Linear regression model

Content

- Linear regression model
- Ordinary least squares
- R^2
- F-ratio and p-values

Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \varepsilon_i$$

- y is the target variable.
- x are the predictor variables.
- β are the coefficients.
- β_0 is the intercept.

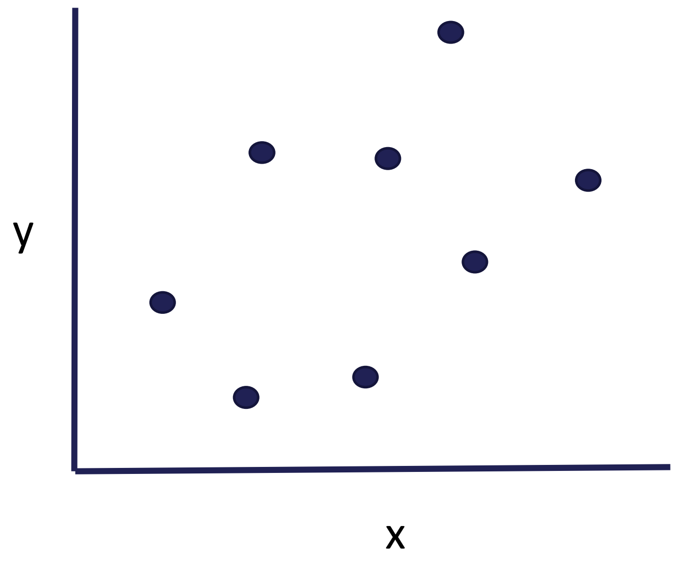
Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \varepsilon_i$$

- ε_i is the difference between the predicted and the observed value of y.
- ε is normally distributed and centred at 0.

Ordinary least squares (OLS)

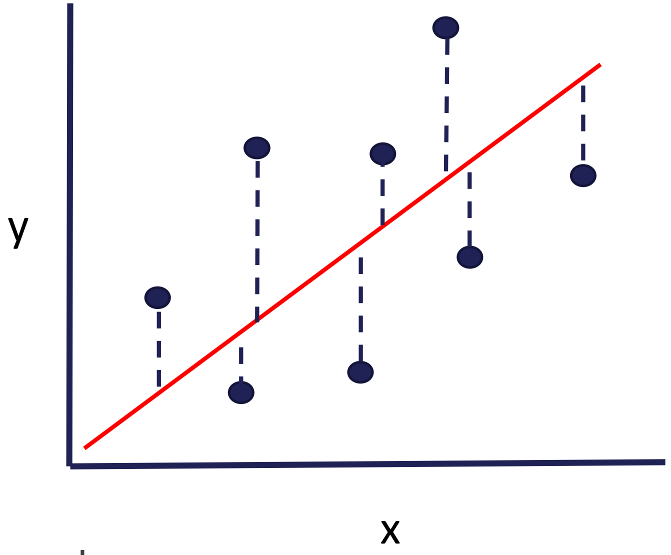
OLS: method to find the line that best fits the data.



Ordinary least squares (OLS)

$$\beta = \min \sum (y - predictions)^2$$

$$\beta = \min \sum \left(y - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni})\right)^2$$



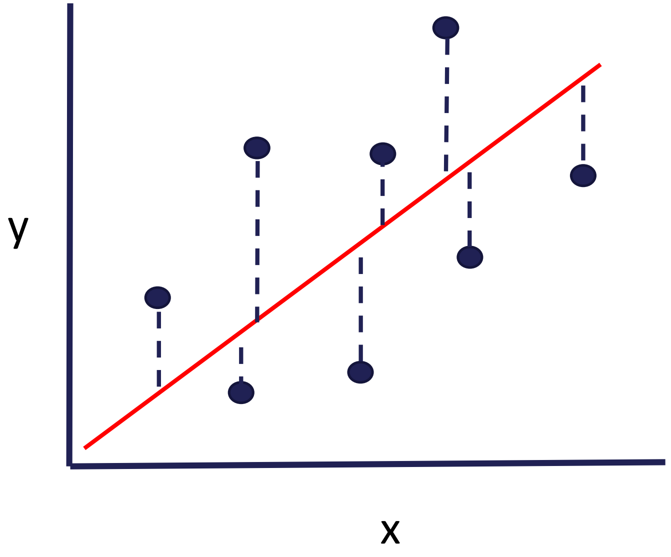
OLS ➔ Find the coefficients that minimize the squared difference of the target and the predictions.

Ordinary least squares (OLS)

Residuals: difference between the target and the predictions (---).

$$residuals = y - predictions = \varepsilon$$

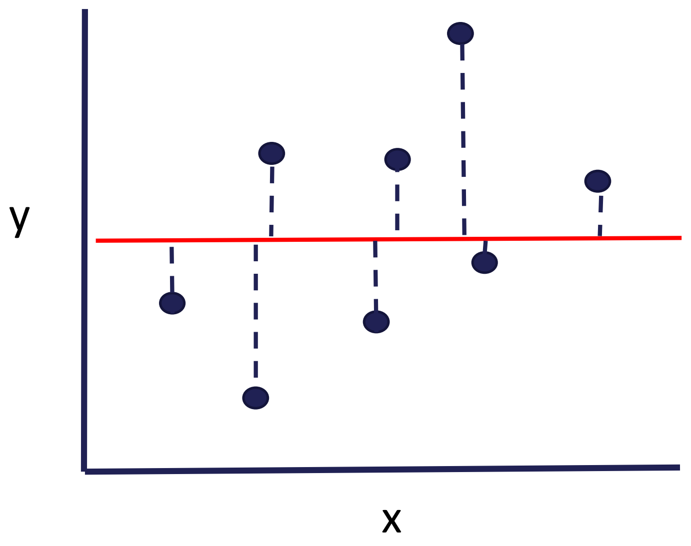
$$residuals = y - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni})$$

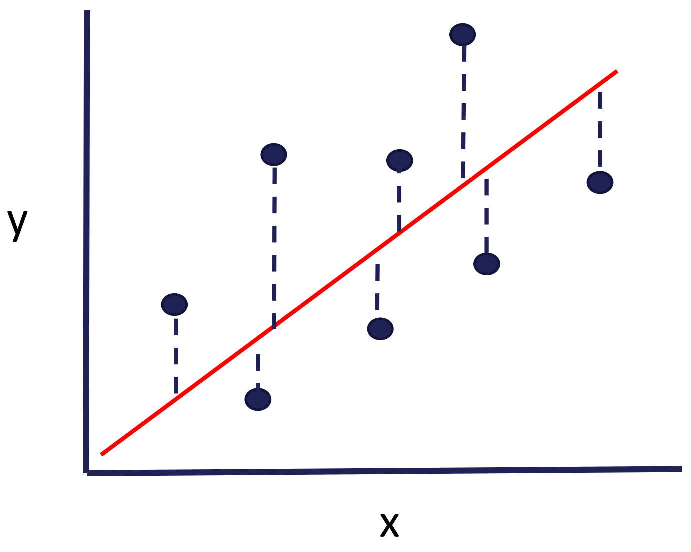


Model assessment

How do we know that the line is a good fit?

Goodness of fit





$$SST = \sum (y - y_{mean})^2$$

$$SSR = \sum (y - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni}))^2$$

$$SSM = SST - SSR$$

Goodness of fit

SST: total sum of squares ➔ total variability.

SSR: residual sum of squares ➔ variability not explained by the model.

SSM: SST – SSR : variability explained by the model.

$$R^2 = \frac{SSM}{SST}$$

Goodness of fit

R^2: fraction of variability that is explained by the model.

If SSM = SST, the model is a perfect fit.

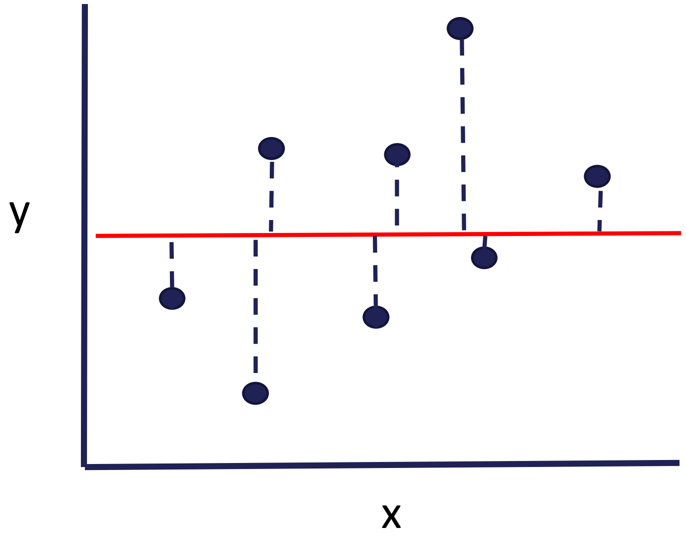
Usually, SSM < SST.

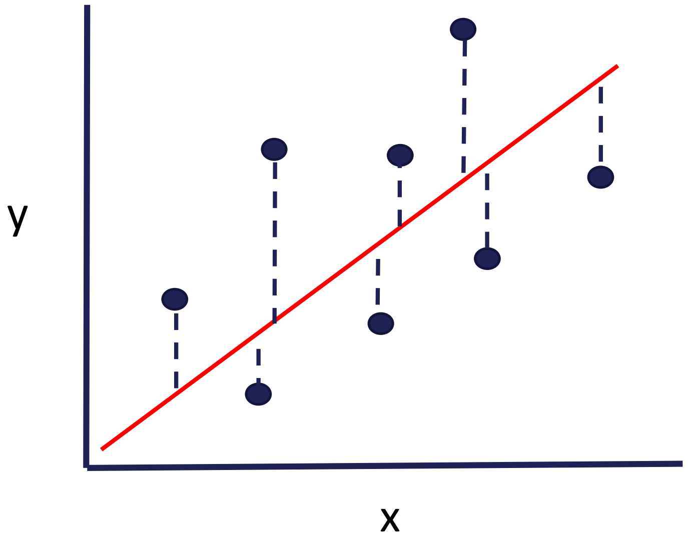
$$R^2 = \frac{SSM}{SST}$$

Model assessment

Can I say it is a good fit with confidence?

Sums of squares





$$SST = \sum (y - y_{mean})^2$$

$$SSR = \sum (y - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni}))^2$$

$$SSM = SST - SSR$$

F-test

SSM: variance explained by the model.

SSR: variance not explained by the model.

$$MS_M = \frac{SSM}{number\ of\ variables}$$

$$MS_R = \frac{SSR}{n\ obs\ -\ number\ of\ coefficients\ (\beta s)}$$

F-test

SSM: variance explained by the model.

SSR: variance not explained by the model.

$$MS_M = \frac{SSM}{number\ of\ variables}$$

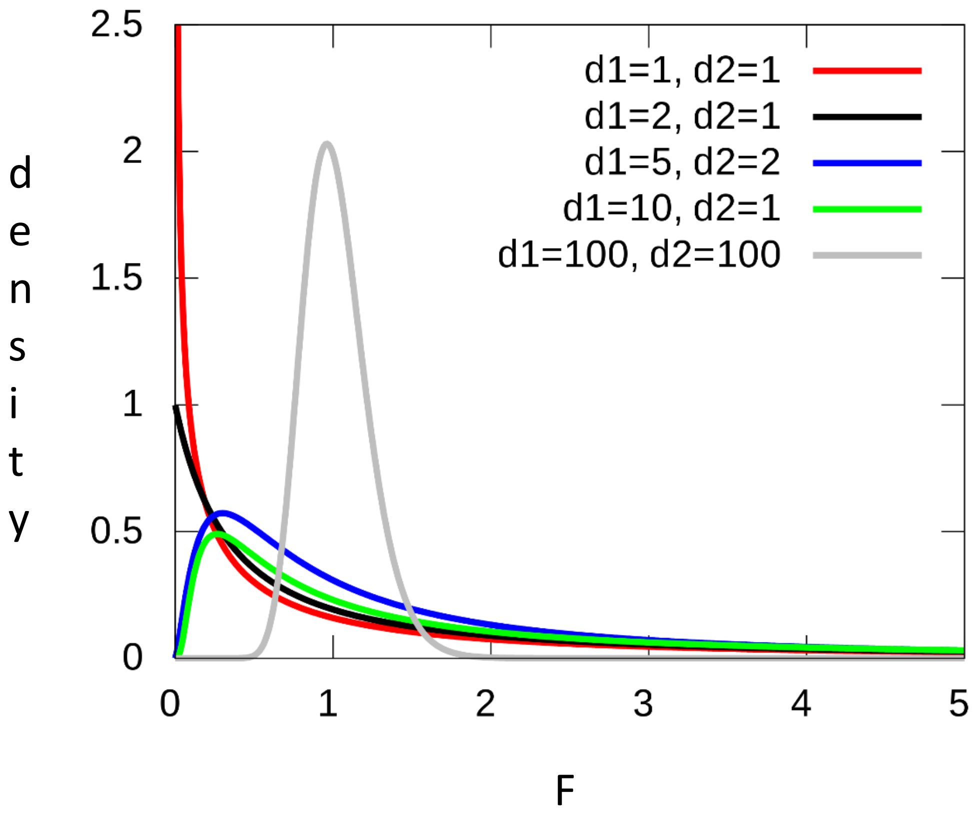
$$MS_R = \frac{SSR}{n\ obs\ -\ number\ of\ coefficients\ (\beta s)}$$

$$F = \frac{MS_M}{MS_R}$$

F-test

$$F = \frac{MS_M}{MS_R}$$

F follows a known probability distribution for situations where a model is not a good fit (null hypothesis) ➔ p-value



https://en.wikipedia.org/wiki/F-distribution#/media/File:F-distribution_pdf.svg

Summary

OLS	$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \varepsilon_i$
	The residuals are normally distributed and centred at 0.
	The R^2 indicates the fraction of variability explained by the model.
	The F-ratio indicates if the model has a significant fit.

OLS in Python

- **Sklearn**
  - Coefficients
  - R^2
- **Statsmodels**
  - Coefficients
  - R^2
  - F-ratio
  - P-values

THANK YOU

www.trainindata.com