BUDAPESTI UNIVERSITY OF TECHNOLOGY AND ECONOMICS

INSTITUTE OF MATHEMATICS

FACULTY OF MATHEMATICS

Linear Regression through Origin

Author:

Dyussenov Nuraly

Supervisor:

Dr. Jozsef Mala

Associate Professor, BME Fac. of Nat. Sci.

Budapest, October 24, 2023

MŰEGYETEM 1782

Contents

List of Tables

List of Figures

1. Introduction

”Bla-bla-bla”

– XY

2. Theoretical background

2.1 Statistics Basics

Definition (Data) Let $(x_1,\dots,x_n)$, where $x_i \in S$ for $i = 1,\dots,n$. The set S is typically $\mathbb{R}$, $\mathbb{R}^d$, or it can be any abstract set. However, for our purposes, S (the sample space) will usually be $\mathbb{R}$.

Definition (Sample) In statistics, our data are often modeled by a vector $\mathbf{X} = (X_1, X_2, \dots, X_n)$ of i.i.d. (independent, identically distributed) random variables, called the sample (of which size is n), where the random variables X_i take values in $\mathbb{Z}$ or $\mathbb{R}$. The common distribution of the X_i is called the parent distribution, and we say that the sample is from that parent distribution.

Definition (Model) A statistical model is a family $\{P_\theta \mid \theta \in \Theta\}$ of distributions on the sample space. When $\Theta \subset \mathbb{R}^d$, we say that we have a parametric model, and we call Θ the parameter set (space).

Definition (p-th Quantile of Data) If $p \in (0,1)$, then a p-th quantile (or a p-th percentile) of the data $(x_1,\dots,x_n)$ is a p-th quantile of the corresponding empirical distribution function $\hat{F}_n$.

Definition (Sample mean) Let $(X_1,\dots,X_n)$ be a sample. Then the random variable

$$\bar{X} = X = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is called the sample mean.

Definition (Estimator) An estimator is a statistic (a function of the sample data) used to estimate an unknown parameter in a statistical model. An estimator for the parameter θ, denoted as $\hat{\theta}$, is any measurable function of the random variables $X_1, X_2, \dots, X_n$.

Definition (Biased) If $\hat{\theta}$ is an estimator of θ, then we can define the quantity $Bias(\hat{\theta})$ = $\mathbb{E}_\theta[\hat{\theta}] - \theta$. The estimator $\hat{\theta}$ is called unbiased if its bias is 0.

Definition (MSE of an Estimator) Let us have the model $\{P_\theta \,|\, \theta \in \Theta\}$ and let us have the sample $(X_1,\dots,X_n)$ from it. The mean square error (or the quadratic risk) of an estimator $\hat{\theta} = \hat{\theta}(X_1,\dots,X_n)$ for the parameter θ is defined by

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta((\hat{\theta}-\theta)^2)$$

when θ is the true parameter.

Steiner's identity: $\mathbb{E}((X-a)^2) = \mathrm{Var}(X) + (a - \mathbb{E}(X))^2$

Interpretation in the context of mean square error (MSE):

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}) + (\mathrm{Bias}_\theta(\hat{\theta}))^2$$

Definition (Sufficiency) Let the model be $\{P_\theta \,|\, \theta \in \Theta\}$ and $\mathbf{X} = (X_1,\dots,X_n)$ be a sample from it. The statistic T is called *sufficient* for the parameter θ (or, for the model $\{P_\theta \,|\, \theta \in \Theta\}$) if the conditional distribution $P_\theta(\mathbf{X} \in \cdot \,|\, T = t)$ does not depend on θ.

Theorem (Neyman-Fisher Factorization Theorem) If the model is $\{p(x|\theta) \,|\, \theta \in \Theta\}$ where $p(x|\theta)$ is a probability mass/density function and $\mathbf{X} = (X_1,\dots,X_n)$ is a sample from it, then the statistic T is *sufficient* for the parameter θ if and only if we can find nonnegative functions g and h such that

$$p_\mathbf{X}(x|\theta) = g(T(x),\theta)h(x).$$

Definition (Likelihood) Let $\{p(x,\theta), \theta \in \Theta\}$ be a model. If the observed value of X is x, we say that $p(x|\theta)$ is the *likelihood* of θ: $L(\theta) = p(x|\theta)$. Thus, we are considering the mass/density as a function of θ, for a fixed x. If $x = (x_1,\dots,x_n)$ is a realization of the sample $\mathbf{X} = (X_1,\dots,X_n)$, then $p(x|\theta)$ is the product of the marginals,

$$L(\theta) = p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

Theorem (Rao-Blackwell) Let $\{P_\theta \,|\, \theta \in \Theta\}$ be a model and $(X_1,\dots,X_n)$ be a sample. Let $\hat{\theta}$ be an estimator of θ with $\mathrm{Var}_\theta(\hat{\theta})$ finite for each θ. If T is a sufficient statistic for θ, then $\theta^* = \mathbb{E}_\theta(\hat{\theta}|T)$ is a statistic, and we have for all θ that

$$\mathrm{MSE}_\theta(\theta^*) \leq \mathrm{MSE}_\theta(\hat{\theta}) \quad (1)$$

and the inequality is strict unless $\hat{\theta}$ is a function of T with probability 1.

2.2 Simple Linear Regression

2.3 Simple Linear Regression with no intercept term

2.4 Comparative Analysis

3. Applications to Linear Regression through Origin

3.1 Something to add 1

3.2 Something to add 1

4. Theoretical results

4.1 A theoretical resilt

4.2 Towards some advanced topic

5. Programming simulations

6. Summary and closing words

?⟨ch:closing⟩?

Bibliography

A. Program Codes

?⟨ap:codes⟩**?**