TDAB01 Probability and Statistics

Maryna Prus
IDA, Linköping University

Lecture 11: Regression

Overview

- **Linear regression**
- **Estimation: Least squares method**
- **Multivariate regression**

Regression

- So far: Distribution of **one** random variable
- Data: $x_1, \ldots, x_n$
- Relation between **two** (or more) variables
- Data: $(x_1, y_1), \ldots, (x_n, y_n)$
- **Regression**: Type of relation between variables
- Y - **response** variable or dependent variable
- X - **explanatory** variable, independent variable, also called predictor
- Example: X - year, Y - population

Linear regression

- One explanatory variable X, assumed **known**, i.e. **not random**.
- Regression model / function:

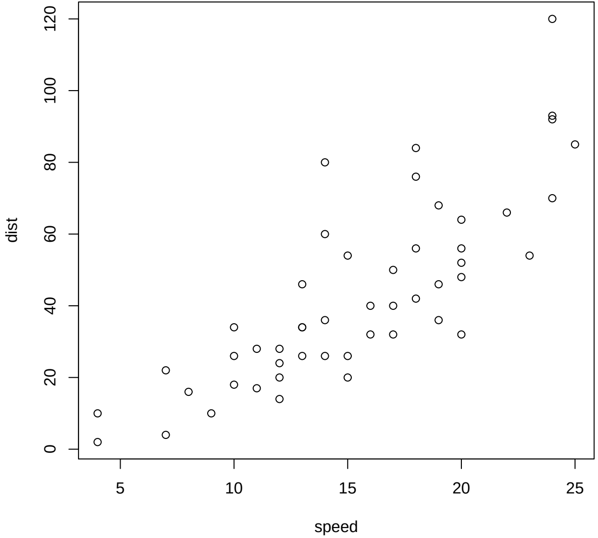
$$\hat{y}(x) = E(Y|X = x) = \beta_0 + \beta_1 x$$

- Can also be written as

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

- ε is random variable with zero mean, often $\varepsilon \sim N(0, \sigma^2)$
- ε called error term or random error

Example: Cars data



Estimation: Least squares method

- Data: $(x_1, y_1), \ldots, (x_n, y_n)$
- **Regression line** $\beta_0 + \beta_1 x$ provides the forecasts

$$\hat{y}_i = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n$$

- **Residual** at x_i:

$$e_i = y_i - \hat{y}_i$$

- **Least squares method**: Choose β_0 and β_1 that minimize sum of the squared residuals

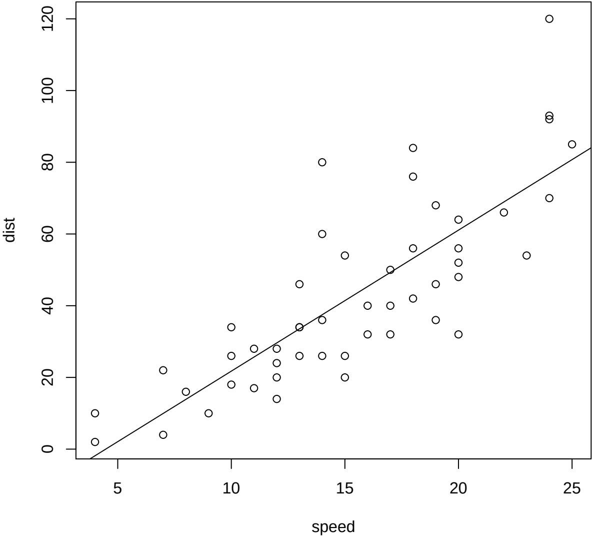
$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Solution:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Example: Cars data, cont.



R-code: See file LinReg

Estimation: ML method

- ML method: Choose values of β_0 and β_1 that maximize the probability (density) of the data. Assume independent normally distributed error terms $(\varepsilon_1, \ldots, \varepsilon_n)$
- Then $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
- Likelihood function:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} f_{Y_i}(y_i)$$

$$L(\beta_0, \beta_1) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

- Log-likelihood function:

$$\ln L(\beta_0, \beta_1) = c - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2,$$

  where $c = -n \ln\left(\sqrt{2\pi\sigma^2}\right)$ is constant, i.e. independent of β_0 and β_1
- Maximizing $\ln L(\beta_0, \beta_1)$ is the same as minimizing $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$
- **ML estimators are the same as LS estimators!**

Multivariate regression

- More than one explanatory variables
- Regression function:

$$\hat{y} = E(Y|X^{(1)} = x^{(1)}, \dots, X^{(k)})$$

  and explicitly

$$\hat{y} = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_k x^{(k)}$$

- Can also be written

$$y = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_k x^{(k)} + \varepsilon$$

- Least squares: $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ where

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1^{(1)} & \dots & x_1^{(k)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & \dots & x_n^{(k)} \end{pmatrix} \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

- **Polynomial regression**

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$$

Thank you for your attention!