Lecture 3 | 12.03.2024

Linear regression model

with one predictor variable

Simple supervised learning

- Linear regression is a simple model of supervised learning...
- The simplest regression model fits a straight line through the data
- However, the true underlying model is hardly a linear line...

Simple supervised learning

- Linear regression is a simple model of supervised learning...
- The simplest regression model fits a straight line through the data
- However, the true underlying model is hardly a linear line...

- The dependent variable Y is assumed to be continuous ($Y \in \mathbb{R}$)
- The explanatory variable can be either continuous or binary

Simple supervised learning

- Linear regression is a simple model of supervised learning...
- The simplest regression model fits a straight line through the data
- However, the true underlying model is hardly a linear line...

- The dependent variable Y is assumed to be continuous ($Y \in \mathbb{R}$)
- The explanatory variable can be either continuous or binary

- The main goal is to learn what is the underlying relationship $Y \approx f(X)$
- where, in addition, we assume that $f \in \mathcal{C} = \{f(x) = a + bx;\; a, b \in \mathbb{R}\}$

Simple (ordinary) linear regression model

- Theoretical (population model)

$$Y = a + bX + \varepsilon$$

- Random sample from the population (i.e., a joint distribution $F_{Y,X}$):

$$\mathcal{S} = \{(Y_i, X_i);\ i = 1, \dots, n\}$$

- Empirical (data) model counterpart

$$Y_i = a + bX_i + \varepsilon_i \qquad i = 1, \dots, n \in \mathbb{N}$$

Principal goals:

- Estimation of the unknown parameters $\alpha, \beta \in \mathbb{R}$
- Estimation of distributional characteristics of $Y|X$ – e.g., $E[Y|X = x]$
- Prediction of the future outcome Y_0, for an observed $X_0 = x_0$ (known)

Simple (ordinary) linear regression model

- Theoretical (population model)

$$Y = a + bX + \varepsilon$$

- Random sample from the population (i.e., a joint distribution $F_{Y,X}$):

$$\mathcal{S} = \{(Y_i, X_i);\ i = 1, \dots, n\}$$

- Empirical (data) model counterpart

$$Y_i = a + bX_i + \varepsilon_i \qquad i = 1, \dots, n \in \mathbb{N}$$

Principal goals:

- Estimation of the unknown parameters $\alpha, \beta \in \mathbb{R}$
- Estimation of distributional characteristics of $Y|X$ – e.g., $E[Y|X = x]$
- Prediction of the future outcome Y_0, for an observed $X_0 = x_0$ (known)

$\hookrightarrow$ both, the estimation and the prediction can be given in terms of some specific point (point estimate, point prediction) or in terms of some region (interval estimate, interval prediction respectively)

Linear regressing line | Examples

- Quality of the fit – the "goodness-of-fit" criterion:
  - **Mean Squared Error:** $f = \mathrm{Arg\,min}_{g \in \mathcal{C}} E[Y - g(X)]^2$ (theoretical functional)
  - **Least Squares:** $\hat{f}_N = \mathrm{Arg\,min}_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - g(X_i)]^2$ (empirical functional)

- Specific class of functions $\mathcal{C} = \{f(x);\ f(x) = a + bx; a, b \in \mathbb{R}\}$
  - linear line with the intercept parameter a and the slope parameter b
  - for $b = 0$ everything reduces to a simple mean (sample average)

- How to find $\hat{f}_N \in \mathcal{C}$ if we only know the data $\{(Y_i, X_i);\ i = 1, \dots, n\}$?
  - restricting on $\mathcal{C}$ we are looking for $\widehat{a}, \widehat{b} \in \mathbb{R}$, such that $\hat{f}_N(x) = \widehat{a} + \widehat{b}x$
  - solving a convex minimization problem

$$\min_{a,b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2 \equiv \min_{a,b \in \mathbb{R}} \mathcal{L}(a, b, \mathcal{S})$$

Least squares solution

Convex minimization problem

- minimization of a convex function
- minimization with respect to a convex set

Normal equations (score equations)

- partial derivative of $\mathcal{L}(a, b, \mathcal{S})$ with respect to the argument $a \in \mathbb{R}$
- partial derivative of $\mathcal{L}(a, b, \mathcal{S})$ with respect to the argument $b \in \mathbb{R}$

Solutions of the normal equations

- Intercept parameter estimate:

$$\widehat{a} = \overline{Y}_n - \widehat{b}\overline{X}_n$$

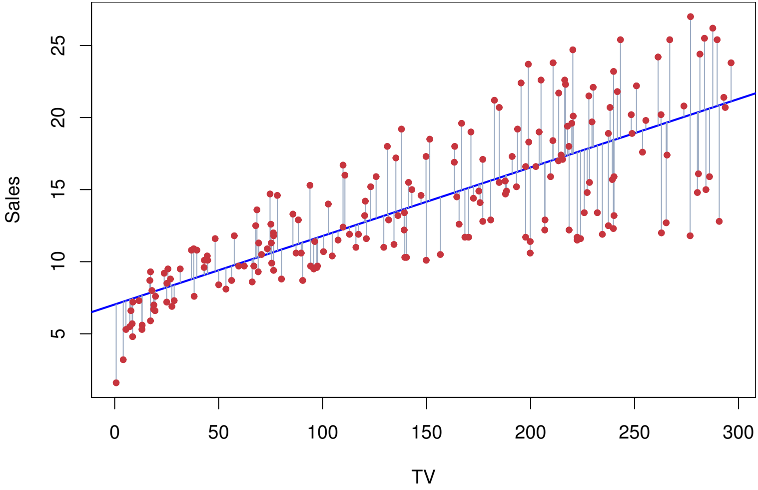
- Slope parameter estimate:

$$\widehat{b} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)(X_i - \overline{X}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

Some useful jargon

- **Fitted values**: $\widehat{Y}_i = \widehat{a} + \widehat{b}X_i$
  ("estimates" for Y values, projected Y_i values onto a line $a + bx$)
- **Residuals**: $\widehat{u}_i = Y_i - \widehat{Y}_i$
  ("estimates" for ε_i, projections of Y_i onto an orthogonal complement)
- **Residual sum of squares (RSS)**: $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$
  (the sum of squared residuals – minimization criterion)
- **Residual standard error (RSE)**: $\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$
  (residual sum of squares divided by the corresponding degrees of freedom)
- **Total sum of squares (SST)**: $\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$
  (the overall data variability with respect to Y when divided by $n - 1$)
- **Multiple R^2 value**: $1 - RSE/SST$
  (the proportion of the explained variability by considering the given model)

Regression example



Statistical properties of $\hat{a}$ and $\hat{b}$

- Assumptions: $E\varepsilon = 0$ and $Var\varepsilon = \sigma^2 < \infty$

Considering the model $Y_i = a + bX_i + \varepsilon_i$ with at least two unique values of X_i for $i = 1, \ldots, n$ and the assumptions above, we have the following:

Statistical properties of $\hat{a}$ and $\hat{b}$

- Assumptions: $E\varepsilon = 0$ and $Var\varepsilon = \sigma^2 < \infty$

Considering the model $Y_i = a + bX_i + \varepsilon_i$ with at least two unique values of X_i for $i = 1, \dots, n$ and the assumptions above, we have the following:

1. **Unbiased estimates:** $E\widehat{a} = a$ and $E\widehat{b} = b$ for all $a, b \in \mathbb{R}$
2. **Linear estimates:** $\widehat{a}$ and $\widehat{b}$ can be expressed as linear functions of Y_i
3. **Best estimates:** $\widehat{a}$ and $\widehat{b}$ are the best linear estimates in terms of the mean squared error criterion

- The result is also known as Gauss–Markov theorem – the estimates are so called **BLUE** – Best Linear Unbiased Estimates (a formal proof will be given for a multiple linear regression model with multiple predictor variables) (**BLUE – nejlepší nestranný lineárný odhad**)

Maximum likelihood estimation

- Assumption: $\varepsilon \sim N(0, \sigma^2)$

Considering the model $Y_i = a + bX_i + \varepsilon_i$ for $\varepsilon_i \sim N(0, \sigma^2)$, the maximum likelihood estimates of $a, b \in \mathbb{R}$ are given as

- Intercept and slope parameter estimates:

$$\widehat{a} = \overline{Y}_n - \widehat{b}\overline{X}_n \qquad \text{and} \qquad \widehat{b} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)(X_i - \overline{X}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

- Variance parameter estimate:

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (\widehat{a} + \widehat{b}X_i)^2)$$

and, moreover, it holds that

- $\widehat{a} \sim N\left(a, \sigma^2\left[\frac{1}{n} + \frac{\overline{X}_n}{\sum_i (X_i - \overline{X}_n)^2}\right]\right) \qquad \text{and} \qquad \widehat{b} \sim N\left(b, \frac{\sigma^2}{\sum_i (X_i - \overline{X}_n)^2}\right)$

Likelihood and log-likelihood

- density of a normal $N(\mu, \sigma^2)$ distribution

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- likelihood $L(\mu, \sigma^2, \mathcal{S})$ for the data $\mathcal{S} = \{(Y_i, X_i);\ i = 1, \ldots, n\}$

$$L(\mu, \sigma^2, \mathcal{S}) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(Y_i - (a + bX_i))^2}{2\sigma^2}\right\}\right]$$

- the corresponding log-likelihood function $\ell(\mu, \sigma^2, \mathcal{S})$

$$\ell(\mu, \sigma^2, \mathcal{S}) = (-n/2)\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(Y_i - (a + bX_i))^2}{2\sigma^2}$$

Statistical inference in a simple model

- **Confidence intervals**
  (random interval which covers unknown but non-random quantity with a pre-defined probability)
  - typically for the unknown parameters $a, b \in \mathbb{R}$
  - also for the conditional mean parameter $\mu_x = E[Y|X = x]$
  - or some reasonable linear combination, e.g. $c_1 a + c_2 b$, for $c_1, c_2 \in \mathbb{R}$

- **Hypothesis tests**
  (null vs. alternative hypothesis about the unknown but non-random parameters)
  - typically in the form $H_0 : c_1 a + c_2 b = d$ against a general (both-sided) alternative $H_A : c_1 a + c_2 b \neq d$
  - performed in terms of a test statistic which is sensitive (large) under the violation of the null hypothesis H_0

Model utilization for prediction

- **Point prediction**
  (one realization of the random variable to somehow characterize another random quantity)
  - what can be the expected outcome/realization of Y if we restrict to a sub-population given by $X = x_0$
  - typically, Y_0 (an outcome of Y when $X = x_0$) is predicted as the estimated conditional mean of Y given $X = x_0$ (i.e., $\widehat{Y}_0 = \widehat{a} + \widehat{b}x_0$)
  - other characteristics can be used of course
- **Interval prediction**
  (random interval which covers unknown but random quantity with a pre-defined probability)

Binary explanatory variable

- Until now, the explanatory variable $X \in \mathbb{R}$ was assumed to be a continuous one (taking infinitely/uncountable many values). The regression model $f(x) = a + bx$ can be, however, also considered for a binary variable X (taking only two different values)

- Let X takes value one (e.g., TRUE) and zero otherwise (e.g., FALSE)
  - For $X = 0$, the model reduces to $E[Y|X = 0] = f(0) = a$
    (*i.e., $a \in \mathbb{R}$ stands for the mean of the sub-population for which we have FALSE*)
  - For $X = 1$, the model reduces to $E[Y|X = 1] = f(1) = a + b$
    (*i.e., $a + b \in \mathbb{R}$ stands for the the mean of the sub-population for which we have TRUE*)

- Infinitely many different parametrizations can be used to encode the binary variable X – for instance, it can take two values ± 1
  (*thus, $a - b$ stands for the mean of the first and $a + b$ for the second sub-population*)
- In other words, the binary explanatory variable X reduces the ordinary linear regression model into a standard two sample problem

$$Y = a + b\mathbb{I}_{\{TRUE\}} + \varepsilon = a + b\mathbb{I}_{\{X_i=1\}} + \varepsilon = \dots$$

Summary

- simple linear regression model $Y = a + bX + \varepsilon$ (population version)
  (for a continuous response $Y \in \mathbb{R}$ and continuous or binary $X \in \mathbb{R}$)
- random sample (Y_i, X_i), $i = 1, \dots, n \Longrightarrow Y_i = a + bX_i + \varepsilon_i$ (data model)
  (realizations $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}$ drawn from a joint distribution of (Y, X))
- estimates for the unknown parameters $a, b \in \mathbb{R}$ via convex minimization
  (minimization based on the mean squared error/least squares respectively)
- under the normal model the estimation based on the maximum likelihood
  (distribution properties of the estimates $\widehat{a}$ and $\widehat{b}$ given straightforwardly)
- typical inference regarding the parameters $a, b \in \mathbb{R}$ or $E[Y|X = x]$
  (performed in terms of confidence intervals or statistical tests respectively)