands-on practice for classification regression modelling.*

Week 3 Units
3-01 Linear Regression
3-02 Regression Evaluation
3-03 Intro to classification
3-04 Logistic Regression
3-05 Grid searching & Decision Tree

Schedule

Time	Topics
5:00 - 6:30	Lesson 1: Linear Regression Recap and Evaluation
6:30 - 6:45	Break
6:45 - 7:45	Lesson 2: Linear Regression Model Interpretation
7:45 - 8:00	Wrap up and Q&A

Linear Regression



Regression Evaluation

Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

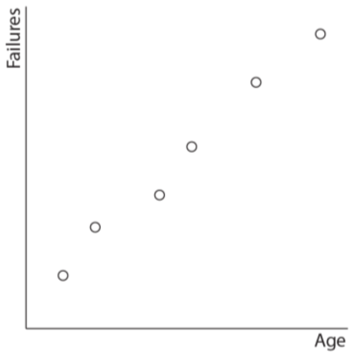
1. Mean Absolute Error (MAE)

$$MAE = \frac{\Sigma|y - \hat{y}|}{N}$$

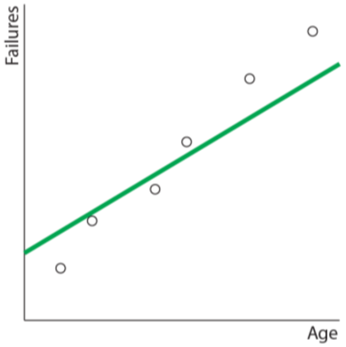
where y is the actual value $\hat{y}$ is the predicted value and $|y - \hat{y}|$ is the absolute value of the difference between the actual and predicted value. N is the number of sample points.

Let's dig into this a bit deeper to understand what this calculation represents.

Take a look at the following plot, which shows the number of failures for a piece of machinery against the age of the machine:



Age	Failures
10	15
20	30
40	40
50	55
70	75
90	90



Age	Failures	Prediction
10	15	26
20	30	32
40	40	44
50	55	50
70	75	62
90	90	74

1. Mean Absolute Error (MAE)

Age	Failures	Prediction	Error
10	15	26	11
20	30	32	2
40	40	44	4
50	55	50	-5
70	75	62	-13
90	90	74	-16

abs(Error)
11
2
4
5
13
16

Mean abs(Error)	8.5

The mean of the absolute errors (MAE) is 8.5.

1. Mean Absolute Error (MAE)

Age	y Failures	$\hat{y}$ Prediction	$y-\hat{y}$ Error
10	15	26	11
20	30	32	2
40	40	44	4
50	55	50	-5
70	75	62	-13
90	90	74	-16

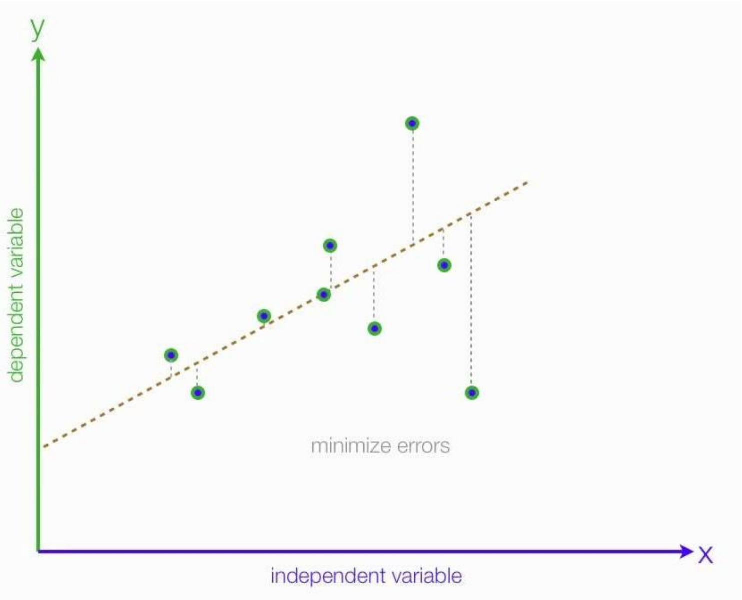
$\lvert y-\hat{y}\rvert$ abs(Error)
11
2
4
5
13
16

Mean abs(Error)	$\frac{\Sigma\lvert y-\hat{y}\rvert}{N}$	**8.5**

Mean Absolute Error (MAE) tells us the average error in units of y, the predicted feature. A value of 0 indicates a perfect fit, i.e. all our predictions are spot on.

2. Root Mean Square Error (RMSE)

- Compared to MAE, RMSE gives a higher total error and the gap increases as the errors become larger. It penalizes a few large errors more than a lot of small errors. If you want your model to avoid large errors, use RMSE over MAE.
- **Root Mean Square Error (RMSE)** indicates the average error in units of y, the predicted feature, but penalizes larger errors more severely than MAE. A value of 0 indicates a perfect fit.



3. R-Squared

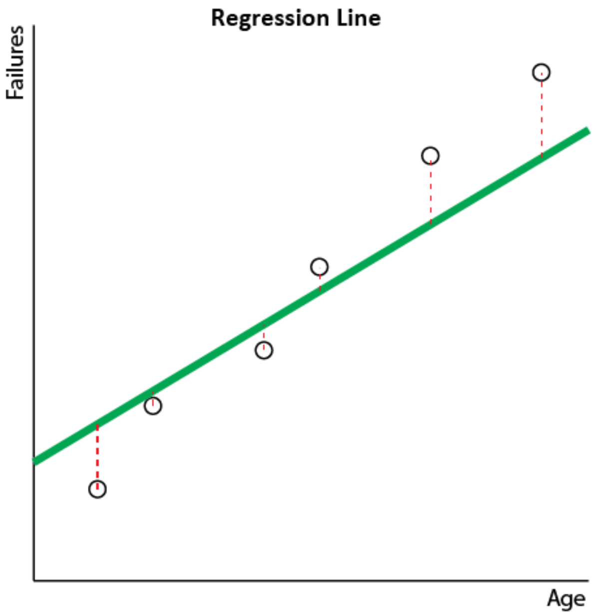
This is where R-squared or R^2 comes in. Here is the formula for R^2 :

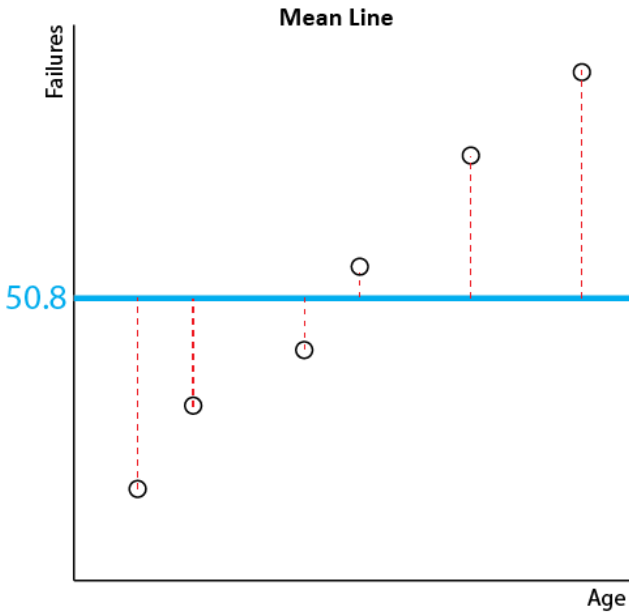
$$R^2 = \frac{\Sigma(y-\bar{y})^2 - \Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2}$$

R^2 computes **how much better the regression line fits the data than the mean line**. Another way to look at this formula is to compare the *variance* around the mean line to the variation around the regression line:

$$R^2 = \frac{var(mean) - var(line)}{var(mean)}$$

3. R-Squared





3. R-Squared

			Regression Line	Mean Line	Regression Line	Mean Line
	y	$\hat{y}$	$y-\hat{y}$	$y-\bar{y}$	$(y-\hat{y})^2$	$(y-\bar{y})^2$
Age	Failures	Prediction	Error	Error	Error^2	Error^2
10	15	26	11	-35.8	121	1281.6
20	30	32	2	-20.8	4	432.6
40	40	44	4	-10.8	16	116.6
50	55	50	-5	4.2	25	17.6
70	75	62	-13	24.2	169	585.6
90	90	74	-16	39.2	256	1536.6

Mean of Error^2	$\frac{\Sigma(y-\hat{y})^2}{N}$	98.5	$\frac{\Sigma(y-\bar{y})^2}{N}$	661.8

R^2	$\frac{\Sigma(y-\bar{y})^2-\Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2}$	**0.85**

So we have an R-squared of 0.85. Without even worrying about the units of y, we can say this is a decent model. Why? Because the model explains 85% of the variation in the data. That's exactly what an R-squared of 0.85 tells us!

Summary

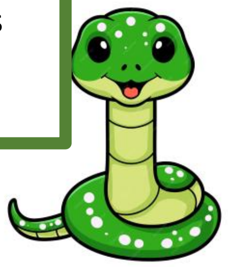
- **Mean Absolute Error (MAE)** tells us the average error in units of y, the predicted feature. A value of 0 indicates a perfect fit.

- **Root Mean Square Error (RMSE)** indicates the average error in units of y, the predicted feature, but penalizes larger errors more severely than MAE. A value of 0 indicates a perfect fit.

- **R-squared (R2)** tells us the degree to which the model explains the variance in the data. In other words how much better it is than just predicting the mean.

  - A value of 1 indicates a perfect fit.
  - A value of 0 indicates a model no better than the mean.
  - A value less than 0 indicates a model worse than just predicting the mean.

Python does the hard work and calculates these metrics for us from our model outputs.



Schedule

Time	Topics
5:00 - 6:30	Lesson 1: Linear Regression Recap and Evaluation
6:30 - 6:45	Break
6:45 - 7:45	Lesson 2: Linear Regression Model Interpretation
7:45 - 8:00	Wrap up and Q&A

Q&A