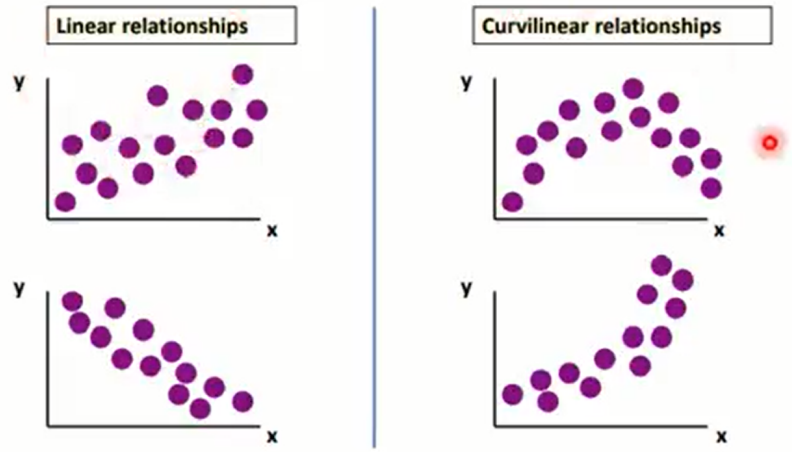
REGRESSION ANALYSIS

Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable
- Dependent variable: the variable we wish to explain
- Independent variable: the variable used to explain the dependent variable

- A scatter plot is used to show the relationship between two variables
- Correlation analysis is used to measure strength of the association(linear) between two variables

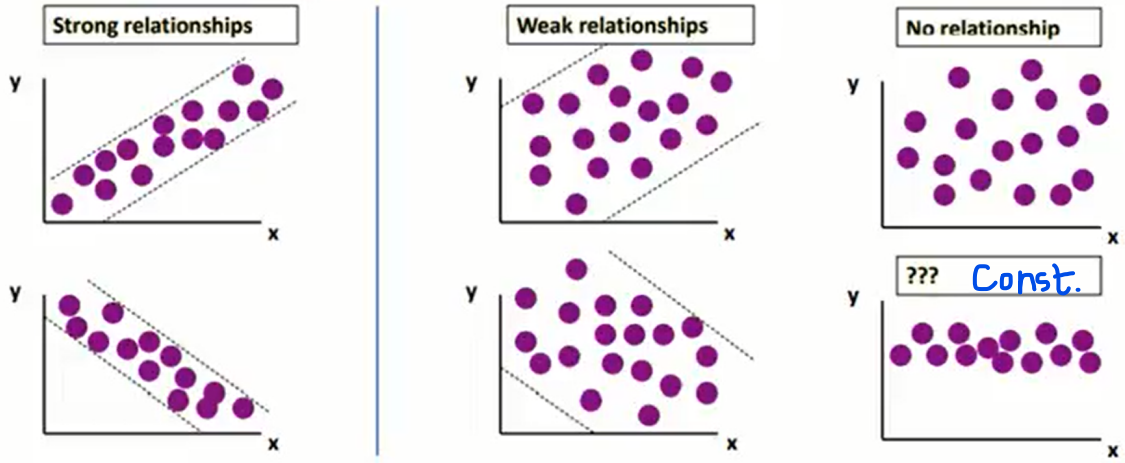
Scatter Plot -- Relationship



Correlation Analysis

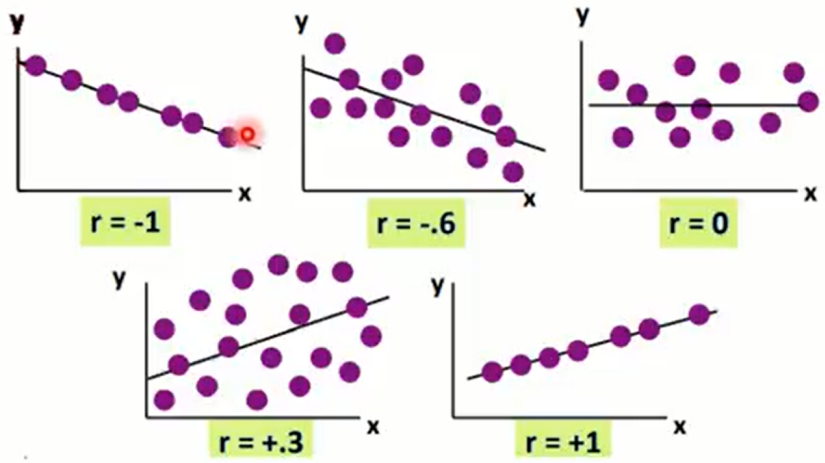
- The population correlation coefficient 'ρ' measures the strength of the association/relationship between two variables
- The sample correlation coefficient 'r' is an estimate of ρ and is used to measures the strength of the linear relationship in the sample observations

Scatter Plot -- Correlation



Correlation Coefficients -- ρ and r

- Unit free
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship



regression analysis.mp4

Estimated Regression Line/Model

- The sample regression line gives an estimate of the population regression line

Estimated (or predicted) y value

Estimate of the regression intercept

Estimate of the regression slope

Independent variable

$$\hat{y}_i = b_0 + b_1 x$$

- The individual random error has mean of zero

15:49 / 59:29

Least Squares Criterion

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared residuals

$$\sum e^2 = \sum (y - \hat{y})^2$$
$$= \sum (y - (b_0 + b_1 x))^2$$

Least Square Equations

The formulas for b_1 and b_0 are:

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

algebraic equivalent:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

Interpretation – slope and intercerpt

- b_0 is the estimated average value of y when the value of x is zero.
- b_1 is the estimated change in the average value of y as a result of a one-unit change in x.
- They are presumed constant in the population, so that the effect of a one-unit change in x on y is assumed constant for all values of x.

Understanding bi

- Change in *y* as a function of unit change in x_i
  - all other things being equal
- Example: income in units of \$10K, years in age, b_{age}= 2
  - For the same gender, years of education, and state of residence, a person's income increases by 2 units (20K)for every year older

$$\text{income} = b_0 + b_1\text{age} + b_2\text{yearsOfEducation} + b_3\text{gender} + b_4\text{state}$$

0:30:28 0:29:01

What formula tells??

Example

Find out the regression line that best explain the given data.

- ❑ Find b0
- ❑ Find b1
- ❑ Find line

$$b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} \qquad b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

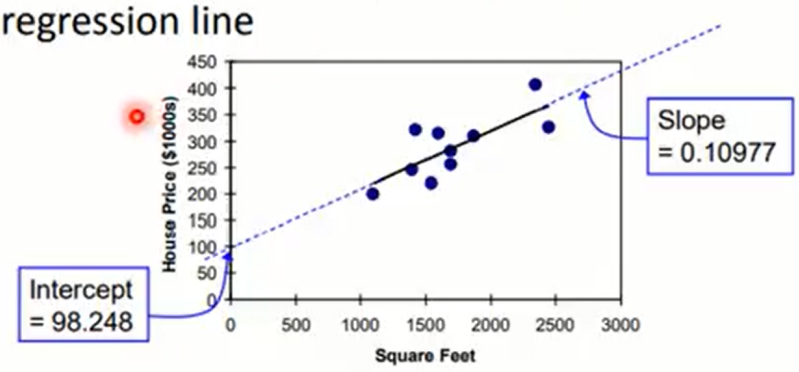
Sample Data for House Price Model

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Regression Line Obtained

houseprice = 98.24833 + 0.10977(squarefeet)

- House price model: scatter plot and regression line



(1) Interpreting – b_0

$$\widehat{\text{houseprice}} = 98.24833 + 0.10977(\text{squarefeet})$$

intercept

b_0 is the estimated average value of Y when the value of X is zero (if x = 0 is in the range of observed x values)

- Here, no houses had 0 square feet, so b_0 = 98.24833 just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

(2) Interpreting – b_1

$$\widehat{\text{houseprice}} = 98.24833 + 0.10977(\text{squarefeet})$$

slope

b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X

- Here, b_1 = .10977 tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Variation – Explained and Unexplained

Total variation is made up of two parts:

$$SST = SSE + SSR$$

Total sum of Squares	Sum of Squares Error	Sum of Squares Regression
$SST = \sum(y - \bar{y})^2$	$SSE = \sum(y - \hat{y})^2$	$SSR = \sum(\hat{y} - \bar{y})^2$

Unexplained

Explained

where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

- The explained variation can be explained by the relationship between x and y.
- The unexplained variation cannot be explained by the relationship between x and y and is due other variables.

Variation??

Model Evaluation – Coefficient of Determination

$$R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$

Note: In the single independent variable case, the coefficient of determination is

$$R^2 = r^2$$

where:

R^2 = Coefficient of determination

r = Simple correlation coefficient

regression analysis.mp4

- The proportion of variation explained by the model is called the coefficient of determination.
- Let SSR = 220.9 and TSS = 256, then

  R^2=SSR/TSS = 220.9/256 =0.8629
- In other words, estimated model can predict about 86% of the variation in y

49:00 / 59:29

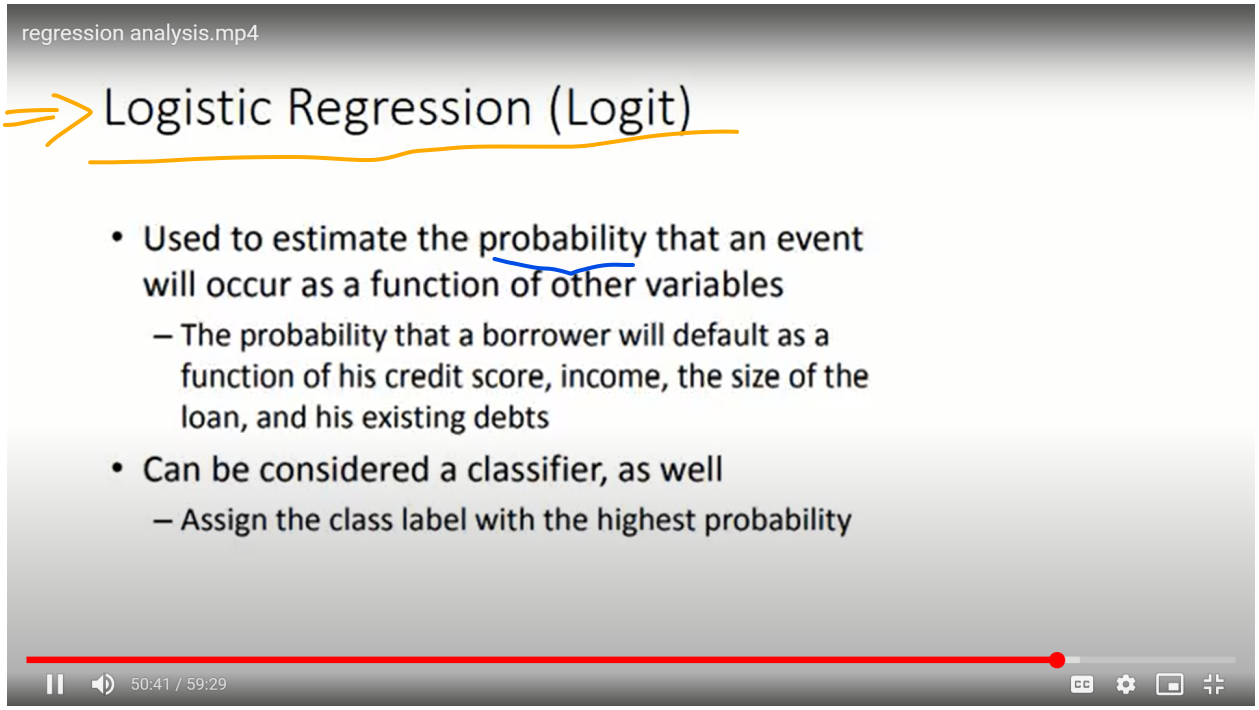
regression analysis.mp4

General Linear Regression Model

- For p independent variables

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

50:33 / 59:29

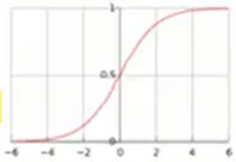


Logit Model

$$\text{default} = f(\text{creditScore}, \text{income}, \text{loanAmt}, \text{existingDebt})$$

- Training data: default is 0/1
  - default=1 if loan defaulted
- The model will return the probability that a loan with given characteristics will default

$$\ln \frac{P(y=1)}{1-P(y=1)} = b_0 + b_1 x_1 + b_2 x_2 \ldots$$



- y=1 is the case of interest: 'TRUE'
- LHS is called logit(P(y=1))
  - hence, "logistic regression"
- logit(P(y=1)) is inverted by the sigmoid function
  - standard packages can return probability for you
- Categorical variables are expanded as with linear regression