Linear Regression

Nazgul Rakhimzhanova

COURSE SCHEDULE

week	Mid Term (weeks 01-07)	End Term (weeks 08-14)	week
01	Intro: Data Science Area and open source tools for Data Science	Feedback review	08
02	NumPy package for data science	Sampling and Estimation	09
03	Pandas package for data science	Visualization II. Correlation and Covariance	10
04	Visualization with matplotlib	Hypothesis testing	11
05	Statistics: Distribution – Normal	Decision tree	12
06	Exploratory Data Analysis (EDA)	Linear Regression	13
07	**Summary for 6 weeks QA session**	**Summary for 6 weeks QA session**	**14**
15	**Course summary**		

PREVIOUSLY

- Confidence Intervals
- Hypothesis Testing (z-test, chi-s.test)

IDEA

- Often, we want to be able to predict an outcome relying on available information. Like our ancestors practiced to predict harvest quality based on 'surrounded natural' features.

Linear Regression

IDEA

- In supervised machine learning we have two different algorithms than can help us to predict - **Regression** and **classification**.
- Regression predicts continuous value outputs
- Classification predicts discrete outputs

REGRESSION

- When there is only one dependent and one explanatory variable, that's **simple regression**
- When we have more than one explanatory variable that’s **multiple regression**
- If there is more than one dependent variable, that's **multivariate regression**

LINEAR REGRESSION

- Linear Regression
  - **Least square method**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Where $\boldsymbol{\beta_0}$ is the intercept, $\boldsymbol{\beta_1}$ is the parameter associated with x1, $\boldsymbol{\beta_2}$ is the parameter associated with x2, and $\boldsymbol{\varepsilon}$ is the residual due to random variation or other unknown factors.

LINEAR REGRESSION

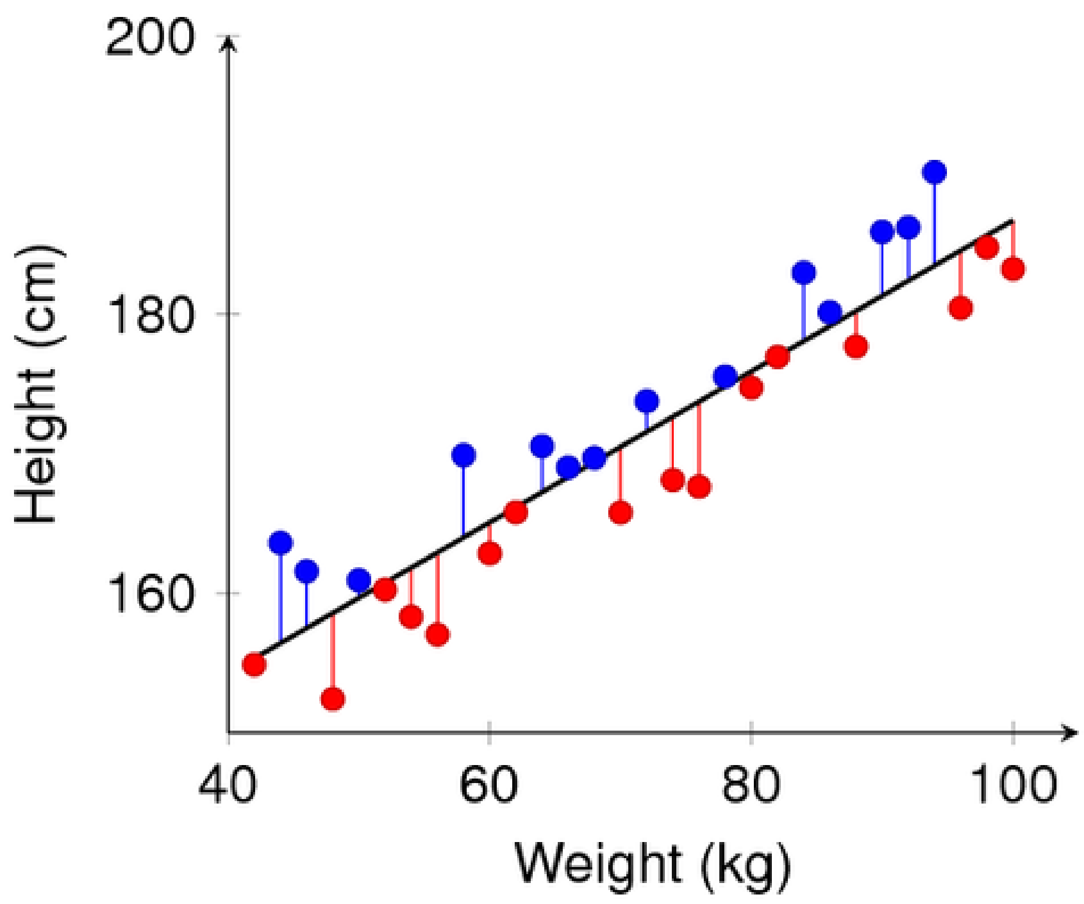
Linear Regression: Single Variable

$$\underbrace{\boxed{\hat{y}}}_{\text{Predicted output}} = \underbrace{\beta_0 + \beta_1}_{\text{Coefficients}} \underbrace{\boxed{x}}_{\text{Input}} + \underbrace{\boxed{\epsilon}}_{\text{Error}}$$

Linear Regression: Multiple Variables

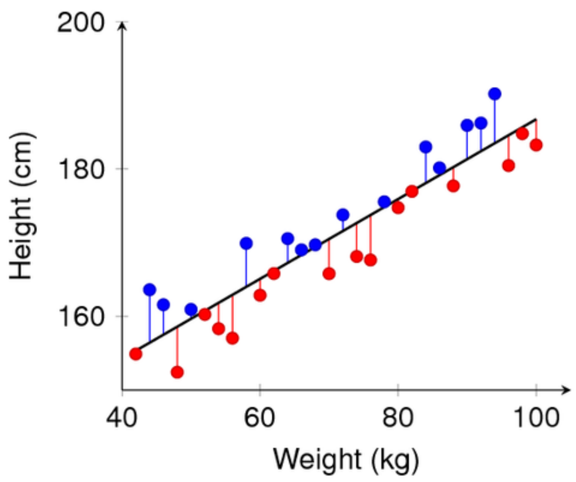
$$\boxed{\hat{y}} = \underbrace{\beta_0 + \beta_1}\boxed{x_1} + \dots + \underbrace{\beta_p}\boxed{x_p} + \boxed{\epsilon}$$

RESIDUALS



LINEAR REGRESSION

Given a sequence of values for y and sequences for x_1 and x_2, we can find the parameters, β_0, β_1 and β_2, that minimize the sum of ε^2. This process is called **ordinary least squares**.



LINEAR REGRESSION metrics

- The various metrics used to evaluate the results of the prediction are:
  - Mean Squared Error(**MSE**)
  - Root-Mean-Squared-Error(**RMSE**).
  - Mean-Absolute-Error(**MAE**).
  - $\mathbf{R^2}$ or Coefficient of Determination.

LINEAR REGRESSION metrics

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \widehat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

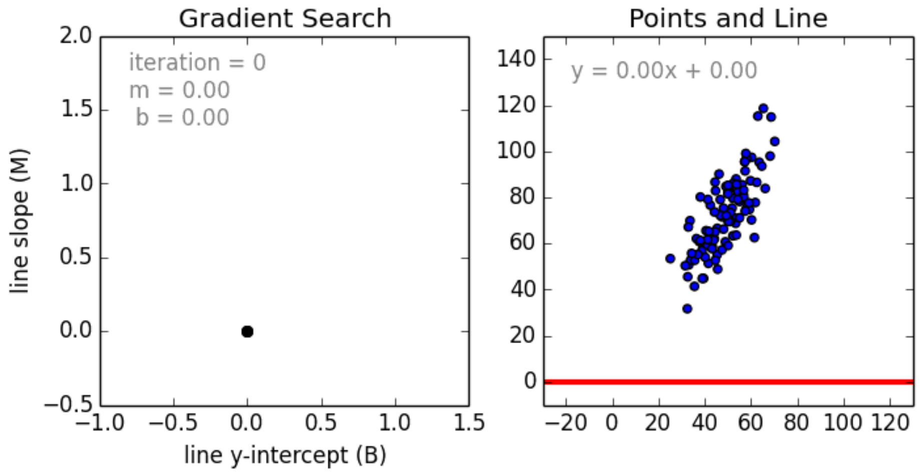
$$MAE = \overbrace{\boxed{\frac{1}{n}}}^{\text{Divide by the total number of data points}}\underbrace{\Sigma}_{\text{Sum of}}\underbrace{\left|\overbrace{\boxed{y}}^{\text{Actual output value}} - \overbrace{\boxed{\widehat{y}}}^{\text{Predicted output value}}\right|}_{\text{The absolute value of the residual}}$$

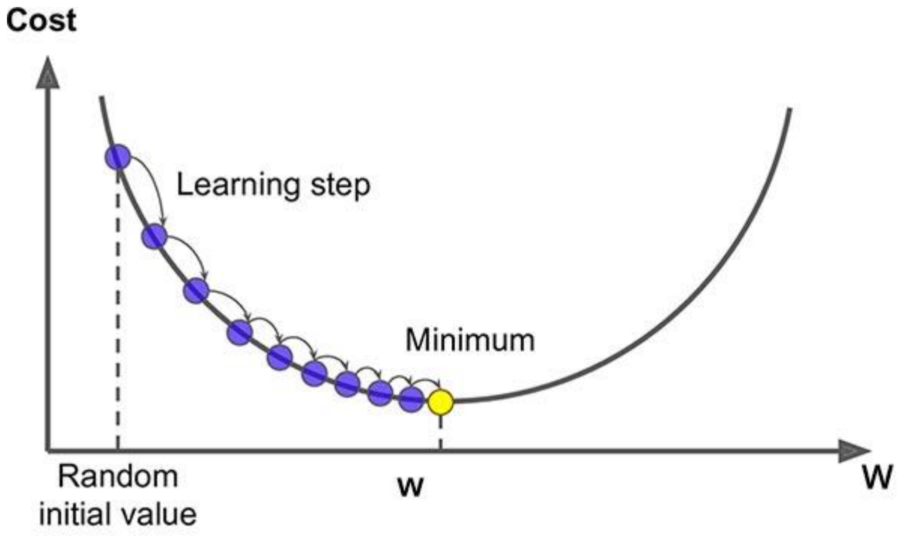
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

LINEAR REGRESSION

- Linear Regression
  - Least square method
    - **Gradient** descent
    - **Stochastic** gradient descent





Steps of Linear Regression model

- 1 step: **import** necessary packages
- 2 step: **read** your data – **handle nans’** and encode categorical values.
- 3 step: **prepare** your data – split your data into training and testing parts
- 4 step: **fit** your model (teach your model) on train data
- 5 step: **predict** on test data
- 6 step: **evaluate** your model with metrics
- 7 step: **improve** your model
- 8 step: **repeat** steps 3-7

REGRESSION

Before you attempt to perform linear regression, Your data must pass through certain required assumptions.

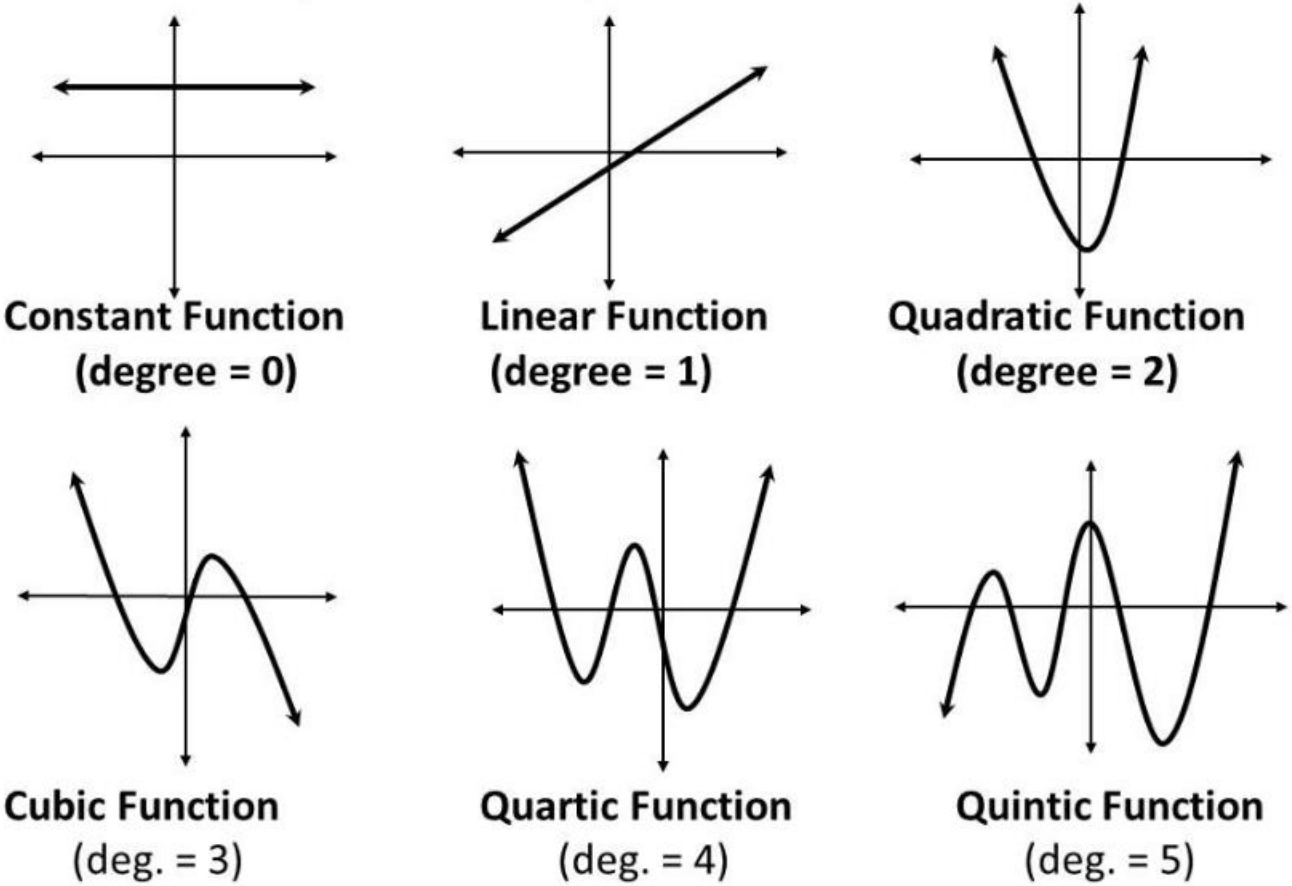
- The variables should be measured at a continuous level. Examples of continuous variables are time, sales, weight and test scores.
- Use a scatterplot to find out quickly if there is a linear relationship between those two variables.
- The observations should be independent of each other (that is, there should be no dependency).
- Your data should have no significant outliers.
- Check for homoscedasticity — a statistical concept in which the variances along the best-fit linear-regression line remain similar all through that line.
- The residuals (errors) of the best-fit regression line follow normal distribution.

HOW TO IMPROVE LR

- Drop outliers
- Drop unnecessary data
- Try polynomial features
- Try to train on different sets of independent variables (cross validation)
- Check correlation coefficient
- Check for confounding variable

Polynominal functions

Graphs of Polynomial Functions:



Output of OLS()

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 4.020e+06
Date:                Fri, 13 Mar 2020   Prob (F-statistic):          2.83e-239
Time:                        13:54:01   Log-Likelihood:                -146.51
No. Observations:                 100   AIC:                             299.0
Df Residuals:                      97   BIC:                             306.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3423      0.313      4.292      0.000       0.722       1.963
x1            -0.0402      0.145     -0.278      0.781      -0.327       0.247
x2            10.0103      0.014    715.745      0.000       9.982      10.038
==============================================================================
Omnibus:                        2.042   Durbin-Watson:                   2.274
Prob(Omnibus):                  0.360   Jarque-Bera (JB):                1.875
Skew:                           0.234   Prob(JB):                        0.392
Kurtosis:                       2.519   Cond. No.                         144.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Output of LinearRegression()

Out:

```
Coefficients:
 [938.23786125]
Mean squared error: 2548.07
Coefficient of determination: 0.47
```

READINGS

- https://machinelearningmastery.com/implement-linear-regression-stochastic-gradient-descent-scratch-python/
- shorturl.at/djLNS
- shorturl.at/ehkqz
- https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics