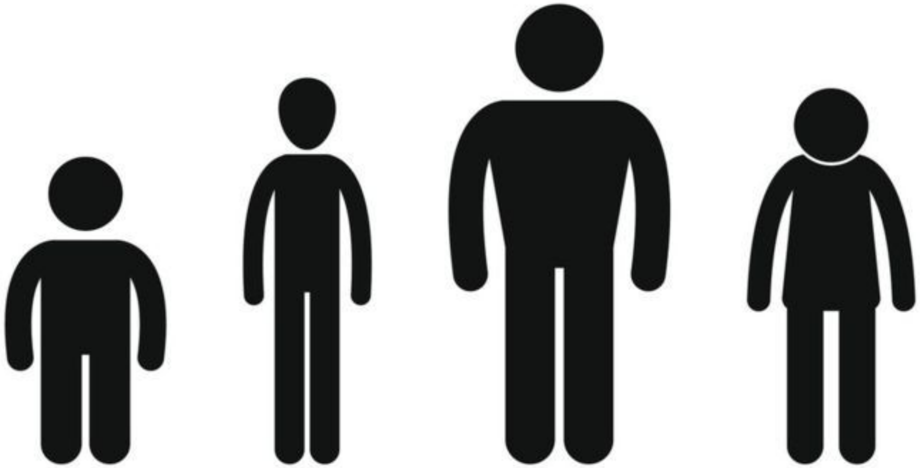
Linear Regression

Supervised Learning

A supervised model is trained on a labeled dataset of (feature, label) pairs.

Regression Model - numerical label

Problem: Predict weight (number) given height and age

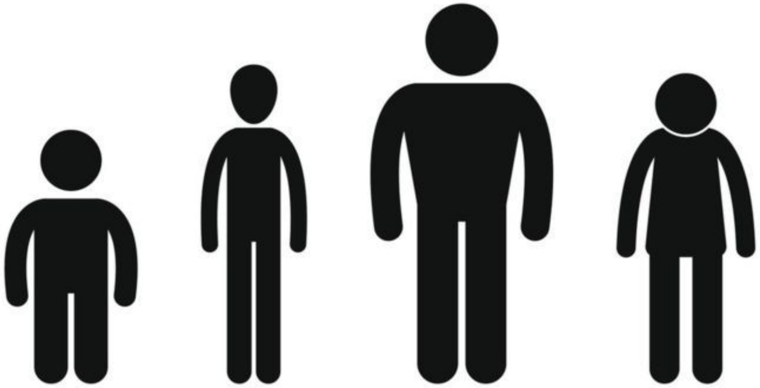


Features:

Height, Age

Label:

Weight



Height:	1.50	1.70	2.10	1.55	1.62
Age:	10	24	40	20	30
Weight:	40	58	80	45	**?**

Training data

Test data

The history of Linear Regression

1800s, Francis Galton

- Study of relationship between parents and children
    - Height of a father VS height of a son

- Son’s height close to fathers height

- But, son’s height is closer to the overall average height of all people

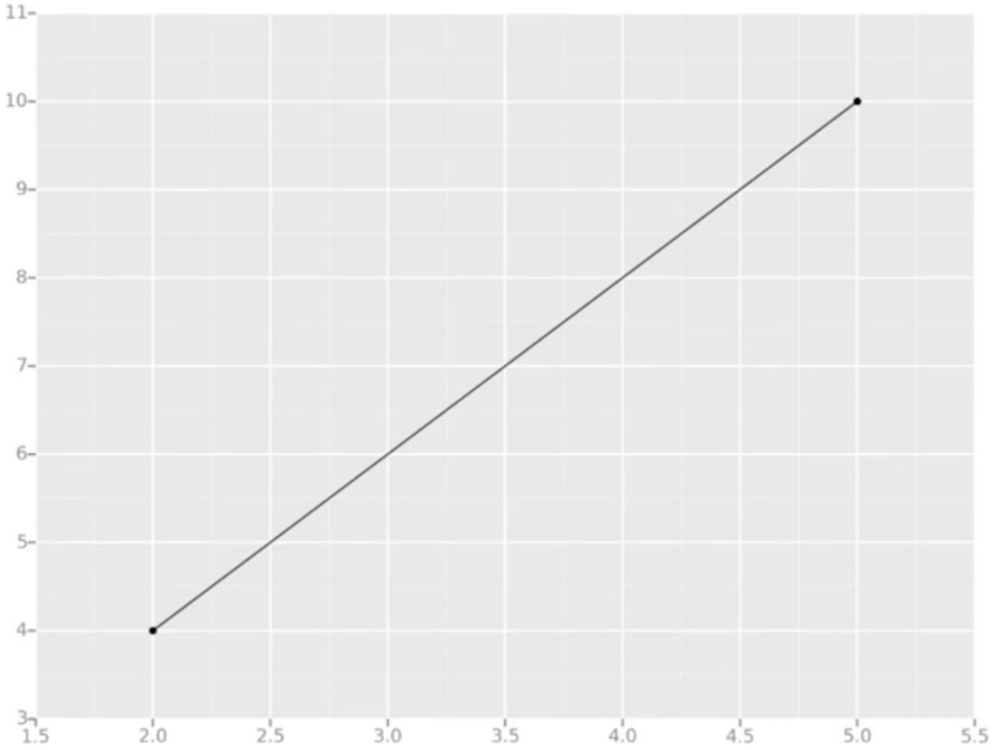
Example: Shaquille O’Neal - 2.2 m VS his son - 2m

- “Father’s son’s height tends to regress (drift towards) the average height”

Linear regression

- Draw a straight line that is as close to all the data points as possible

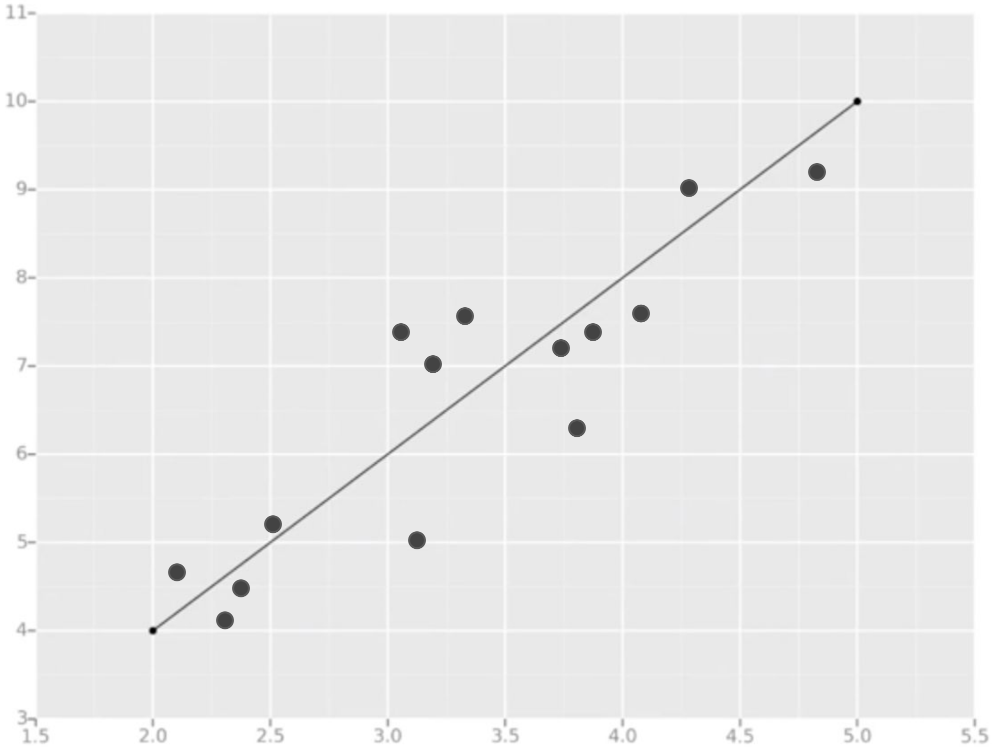
- Our line fits the data points perfectly



Linear regression

- Draw a straight line that is as close to all the data points as possible

- More points - some errors



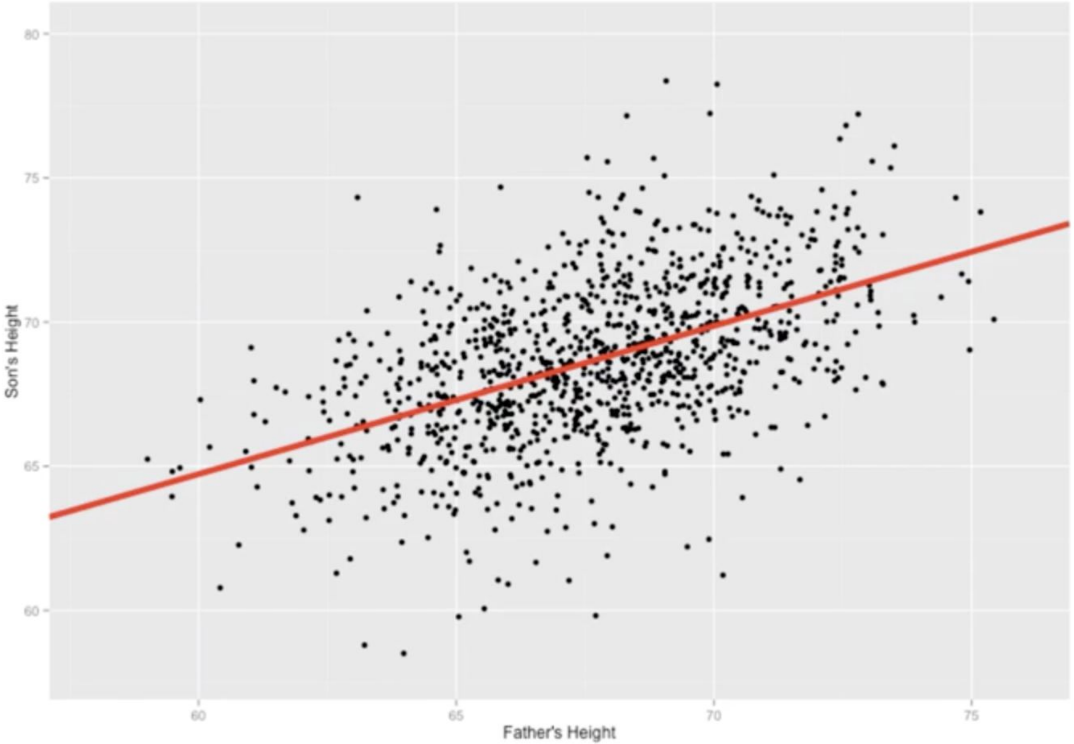
Benefits of Linear Regression

- Runs fast
- Easy to use
- Easily interpretable
- Basis for many other methods

Linear regression

- Our goal with this algorithm is to minimize the vertical distance between all data points and our line i.e. **to find the best line that describes our data**

- Different minimizing methods available: **least squares**, absolute distance, etc.

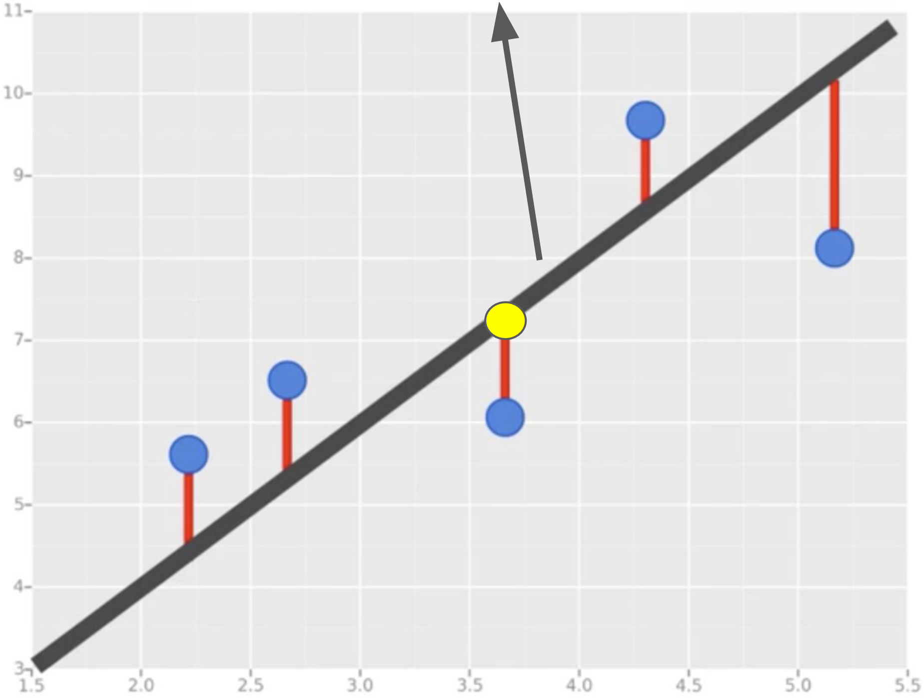


Least squares method

- Minimizing the sum of squares of the residuals
- **Residual** - difference between the observation (actual y-value) and the fitted line i.e. $y_i - \hat{y}$
- **Slope** - (change in y) / (change in x)
- **Intercept** - value of y when x=0

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\hat{y} = \overset{\text{intercept}}{\hat{\beta}_0} + \overset{\text{slope}}{\hat{\beta}_1} x$$



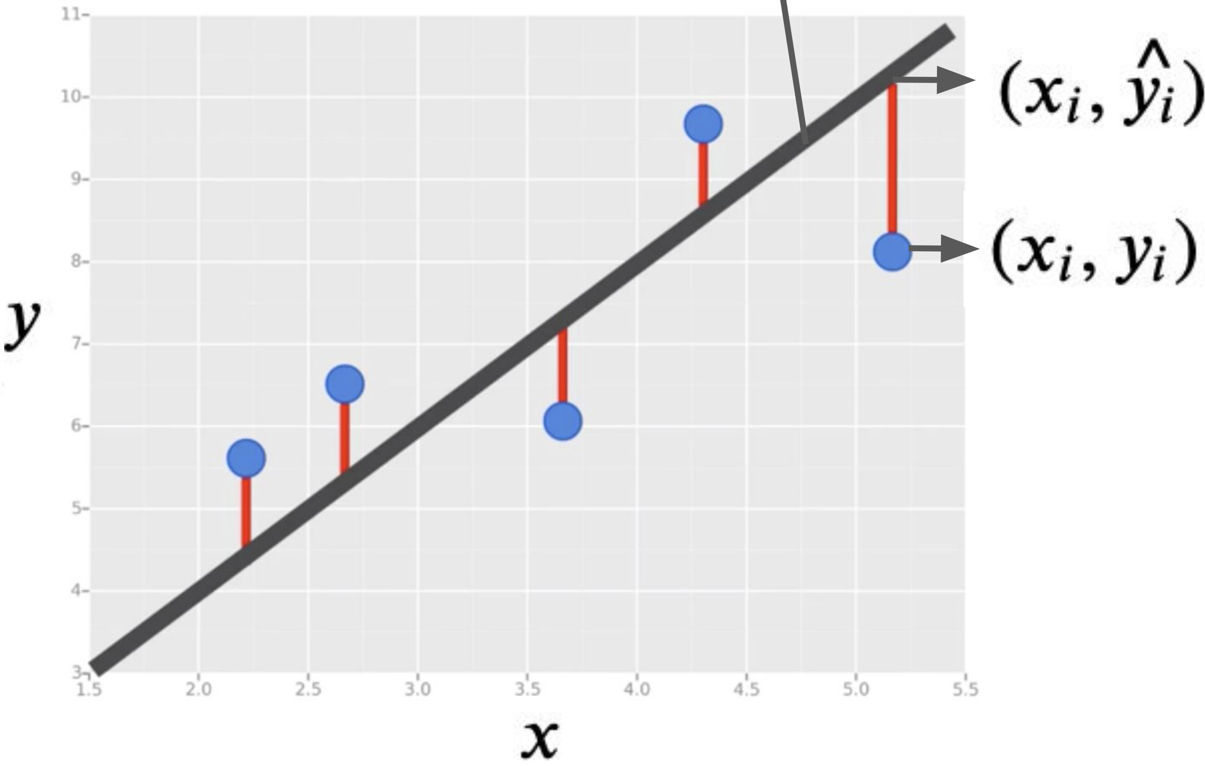
Bivariate Linear Regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$Y = \widehat{Y} + err$$

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

unknowns



$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

1) Take partial derivatives w.r.t. $\hat{\beta}_0$ and $\hat{\beta}_1$

2) Set the partial derivatives equal to 0

3) Solve the resulting equations for $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial}{\partial\hat{\beta}_0} \sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{N} \frac{\partial}{\partial\hat{\beta}_0}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{N} 2 * (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) * (-1) =$$

$$= -2 * \sum_{i=1}^{N}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$$

$$\frac{\partial}{\partial\hat{\beta}_1} \sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{N} \frac{\partial}{\partial\hat{\beta}_1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{N} 2 * (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) * (-x_i) =$$

$$= -2 * \sum_{i=1}^{N} x_i * (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$$

$$\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = -2 * \sum_{i=1}^{N} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0$$

$$\sum_{i=1}^{N} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0$$

$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} \hat{\beta}_0 - \sum_{i=1}^{N} \hat{\beta}_1 x_i = 0$$

$$\sum_{i=1}^{N} y_i - N\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{N} x_i = 0$$

$$N\hat{\beta}_0 = \sum_{i=1}^{N} y_i - \hat{\beta}_1 \sum_{i=1}^{N} x_i$$

$$\boxed{\hat{\beta}_0 = \frac{\sum_{i=1}^{N} y_i - \hat{\beta}_1 \sum_{i=1}^{N} x_i}{N}}$$

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \ = \ -2 * \sum_{i=1}^{N} x_i * (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) \ = \ 0$$

$$\sum_{i=1}^{N} x_i * (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) \ = \ 0$$

$$\sum_{i=1}^{N} (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^{\,2}) \ = \ 0$$

$$\sum_{i=1}^{N} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{N} x_i - \hat{\beta}_1 \sum_{i=1}^{N} x_i^{\,2}) \ = \ 0 \quad \longleftarrow \quad \hat{\beta}_0 \ = \ \frac{\sum_{i=1}^{N} y_i - \hat{\beta}_1 \sum_{i=1}^{N} x_i}{N}$$

$$\sum_{i=1}^{N} x_i y_i - \frac{(\sum_{i=1}^{N} y_i - \hat{\beta}_1 \sum_{i=1}^{N} x_i) * \sum_{i=1}^{N} x_i}{N} - \hat{\beta}_1 \sum_{i=1}^{N} {x_i}^2 = 0$$

$$\sum_{i=1}^{N} x_i y_i - \frac{1}{N} \sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i + \frac{\hat{\beta}_1}{N} (\sum_{i=1}^{N} x_i)^2 - \hat{\beta}_1 \sum_{i=1}^{N} {x_i}^2 = 0$$

$$\sum_{i=1}^{N} x_i y_i - \frac{1}{N} \sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i = \hat{\beta}_1 \sum_{i=1}^{N} {x_i}^2 - \frac{\hat{\beta}_1}{N} (\sum_{i=1}^{N} x_i)^2$$

$$\sum_{i=1}^{N} x_i y_i - \frac{1}{N} \sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i = \hat{\beta}_1 (\sum_{i=1}^{N} {x_i}^2 - \frac{1}{N} (\sum_{i=1}^{N} x_i)^2)$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i - \frac{1}{N} \sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} {x_i}^2 - \frac{1}{N} (\sum_{i=1}^{N} x_i)^2}}$$

Multivariate Linear Regression

Suppose we have **n** data points of **k** dimensions - each data point is described with **k** features. A general multivariate model can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i \qquad \text{for } i = 1, \ldots, \text{n.}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \\ u_n \end{bmatrix}$$

n x 1   n x (k+1)   (k+1) x 1   n x 1

$$Y = X\beta + u \qquad \boldsymbol{u = Y - X\beta}$$

Error term: how far is the actual y from regression line

$$\min_{\beta} \quad u'u = (Y - X\beta)'(Y - X\beta)$$

$$u'u = (Y' - \beta'X')(Y - X\beta)$$

$$= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta$$

$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

$$Y'X\beta$$

1xn nx(k+1) (k+1)x1

1xk+1 k+1x1

1x1

Transpose of a scalar is equal to itself

$$Y'X\beta = (Y'X\beta)' = \beta'X'Y$$

$$u'u = Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

$$\frac{\partial(u'u)}{\partial\beta} = -2X'Y + 2X'X\beta$$

$$-2X'Y + 2X'X\beta = 0$$

$$X'X\beta = X'Y$$

$$(X'X)^{-1}(X'X)\beta = (X'X)^{-1}X'Y$$

$$\boxed{\beta = (X'X)^{-1}X'Y}$$

Training, test and validation sets

- **Training set -** a subset to train a model
- **Validation set -** a set used for parameter tuning
- **Test set -** a subset to test the trained model

Test set criteria:

- Large enough to give statistically meaningful results
- Representative of the data set as a whole (has same characteristics as the training set)

Never train on the test set

Model Evaluation

- **R-squared** - proportion of variance explained (0,1)
- What is a good R-squared value? - hard to say
- More features - higher R-square => not a reliable approach for choosing the best model

$$Unexplained\ variance = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$$Total\ variance = \sum_{i=1}^{n} (y_i - avg(y))^2$$

$$R^2 = 1 - \frac{Unexplained\ variance}{Total\ variance}$$

Model Evaluation

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

Mean Squared Error (MSE) is the mean of the squared errors:

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

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

✔ **RMSE** is the most popular since it is interpretable in “y” units.

Thank you!