

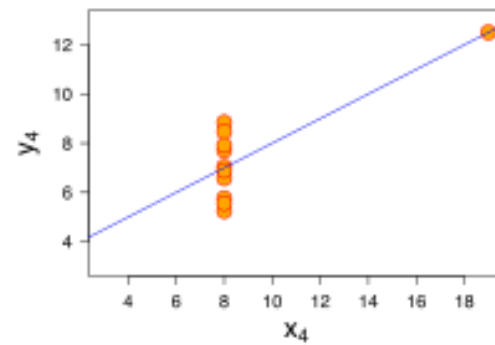
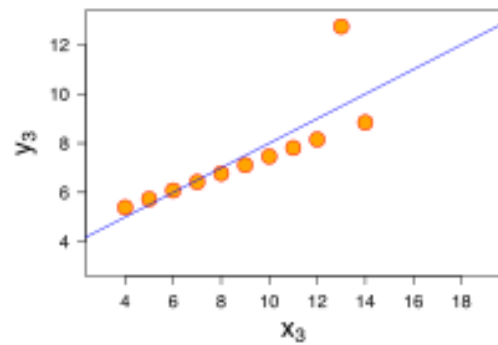
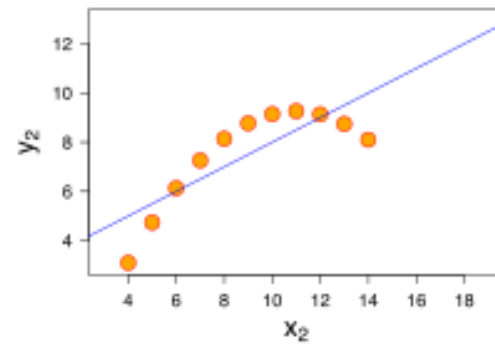
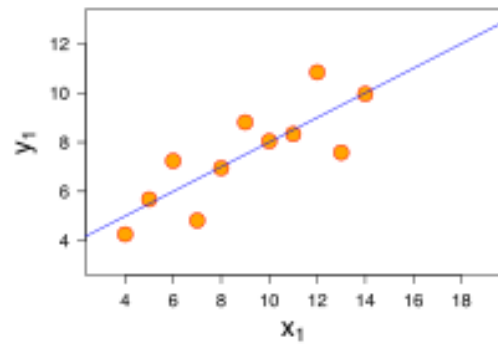
Machine Learning and Artificial Intelligence

Lab 08 – Linear and Polynomial Regression

03/05/2022

Regression analysis

- Regression analysis is a set of statistical processes for estimating the relationships between a **dependent variable** and one or more **independent variables**.
- In these problems, the dependent variable is a continuous quantity, rather than a categorical one, which is what we have seen so far in our classification problems.
- Serves two main purposes:
 1. *Prediction and forecasting*
 2. *(Pseudo)causal inference*



Components

Regression models involve the following components:

- The **unknown parameters**.
- The **independent variables**, which are observed in data and are often represented by a vector.
- The **dependent variable**, which is observed in data and is often denoted using a scalar.
- The **error terms**, which are *not* directly observed in data and are often denoted using the scalar.

Linear regression

- One of the simplest and most widely used techniques for regression.
- Given $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^n$, fits a line to the data:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

- The error is represented by the *total* sum of squared errors:

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

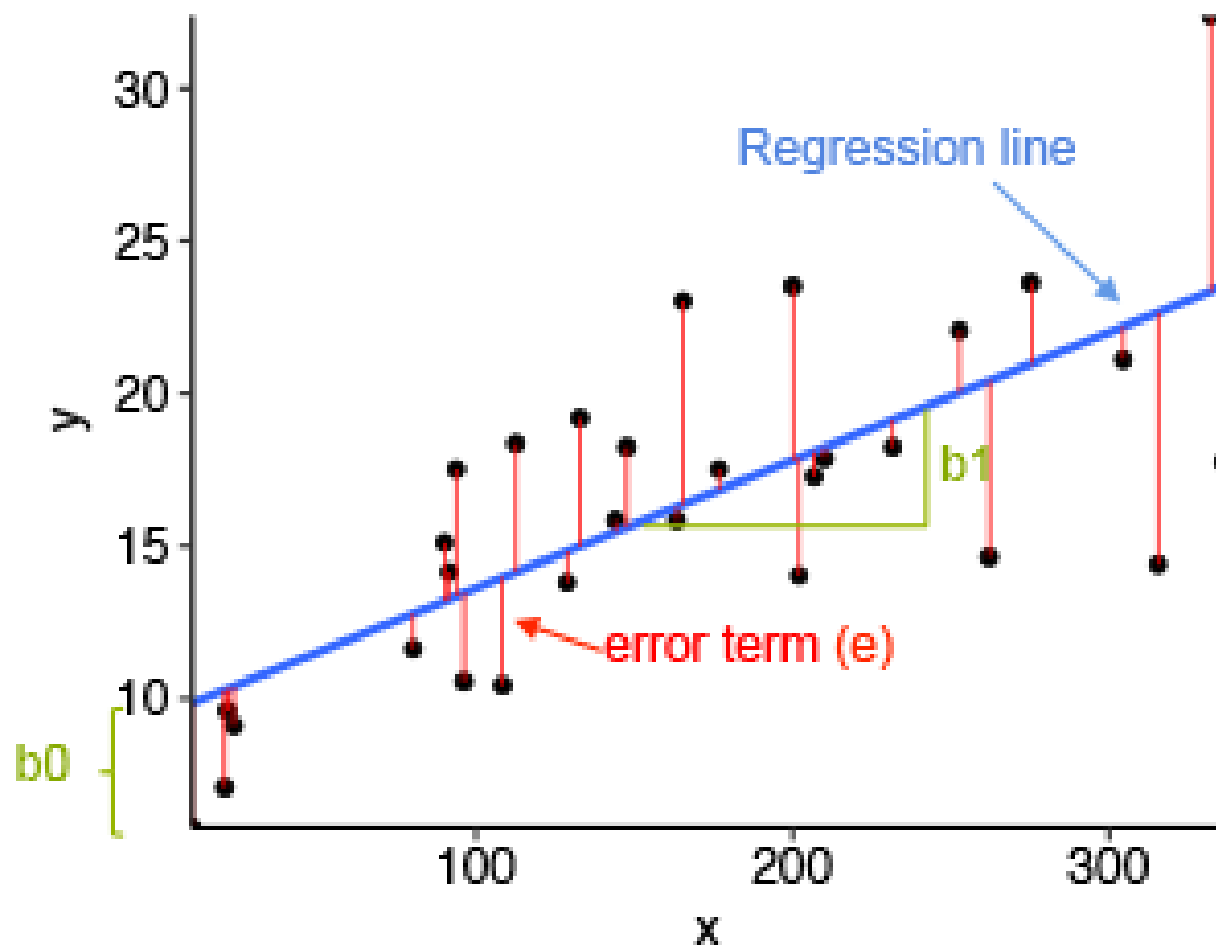
- Fit the line in such a manner that this error is minimal!

Least squares estimation

- One of the most optimal attributes of linear regression is that it **has a closed form solution**, given by ordinary *least squares*:

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

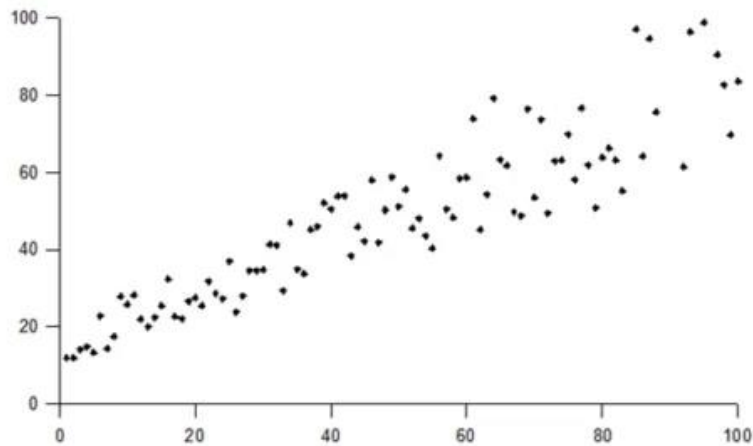


Assumptions (a whole lot)

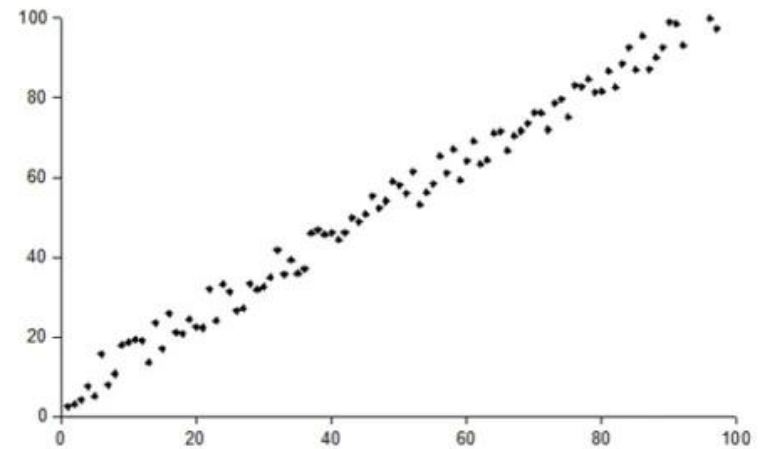
- The ordinary least squares solution makes a lot of assumptions in order to guarantee that the line is the one of best fit (least squared error):
 1. Linearity: The mean of the dependent variable is a linear combination of the independent variables (regressors).
 2. No linear dependence between regressors.
 3. Errors (Residuals) are independent and follow a normal distribution centered at 0 ($\mu = 0$)
 4. Homoskedasticity: Different values of the dependent variable have the same variance in their errors, regardless of the values of the predictor (independent) variables.

Assumptions

Heteroskedasticity



Homoskedasticity



Polynomial regression

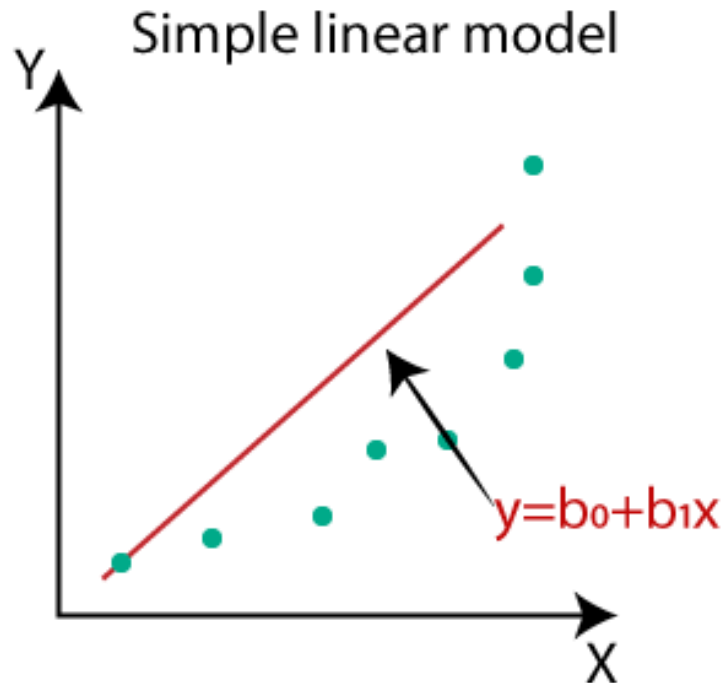
- Sometimes the relationship between the dependent and independent variables may not be linear.
- The relationship between the independent variable x and the dependent variable y can be modelled as an n th degree polynomial in x .

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_m x_i^m + \epsilon_i \quad (i = 1, 2, \dots, n)$$

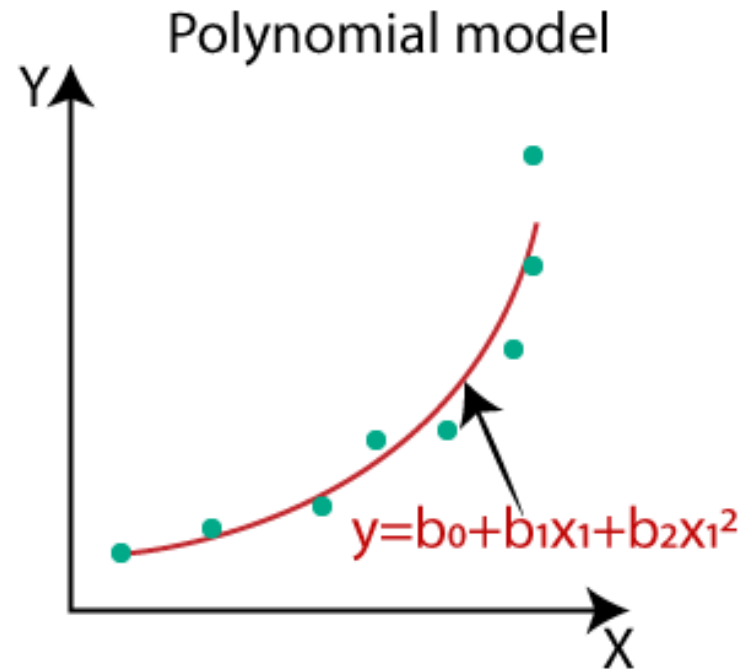
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ 1 & x_3 & x_3^2 & \dots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

- Fit the line in such a manner that SSE is minimal! (again)

Polynomial regression



OLS still holds



$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Goodness-of-fit

- In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.
- The most widely used statistic to test for a good model fit is R^2 : it is the percentage of the response variable variation that is explained by a linear model.

$$R^2 = 1 - \frac{RSS}{TSS}$$

R^2 = coefficient of determination

RSS = sum of squares of residuals

TSS = total sum of squares

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Exercises