Lecture 3 | 03.03.2025

Linear regression model

with one predictor/explanatory variable

Simple supervised learning

- Linear regression is a simple version of supervised learning...
  (fitting algorithms are trained on labeled data – the label is the information in Y)
- The simplest regression model fits a straight line through the data...
  (linear regression, classification, logistic regression, decision trees, SVM, NN, ...)
- Although, the true underlying model is hardly a linear line...
  (the model fitted is only a (simple) approximation of the unknown (more complex) truth)

Ordinary (simple) linear regression

- The dependent(random) variable – label Y – is assumed to be continuous
- The explanatory variable $X \in \mathbb{R}$ can be either continuous or discrete

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

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 $a, b \in \mathbb{R}$
- Estimation of distributional characteristics of $Y|X$ – e.g., $E[Y|X = x]$
- Prediction of a future outcome of Y_0, for an observed $X_0 = x_0$ (known)
- Forecasting outcomes of Y_0 given $X_0 = x_0$ (uncertainty statement)

$\hookrightarrow$ both, the estimation and the prediction can be given in terms of some specific point (e.g., point estimate, point prediction) but the forecasting is typically given in terms of some region (interval estimate, interval prediction respectively) with a given credibility guarantees

Linear regressing line

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

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

↪ the notation $\mathcal{L}(a, b, \mathcal{S})$ is used to denote a general (arbitrary) loss function $\mathcal{L}(\cdot)$, the set of unknown parameters $a, b \in \mathbb{R}$ and, also, the available dataset $\mathcal{S} = \{(Y_i, X_i);\ i = 1, \dots, n\}$. **The loss function $\mathcal{L}(\cdot)$ can be, however, defined differently.**

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 $b \in \mathbb{R}$
- both partial derivatives are set to be equal to zero and solved for $a, b \in \mathbb{R}$

Solutions of the normal equations

- Intercept parameter estimate:

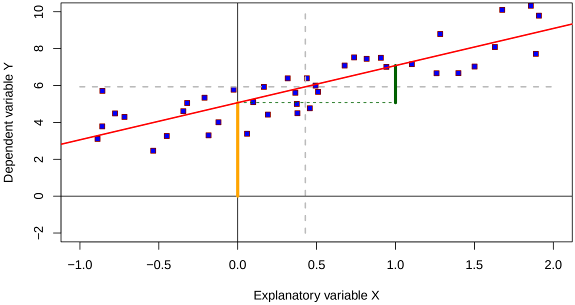
$$\widehat{a} = \overline{Y}_n - \widehat{b}\overline{X}_n$$

- Slope parameter estimate:

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

- the convexity of the optimization problem guarantees a unique solution

Least squares solution – visualization

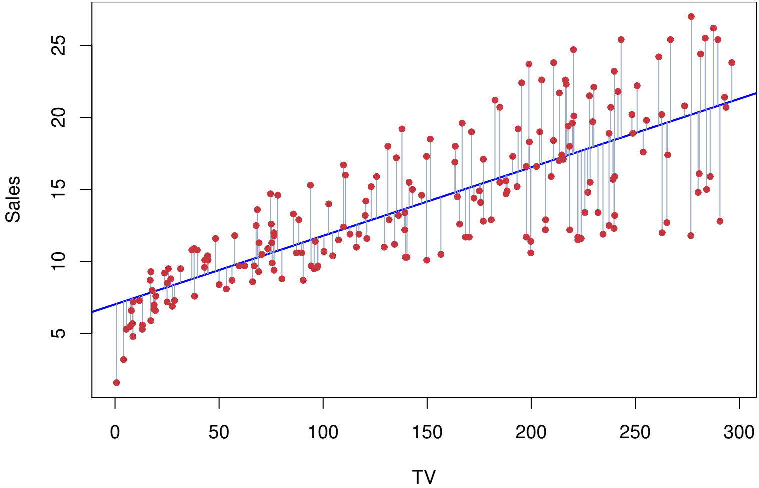


- ❑ random sample from $F_{(Y,X)}$ — the observed data $\{(Y_i, X_i);\ \ i = 1, \dots, n\}$
- ❑ estimated regression model $\hat{f}(x) = \widehat{a} + \widehat{b}x$ $(y = 5.0 + 2.0x)$
- ❑ estimated intercept parameter $\widehat{a} \in \mathbb{R}$ $(\widehat{a} = 5.048)$
- ❑ estimated slope parameter $\widehat{b} \in \mathbb{R}$ $(\widehat{b} = 2.012)$
- ❑ the “unknown” true regression model is $f(x) = 5 + 2x$

Some useful jargon

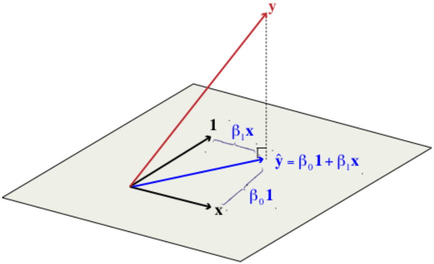
- **Fitted values**: $\widehat{Y}_i = \widehat{a} + \widehat{b}X_i$
  ($\widehat{Y}_i$ are "estimates" for Y_i values, projected Y_i values onto a line $\widehat{a} + \widehat{b}x$)
- **Residuals**: $u_i = Y_i - \widehat{Y}_i$
  (u_i are "estimates" for ε_i, projections of Y_i into orthogonal complement)
- **Residual sum of squares (RSS)**: $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$
  (the sum of squared residuals – minimization criterion – least squares)
- **Residual variance**: $\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ (RSS divided by degrees of freedom)
  (the empirical estimate of the unknown variance of the error term)
- **Residual standard error (RSE)**: $\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}$
  (estimate for the standard error – resp. square root of residual variance)
- **Total sum of squares (SST)**: $\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$
  (the overall data variability with respect to Y when "scaled" by $n-1$)
- **Multiple R^2 value**: $R^2 = 1 - RSE/SST = (SST - RSE)/SST$
  (relative proportion of the variability explained by the model – the value $(SST - RSE)$ represents the overall variability explained by the model and it is given relatively wrt the total variability in the denomitator – SST)

Regression example



Projection from 3D onto 2D – illustration

- For three data points only, $(Y_1, X_1), (Y_2, X_2)$ and (Y_3, X_3), the whole dataset can be represented in terms of two points in the three dimensional (3D) real space, $\boldsymbol{y} = (Y_1, Y_2, Y_3)^\top \in \mathbb{R}^3$ and $\boldsymbol{x} = (X_1, X_2, X_3)^\top \in \mathbb{R}^3$
- the underlying model is (still) the simple regression line $f(x) = a + bx$
- a geometric interpretation of the regression is a projection from $\mathbb{R}^3$ into $\mathbb{R}^2$



Statistical properties of the estimates $\hat{a}$ and $\widehat{b}$

- **The underlying model:** $Y = a + bX + \varepsilon$ (i.e., straight line)
- **Assumptions:** $E\varepsilon = 0$ and $Var\varepsilon = \sigma^2 < \infty$ (random error properties)

Obtaining now the random sample (Y_i, X_i) with at least two unique values of X_i for $i = 1, \dots, n$ (because the straight line is determined by two unique points) it holds, under the assumptions above, that

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 the Gauss–Markov Theorem – the estimates are so called **BLUE** – Best Linear Unbiased Estimates (a complete formal proof will be given for a multiple linear regression model with multiple predictor variables) (**BLUE – nejlepší nestranný lineárný odhad**)

Maximum likelihood estimation

- **The underlying model:** $Y = a + bX + \varepsilon$ (i.e., straight line)
- **Assumption:** $\varepsilon \sim N(0, \sigma^2)$ (distributional properties of ε)

Obtaining again the random sample (Y_i, X_i) with at least two unique values of X_i for $i = 1, \dots, n$ it holds, under the assumptions above, that

- Intercept and slope parameter estimates are equal to

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

- Variance parameter estimate:

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

- Moreover

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

Note that the estimates $\widehat{a}$ and $\widehat{b}$ are linear functions of $Y_1, \dots, Y_n$ and the distribution statements above are considered conditionally on $X_1, \dots, X_n$

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

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

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

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

$\hookrightarrow$ note the notation difference between the likelihood $L(\cdot)$, log-likelihood $\ell(\cdot)$, and a general loss function $\mathcal{L}(\cdot)$

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. $c_1 a + c_2 bx$, 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 $H_0 : c_1 a + c_2 bx = d$ against a general (both-sided) alternative $H_A : c_1 a + c_2 bx \neq d$
  - 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 only values one (e.g., TRUE) and zero (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*)

- There are infinitely many different parametrizations that 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 of the form

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

Summary

- the dependent variable Y and the explanatory variable X are assumed to follow (jointly) some (known/unknown) distribution $F_{(Y,X)}(y,x)$
- 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)
- utilization of the regression model for estimation/prediction/forecasting (the application is relatively straightforward due to intuitive parameters)