BIG DATA STATISTICS FOR BUSINESS

AY 2023-24

Giovanni De Luca
Parthenope University of Naples

1

Supervised learning

- Many problems of statistical learning require a supervised learning technique.
- *Supervised learning*: for each statistical unit we know the response variable, Y and p predictors $X_1, X_2, \dots, X_p$.
- The interest lies in the analysis of the relationship between the predictors and the response variable with the aim of predicting the latter for new observations.

2

2

Unsupervised learning

- Unsupervised learning: for each statistical unit we have *p* variables, but no response variable is observed.
- In this case, the goal is the study of the relationship between the variables or between the observations, or grouping the observations into distinct groups.

3

3

Supervised learning techniques

- Classical linear regression, polynomial regression, logistic regression.

4

4

Linear regression

- Linear regression is a useful and widely used statistical learning method.
- Linear regression analyzes the response of a numerical variable (response variable or dependent variable) to *p* numerical variables (predictors or explanatory variables).
- Simple linear regression: 1 predictor.
- Multiple linear regression: *p* predictors.

Simple linear regression

- In simple linear regression, we denote

  Y = response variable

  X = predictor

Example of simple linear regression

- Suppose we want to study the consumption in mpg for some automobiles as a function of horsepower.
- Y = mpg
- X = horsepower

Simple linear regression

- Simple linear regression is given by

$$Y = \beta_0 + \beta_1 X + e$$

- Y = mpg (response variable)
- X = horsepower
- β_0 and β_1 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 predictor.
- Deterministic relationship between Y and X: only one value of Y is associated to a specific value of X.
- The error e is a continuous random variable with null average.

9

9

Estimate

- Parameters β_0 and β_1 are estimated using the Ordinary Least Squares (OLS) method.
- OLS method identifies the values of β_0 and β_1 which minimize the sum of the squares of the differences between observed values Y and the predicted values $\hat{Y}$,

$$min \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

- Predicted values are $\hat{Y} = b_0 + b_1 X$
- b_0 and b_1 are the estimates of the parameters β_0 and β_1
- $\hat{Y} = b_0 + b_1 X$ is equation of the regression line.

10

10

Parameters

- Parameter b_0 (also known as intercept) is the predicted value of Y when X = 0.
- Parameter b_1 is the predicted change of Y when X increases by one unit.
- If $b_1 > 0$ (< 0), there is a positive (negative) association between $\hat{Y}$ and X.

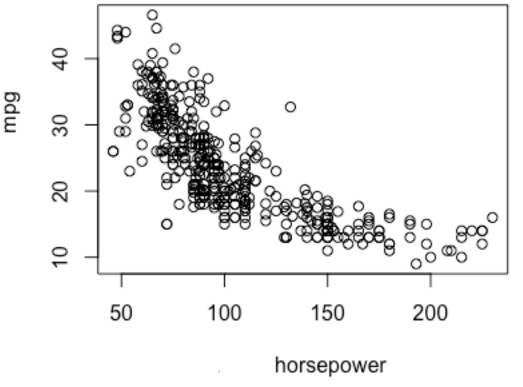
11

11

12

12

Example



$$Y = \beta_0 + \beta_1 X + e$$

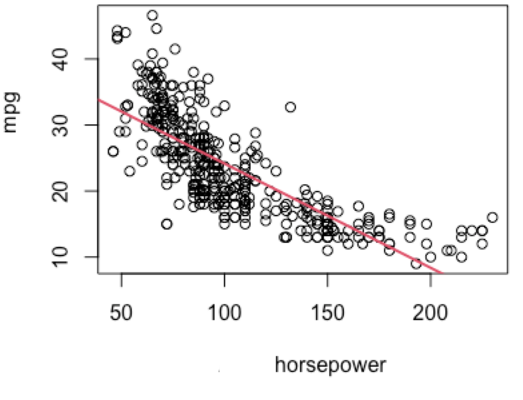
$$\hat{Y} = b_0 + b_1 X$$

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

13

13

Example



$$\hat{Y} = 39.9 - 0.16\,X$$

$$\widehat{mpg} = 39.9 - 0.16\,horsepower$$

14

14

Inference on the parameter

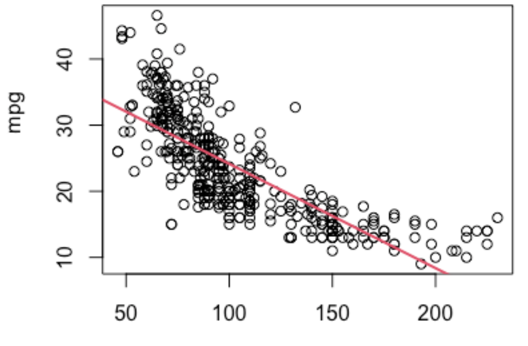
- *t*-test
  $H_0 : \beta_1 = 0$ (lack of association between Y and X)
  $H_1 : \beta_1 \neq 0$ (positive or negative association between Y and X)

- Look at the *p*-value.
- We reject the H_0 if the *p*-value of the test is less than 0.05,

$$p\text{-value} < 0.05$$

15

15




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,	Adjusted R-squared:  0.6049 
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

16

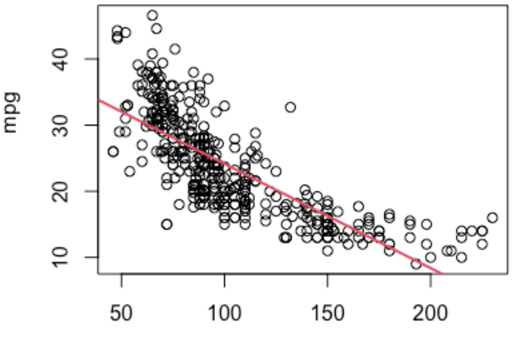
16

R-squared

- R^2 is a statistical measure of fit (how close the predicted values of the model are to the observed data).
- $0 \leq R^2 \leq 1$
- Simple rule: the higher R^2, the better the model fits the data.
- After multiplying by 100, it provides the percentage of the response variable variation that is explained by the linear model (between 0 and 100%).
- It is also known as coefficient of determination.

17

17




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,	Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

18

18

Prediction

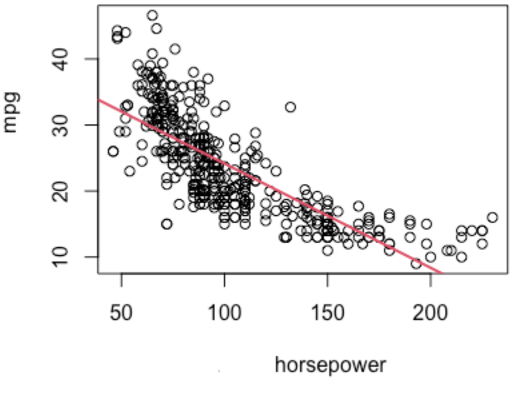
- The prediction for new data is easily obtained applying the regression equation.
- For the value x, the prediction is

$$\hat{Y}(x) = b_0 + b_1 x$$

19

19

Predicted values

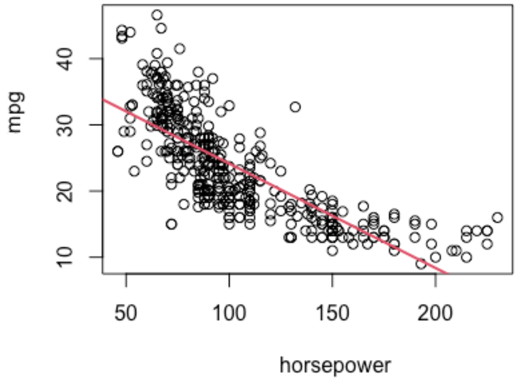


- $horsepower = 70$
- $\widehat{mpg}(70) = 39.9 - 0.16 \cdot 70 = 28.7$

20

20

Predicted values



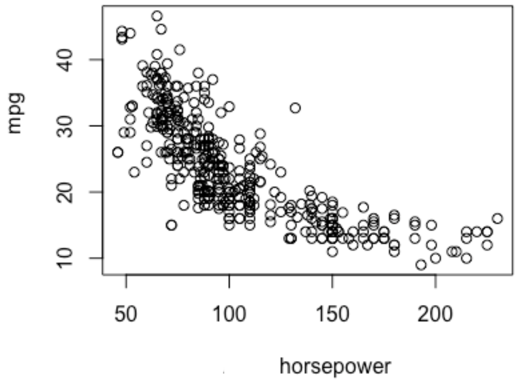
- $horsepower = 100$
- $\widehat{mpg}(100) = 39.9 - 0.16 \cdot 100 = 23.9$

21

Polynomial regression

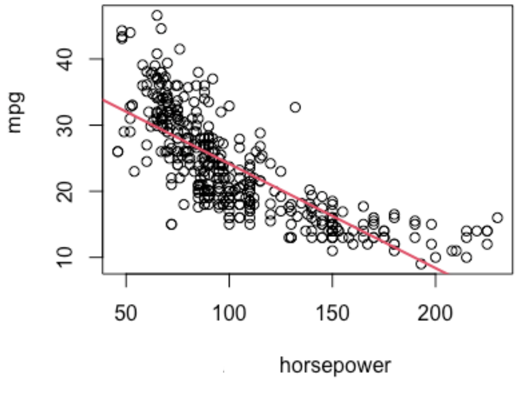
- Linear regression involves a linear relationship between the response variable and the predictors.
- Sometimes, the relationship is not linear.
- Polynomial regression is a simple strategy to extend a linear model to capture a non-linear relationship.
- In practice, polynomial regression adds further terms given by some power of the original predictors.
- Non-linear relationship: relationship not adequately represented by a straight line.

22



23

23



$$\hat{Y} = 39.9 - 0.16\,X$$

$$\widehat{mpg} = 39.9 - 0.16\; horsepower$$

24

24

Polynomial regression (order 2)

- Polynomial regression of order 2 is given by

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e$$

$$\hat{Y} = b_0 + b_1 X + b_2 X^2$$

$$\widehat{mpg} = b_0 + b_1 horsepower + b_2 horsepower^2$$

25

25

Polynomial regression (order 3)

- Polynomial regression of order 3 is given by

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + e$$

$$\hat{Y} = b_0 + b_1 X + b_2 X^2 + b_3 X^3$$

$$\widehat{mpg} = b_0 + b_1 horsepower + b_2 horsepower^2 + b_3 horsepower^3$$

26

26

Polynomial regression (order *d*)

- In general, polynomial regression of order *d* is given by

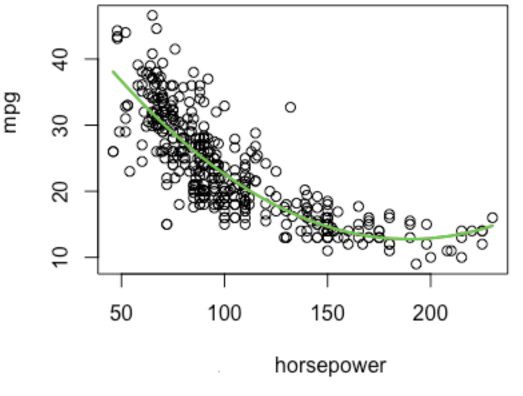
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + e$$

$$\hat{Y} = b_0 + b_1 X_i + b_2 X^2 + \cdots + b_d X^d$$

- *d* is usually not larger than 3.

27

27



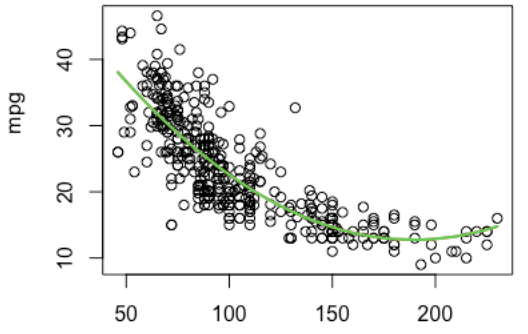
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + e$$

$$\hat{Y} = b_0 + b_1 X + b_2 X^2$$

$$\widehat{mpg} = b_0 + b_1 horsepower + b_2\, horsepower^2$$

28

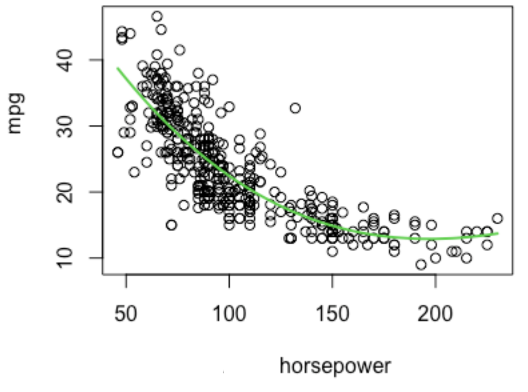
28




```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                56.9000997  1.8004268   31.60   <2e-16 ***
poly(horsepower, 2, raw = T)1 -0.4661896  0.0311246  -14.98   <2e-16 ***
poly(horsepower, 2, raw = T)2  0.0012305  0.0001221   10.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.374 on 389 degrees of freedom
Multiple R-squared:  0.6876,    Adjusted R-squared:  0.686 
F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
```

29

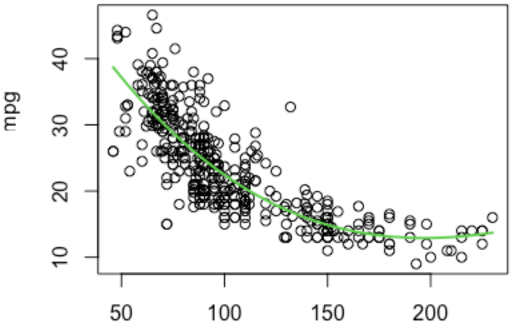


$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + e$$

$$\hat{Y} = b_0 + b_1 X + b_2 X^2 + b_3 X^3$$

$$\widehat{mpg} = b_0 + b_1 horsepower + b_2\, horsepower^2 + b_3\, horsepower^3$$

30




```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.068e+01  4.563e+00  13.298  < 2e-16 ***
poly(horsepower, 3, raw = T)1 -5.689e-01  1.179e-01  -4.824 2.03e-06 ***
poly(horsepower, 3, raw = T)2  2.079e-03  9.479e-04   2.193   0.0289 *
poly(horsepower, 3, raw = T)3 -2.147e-06  2.378e-06  -0.903   0.3673
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.375 on 388 degrees of freedom
Multiple R-squared:  0.6882,    Adjusted R-squared:  0.6858
F-statistic: 285.5 on 3 and 388 DF,  p-value: < 2.2e-16
```

31

Order of the polynomial

- The selection of the order of the polynomial is usually based on Adjusted R^2 ($\bar{R}^2$).
- $\bar{R}^2$ is a measure of the fit that replace R^2 (which always increases with the addition of further predictors).
- $\bar{R}^2$ takes into account the fit but also the number of the parameters of the model (model complexity).
- $\bar{R}^2$ may decrease upon adding a new predictor if this is poorly relevant (a negligible increase of R^2).
- For the selection of the order of the polynomial, we select the model with the highest $\bar{R}^2$.

32

32

Model selection

- The formulation uses the Sum of Squared Errors (*SSE*)

$$\bar{R}^2 = 1 - \frac{SSE(n-1)}{SST(n-d-1)}$$

- n is the number of observations
- d is the order of the polynomial
- SSE is given by

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

33

33

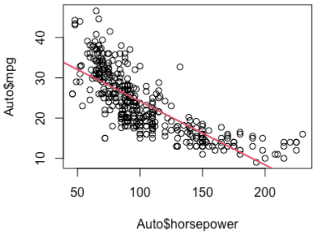
Improvement of the predictions

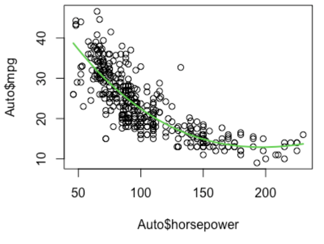
- The choice of a better model involves an improvement of the predictions.

34

34

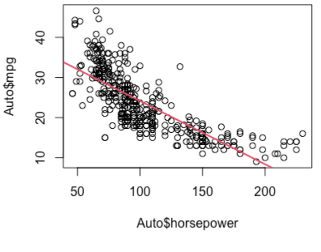
Improvement in predictions

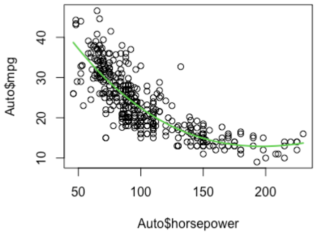




- $horsepower = 50$
- $\widehat{mpg}(50) = 39.9 - 0.16 \cdot 50 = 32.04$
- $\widehat{mpg}(50) = 56.9 - 0.466 \cdot 50 + 0.001 \cdot 50^2 = 36.67$

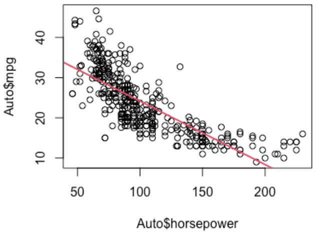
Improvement in predictions

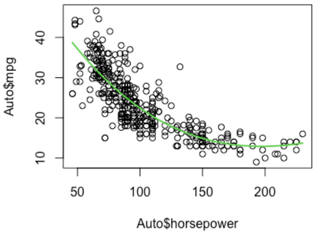




- $horsepower = 100$
- $\widehat{mpg}(100) = 39.9 - 0.16 \cdot 100 = 24.15$
- $\widehat{mpg}(100) = 56.9 - 0.466 \cdot 100 + 0.001 \cdot 100^2 = 22.58$

Improvement in predictions





- $horsepower = 200$
- $\widehat{mpg}(200) = 39.9 - 0.16 \cdot 200 = 8.37$
- $\widehat{mpg}(200) = 56.9 - 0.466 \cdot 200 + 0.001 \cdot 200^2 = 12.88$

37

37