Regression Analysis

Lecture outline

1. The regression problem and types of regression
2. Least squares
3. Linear least squares regression
4. Multivariate linear regression
5. Prediction and inference with linear regression

The regression problem and types of regression

Regression analysis

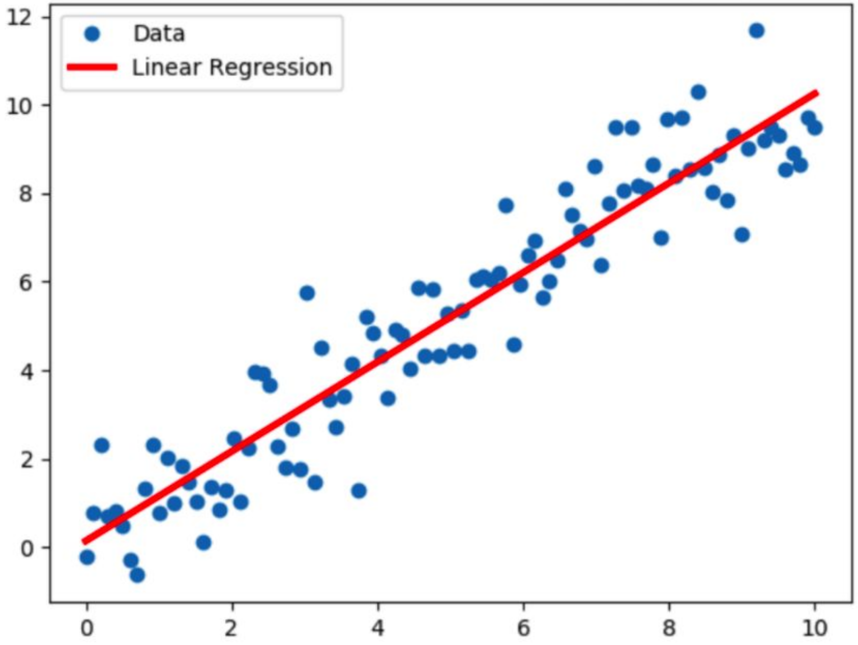
- Regression analysis is about analysing quantitative and predictive relationships between variables
- This is used to predict the value of an unknown variable using the value of known variables
- The task before us is to build predictive mathematical models
- Find a function f(x) that maps the independent variable x to the dependent variable y (target)
    - f(x) is estimated using some training data from the same process we are trying to model
- X represents features/attributes/measurements
    - Called the predictor / covariate / independent variable
    - E.g. House size
- Y represents observations
    - Called the outcome / response / target / dependent variable
    - E.g. House price

Objectives of regression analysis

1. Establish if there is a statistically significant relationship between the dependent variable and independent variables
    a. If yes, find that relationship
2. Use the above relationship that has been estimated to forecast unobserved values of the dependent variable

Simple Linear Regression

Simple Linear Regression



- Find the best <u>linear function</u> f(x) that maps the independent variable x to the dependent variable y (target)
- X : E.g. House size. Usually a real number
- Y : E.g. House price. Usually a real number

$$y = f(x) = b_1 x + b_0$$

The regression coefficients b_0 and b_1 should be learned from the data

Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- y is the dependent variable
- x is the independent variable
- β_0 is the constant or intercept

How to figure out the regression coefficients: Least squares method


5 Minutes with Cyrill
Least Squares

Simple linear regression: Modeling data with a function + noise

- For bivariate data $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$

$$y_i = f(x_i) + \epsilon_i$$

- ϵ_i is a random error term which is ***assumed*** to be:
  - distributed normally $\epsilon_i \sim N(0,\sigma^2)$
  - independent of each other
  - independent of x_i.
- When f(x) is linear:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$Total\ sq.\ err: \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

- Goal: find the value of β_0 and β_1 to get the best fitting line. Also estimate σ
  - β_0 and β_1 are the **y-intercept** and the **gradient** of the straight line we try to find

Estimating the parameters (β_0, β_1) of our model

- For bivariate data $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$Total\ sq.\ err: \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

$$(\beta_0, \beta_1) = \underset{(b_0, b_1)}{\operatorname{argmin}}\, \mathbb{E}\left[(Y - (b_0 + b_1 X))^2\right]$$

Method of least squares

- The above minimization is defined for the whole distribution of data, which we cant access. We only have samples. Therefore, the in-sample/empirical/training Mean Squared Error (MSE) is defined as follows:

$$\widehat{MSE}(b_0, b_1) \equiv \frac{1}{n} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

$$\widehat{MSE}(b_0, b_1) \to MSE(b_0, b_1) \text{ as } n \to \infty$$

Method of Least Squares

$$\widehat{MSE}(b_0, b_1) \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$

$$\frac{\partial \widehat{MSE}}{\partial b_0} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))(-2)$$

$$\frac{\partial \widehat{MSE}}{\partial b_1} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))(-2x_i)$$

At the optimum (by setting to zero), we get the **normal/estimating equations** of the least squares method

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

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

Method of Least Squares(2)

$$\boxed{\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}}$$

$$\begin{aligned} 0 &= \overline{xy} - \bar{y}\bar{x} + \hat{\beta}_1 \bar{x}\bar{x} - \hat{\beta}_1 \overline{x^2} \\ 0 &= c_{XY} - \hat{\beta}_1 s_X^2 \\ \hat{\beta}_1 &= \frac{c_{XY}}{s_X^2} \end{aligned}$$

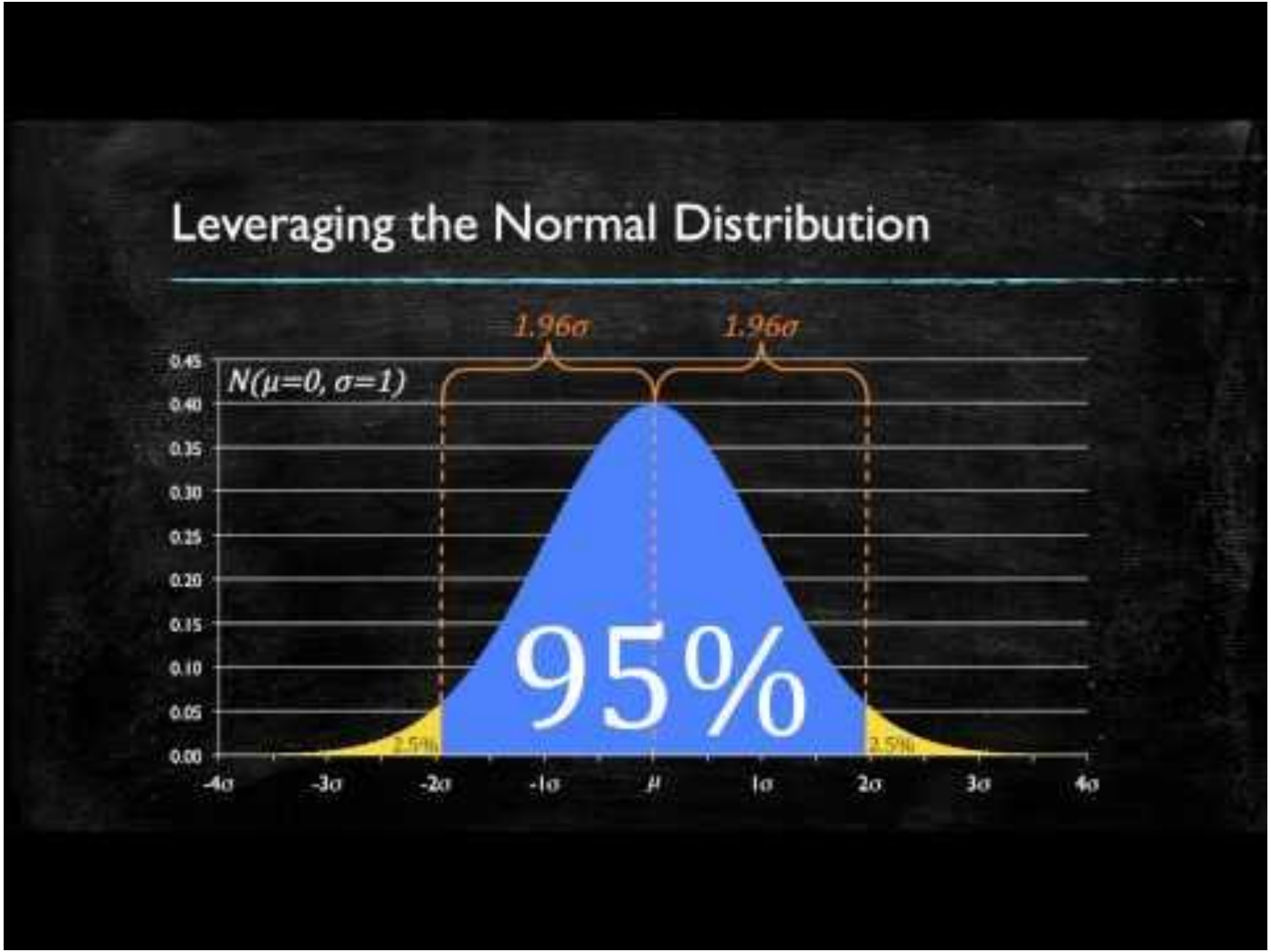
$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

$$C_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Interpreting the coefficients, confidence intervals and the statistical significance



Leveraging the Normal Distribution
1.96σ
1.96σ
N(μ=0, σ=1)
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
95%
2.5%
2.5%
-4σ
-3σ
-2σ
-1σ
μ
1σ
2σ
3σ
4σ

Prediction & inference using linear regression

Constructing the Confidence Interval

- Compute the point estimate of the forecast:

$$\begin{aligned} Consumption &= 49.1334 + 0.8528\ Income + \varepsilon \\ &= 49.1334 + 0.8528 \times 100 + 0 \\ &= 134.4070 \end{aligned}$$

Linear regression: sample code

```
from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=2)

from sklearn.linear_model import LinearRegression

lr = LinearRegression()

lr.fit(X_train,y_train)

y_pred = lr.predict(X_test)
```

[Sklearn documentation]()

Multiple linear regression

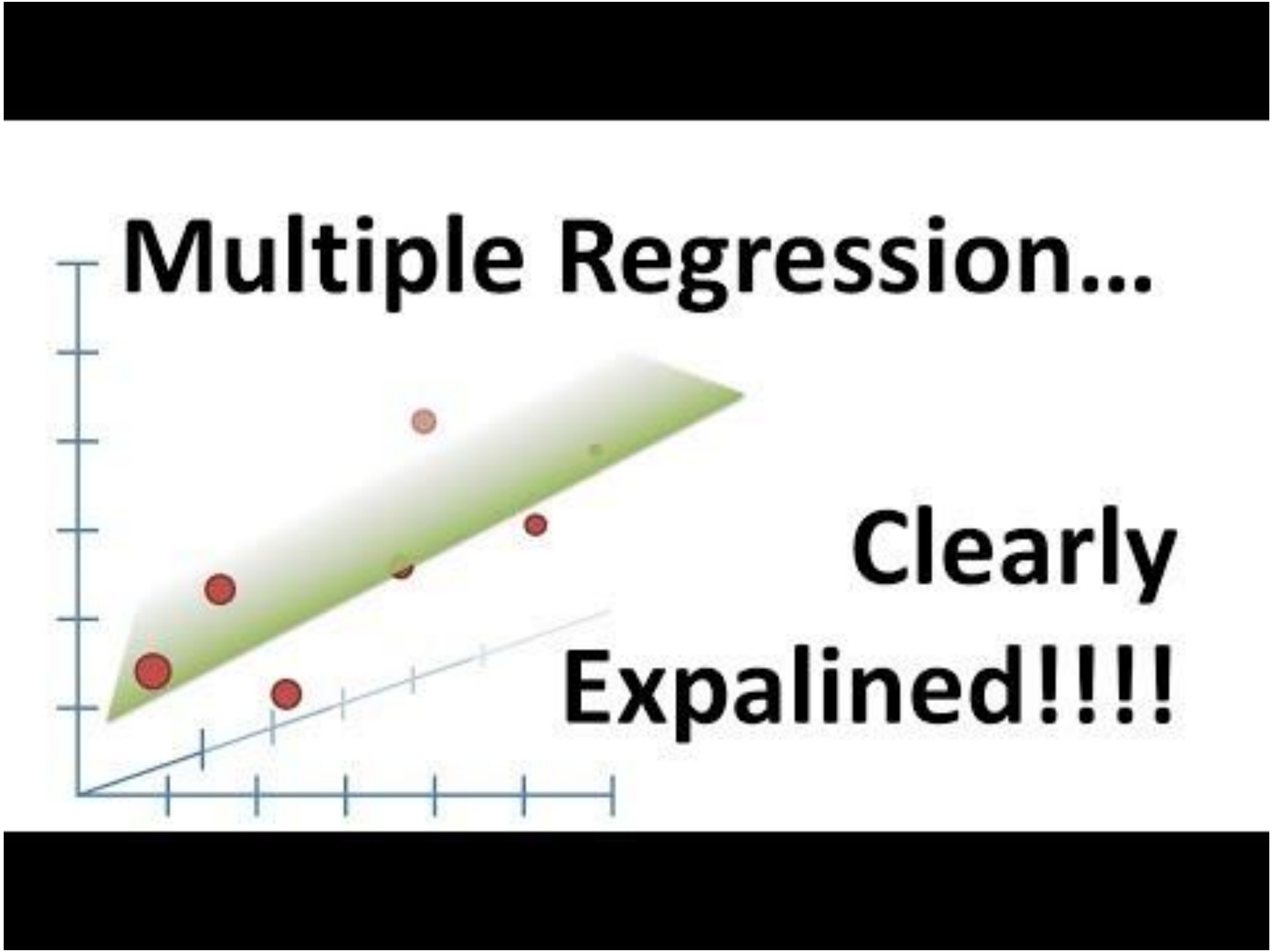
Multivariate/Multiple linear regression

When we have to map multiple observations x_1, x_2, …, x_n to an observation/target y (real number), it is referred to as multivariate regression or multiple linear regression

E.g. what is the temperature tomorrow (y) given today's temperature (x_1), yesterday's temperature (x_2), temperature from two days before (x_3)

$$y = f(x_1, x_2, \ldots x_n) = b_0 + x_1 b_1 + \ldots + x_n b_n$$

The regression coefficients b_0 and b_1, b_2,..., b_n should be learned from the data
For that, we can still use the Least Squares method, just like in simple linear regression.


Multiple Regression...
Clearly
Expalined!!!!

Model Evaluation

Evaluating a regression model

- No model is perfect. We want to know how good it is
- There is no single perfect evaluation metric
    - We use a combination of metrics to analyse different aspects of a model

Some evaluation metrics

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2
- Adjusted R^2

Mean Squared Error

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

- Advantages
  - Differentiable. Therefore can be used as a loss function for optimization tasks
- Disadvantages
  - Units are squared (wrt target variable)
  - Not robust against outliers because of the squaring

Sklearn documentation

Mean Absolute Error (MAE)

$$\frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y_i}\right|$$

- Advantages
  - Has the same units as the target variable
  - Robust to outliers
- Disadvantages
  - Not differentiable. Could be a problem for optimization (finding the best line/hyperplane, using numerical methods)

[Sklearn documentation]

R^2 score (coefficient of determination)

- Provides information about the **goodness of fit** of a model
- Statistical measure of how well the regression line approximates the actual data
- To calculate this:

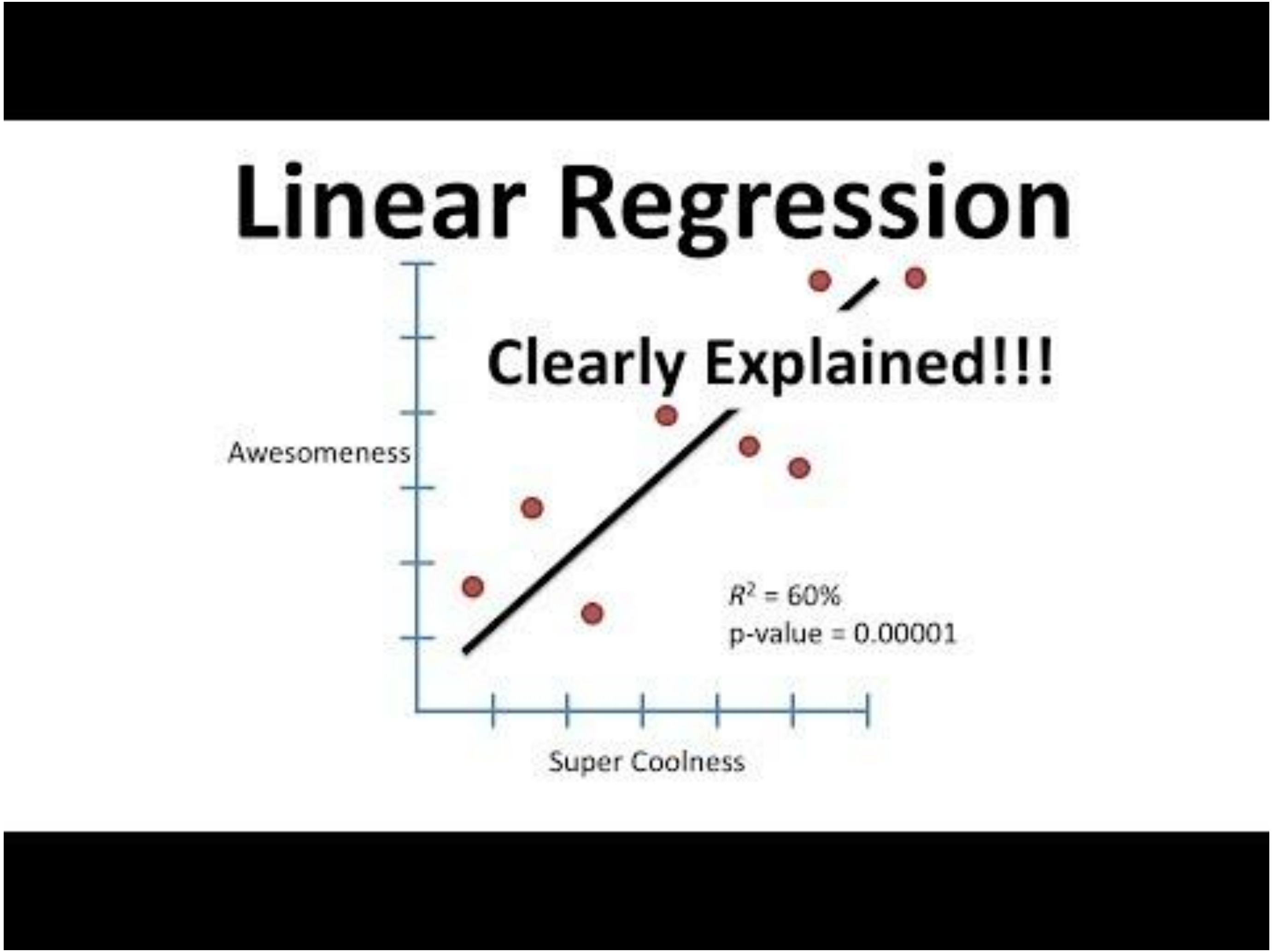
$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

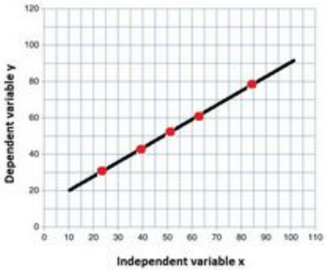
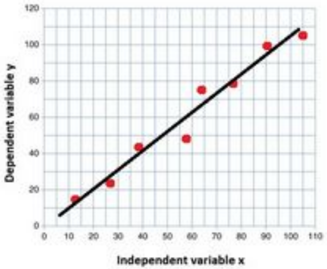
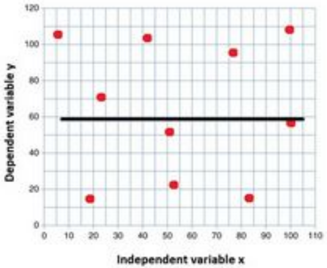
Unexplained variation is captured by the Residual Sum of Squares. **Total Variation** is the Total Sum of Squares

[Sklearn documentation]

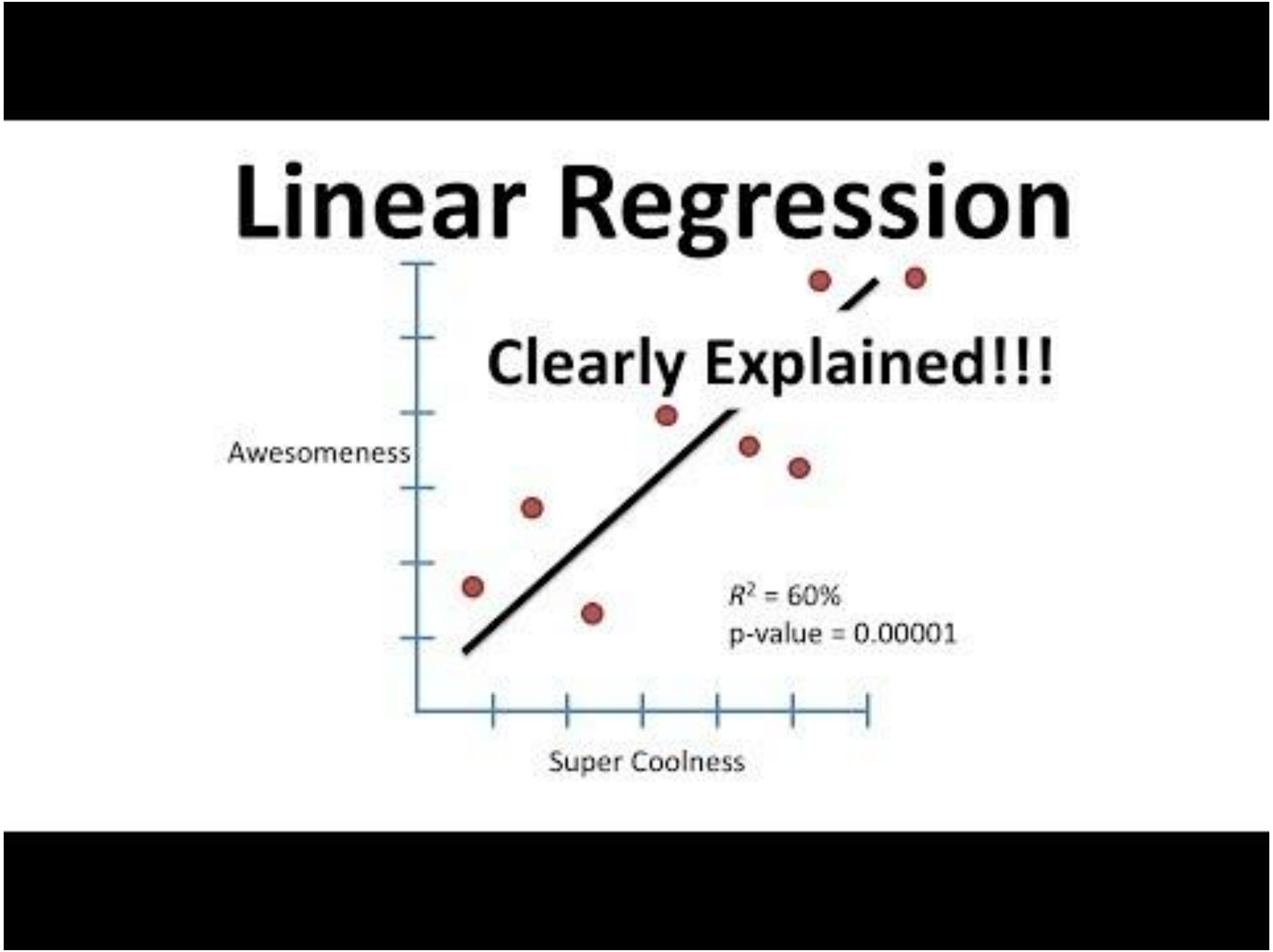
More on R^2


Linear Regression
Clearly Explained!!!
Awesomeness
Super Coolness
R^2 = 60%
p-value = 0.00001

Interpreting R^2

R^2 Values	Interpretation	Graph
$R^2 = 1$	All the variation in the y values is accounted for by the x values	
$R^2 = 0.83$	83% of the variation in the y values is accounted for by the x values	
$R^2 = 0$	None of the variation in the y values is accounted for by the x values	

Limitation of R^2


Linear Regression
Clearly Explained!!!
Awesomeness
Super Coolness
R^2 = 60%
p-value = 0.00001

Adjusted R^2

- The drawback of R2 is that when adding new features (multiple linear regression), R2 either remains constant or increases
  - Remains true, even when we add irrelevant features!

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$$

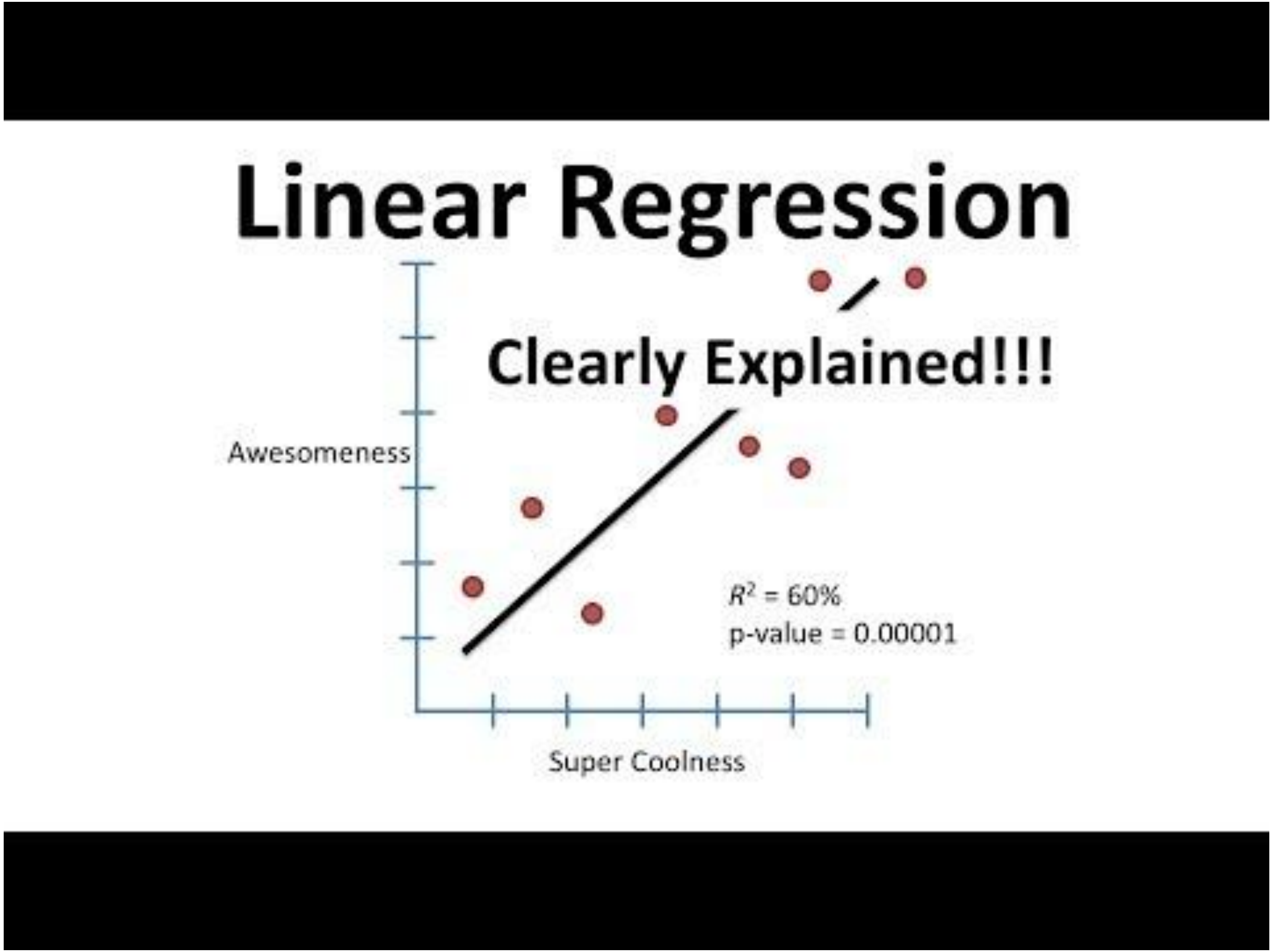
where:

n = number of observations

k = number of independent variables

R_a^2 = adjusted R^2

More on p-values and confidence intervals


Linear Regression
Clearly Explained!!!
Awesomeness
R^2 = 60%
p-value = 0.00001
Super Coolness

Steps in linear regression (summary)

1. Use least squares method to fit a line/plane/hyperplane to the data
2. Evaluate the model
   a. MSE, MAE
   b. Calculate the goodness of fit $\mathbf{R^2}$
   c. Calculate the **p-value** for R^2
   d. If R2 and p-value suggest the model is accurate enough, we can use it to predict unobserved target values y_i, given x_i
3. If you are happy with the performance, use it to predict unobserved values

Other types of regression…

Variations of linear regression, with regularization

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

Linear regression (Ordinary Least Squares or OLS), minimize:

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

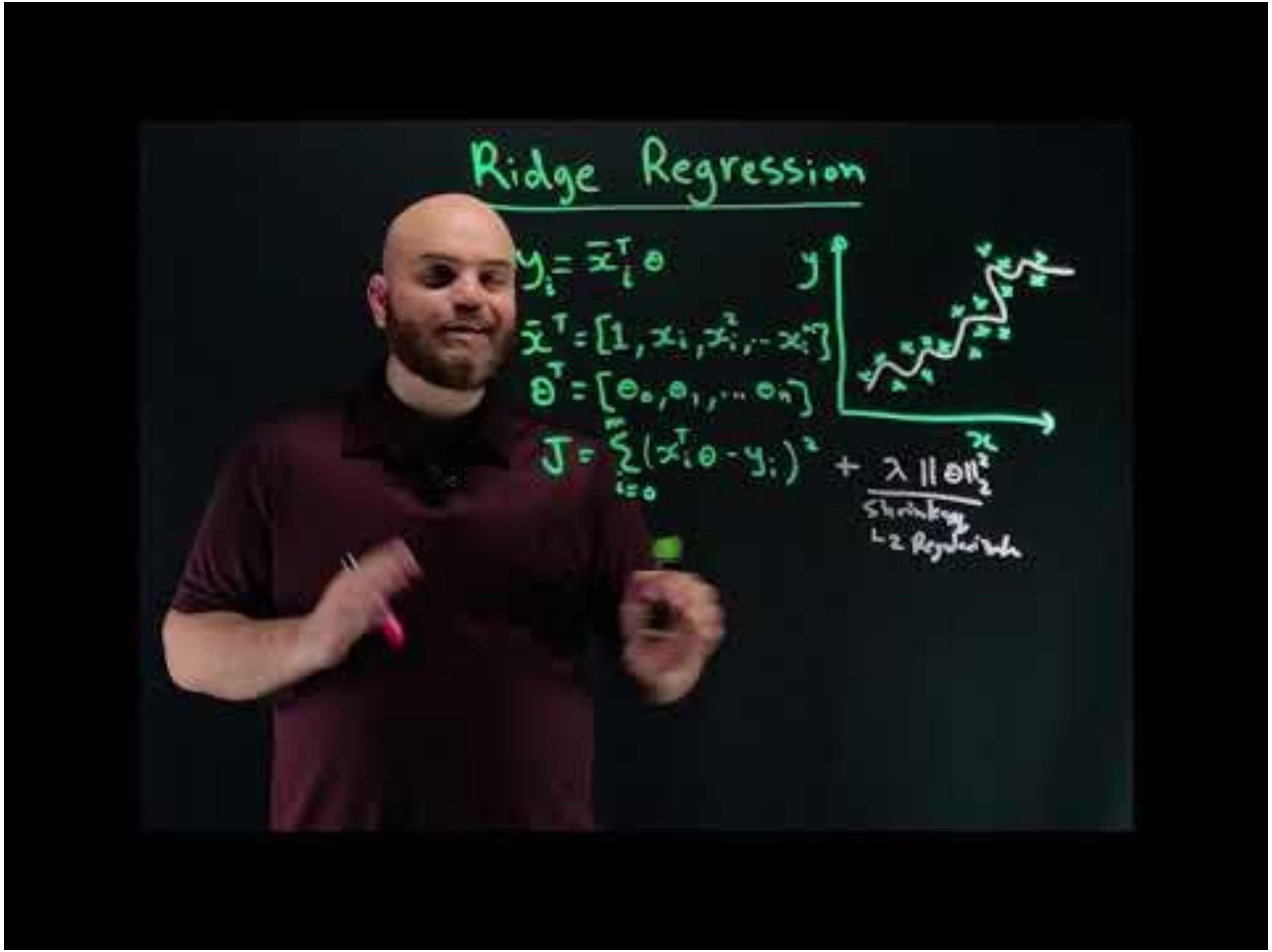
Ridge regression, minimize:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 + \lambda\sum_{j=i}^{n} w_j^2$$

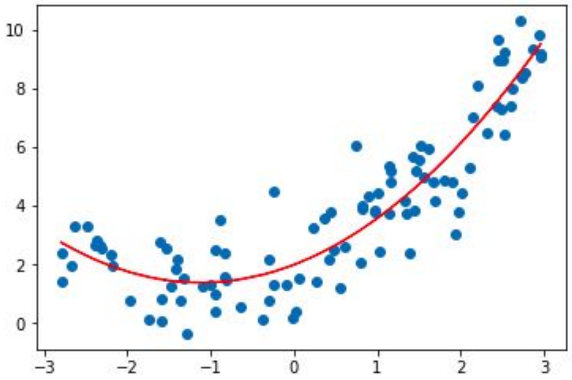
Lasso regression, minimize:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 + \lambda\sum_{j=i}^{n}|w_j|$$

Ridge regression



Polynomial regression

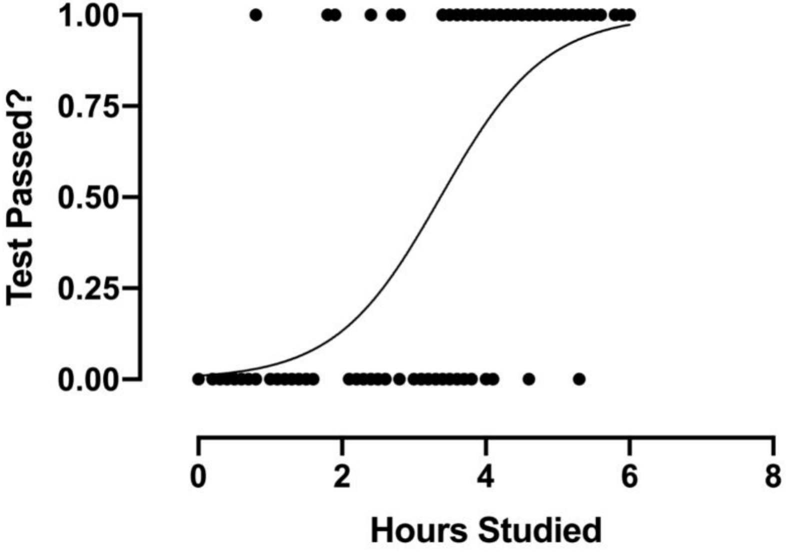


- Linear regression can only handle linear relationships between the target variable y and the independent variables x_i
- To overcome this limitation, we can add polynomial terms such as x^2, x^3,...,x^n to the regression equation, to model any non-linear relationships between y and x_i

$$y = f(x_1, x_2, \ldots x_n) = b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n$$

The regression coefficients b_0 and b_1, b_2,..., b_n should be learned from the data

Logistic regression



- When the target value y is a discrete value (true/false, 1/0)
- f(x) computes the probability of true/false

Demo

- Linear regression sklearn coding example
- Demo video: python coding tutorial

Summary

- What is regression?
- Linear regression
- Simple linear regression
- Least squares method to estimate the regression coefficients
    - Ordinary Least Squares (OLS) regression
- Multiple linear regression
- Evaluating a regression model
    - MSE, MAE, R^2, adjusted R^2, p-values
- Other linear regression methods (with regularization)
    - Ridge
    - Lasso
- Polynomial regression (non-linear)
- Logistic regression