BIG DATA STATISTICS FOR BUSINESS

AY 2023-24

Giovanni De Luca
Parthenope University of Naples

1

Example of multiple regression (two predictors)

- Suppose we want to study the consumption in mpg for some automobiles as a function of horsepower and displacement
- Y = mpg
- X_1 = horsepower
- X_2 = displacement

2

2

Multiple linear regression (two predictors)

- Multiple linear regression with two predictors is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

- Y = mpg (response variable)
- X_1 = horsepower (1st predictor)
- X_2 = displacement (2nd predictor)
- β_0, β_1 and β_2 are parameters of the model
- e = error
- n statistical units

Error

- e is the error which prevents from defining a deterministic relationship between Y and the predictors.
- The error e is a continuous random variable with an average value equal to 0.

Estimate

- Parameters β_0, β_1 and β_2 are estimated using the Ordinary Least Squares (OLS) method.

Predicted values

- Predicted values are given by

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$

- This is not a regression line, but a regression plane!

Parameters

- Parameter b_0 (also known as intercept) is the predicted value of Y when $X_1 = X_2 = 0$.
- Parameter b_1 is the predicted change of Y when X_1 increases by one unit, if X_2 remains constant.
- If $b_1 > 0$ (< 0), there is a positive (negative) association between Y and X_1.
- Parameter b_2 is the predicted change of Y when X_2 increases by one unit, if X_1 remains constant
- If $b_2 > 0$ (< 0), there is a positive (negative) association between Y and X_2.

Inference on the parameters

- *t*-test
  $H_0 : \beta_1 = 0$ (lack of association between Y and X_1)
  $H_1 : \beta_1 \neq 0$ (positive or negative association between Y and X_1)

- *t*-test
  $H_0 : \beta_2 = 0$ (lack of association between Y and X_2)
  $H_1 : \beta_2 \neq 0$ (positive or negative association between Y and X_2)

Inference on the parameters

- *F*-test

  $H_0 : \beta_1 = \beta_2 = 0$ (No overall significance in regression)

  H_1 : Overall significance in regression

9

9

Adjusted R-squared

- The goodness of fit is measured using $\bar{R}^2$ which is a measure of the fit but also takes into account the number of the parameters of the model (model complexity).

10

10

Example

- Let's try to predict the variable *mpg* using *horsepower* and *displacement*.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

$$\widehat{mpg} = b_0 + b_1 horsepower + b_2\ displacement$$

11

11

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.469488   0.727716  51.489  < 2e-16 ***
horsepower   -0.058275   0.013491  -4.319 1.99e-05 ***
displacement -0.040818   0.004963  -8.225 2.95e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.534 on 389 degrees of freedom
Multiple R-squared:  0.6643,	Adjusted R-squared:  0.6626 
F-statistic: 384.9 on 2 and 389 DF,  p-value: < 2.2e-16
```

12

12

Predicted values

$horsepower = 100$

$displacement = 150$

$\widehat{mpg}(100,150) = 37.47 - 0.06 \cdot 100 - 0.04 \cdot 150 = 25.52$

13

13

Multiple linear regression (p predictors)

- Multiple linear regression with p predictors is given by

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$$

- Y = response variable
- X_1 = 1st predictor
- …
- X_p = *p-th* predictor
- $\beta_0, \beta_1, \ldots, \beta_p$ are the parameters of the model

14

14

Model selection

- To better predict the response variable according to the values of some predictors, we have to search for the best model, that is the best subset of predictors.
- Selection is usually based on the Adjusted $\bar{R}^2$ or the Akaike Information Criterion (AIC): we select the model with the highest $\bar{R}^2$ or the lowest AIC.
- The function *step* in *R* uses the AIC.

15

15

Model selection

- AIC uses the Sum of Squared Errors (*SSE*)

$$AIC = n\log(SSE) + 2(p+1)$$

- n is the number of observations
- p is the number of predictors
- SSE is given by

$$SSE = \sum (Y - \hat{Y})^2$$

16

16