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, November 7, 2023

MŰEGYETEM 1782

Contents

List of Tables

List of Figures

1. Introduction

”Let epsilon be negative.”

– Some funny guy

1.1 Introduction

Linear regression through the origin is a specialized form of linear regression where the regression line passes through the origin point $(0,0)$. In this variant, the relationship between the dependent variable y and the independent variable x is represented by the equation:

$$y = \beta x + \varepsilon$$

where β is the slope of the line, and ε represents the error term.

Linear regression through the origin is particularly useful in situations where it is essential to force the regression line to go through the origin, indicating a direct proportional relationship between the variables without any offset. This constraint simplifies the model and is valuable in various scientific and engineering applications.

The objective of this thesis is to delve into the nuances of linear regression through the origin. We aim to explore the mathematical foundations of this regression variant, discuss methods for estimating the parameter β, and analyze its implications on model interpretation and prediction. By conducting a comprehensive study and practical analysis, we intend to provide valuable insights into the application of linear regression through the origin in real-world contexts.

In the following chapters, we will focus on the specific aspects of linear regression through the origin. We will derive the regression equation, explore methods for estimating the slope β, and investigate techniques for evaluating the model's goodness of fit. Practical examples and numerical illustrations will be used to reinforce the concepts discussed.

Additionally, we will compare the results obtained from regression through the origin with those of traditional linear regression models. Finally, we will draw conclusions based on our findings and propose directions for further research in this domain.

This introductory chapter sets the stage for the subsequent discussions and analyses. By examining linear regression through the origin with mathematical rigor, we aim to enhance the understanding of this specialized regression technique and its practical applications in scenarios demanding an origin-centered approach.

2. Theoretical background

2.1 Statistics Basics

Definition (Data) Let $(x_1,\ldots,x_n)$, where $x_i \in S$ for $i = 1,\ldots,n$. 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 $\mathbf{X} = (X_1,X_2,\ldots,X_n)$ 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 $\{P_\theta \,|\, \theta \in \Theta\}$ 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 $(x_1,\ldots,x_n)$ is a p-th quantile of the corresponding empirical distribution function $\hat{F}_n$.

Definition (Sample mean) Let $(X_1,\ldots,X_n)$ 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 $X_1,X_2,\ldots,X_n$.

Definition (Unbiased Estimator) 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

Consider the model function:

$$y = \beta_0 + \beta_1 x,$$

which describes a line with slope β_1 and y-intercept β_0. In general, such a relationship may not hold exactly for the largely unobserved population of values of the independent and dependent variables; we call the unobserved deviations from the above equation the errors. Suppose we observe n data pairs and call them $\{(x_i, y_i), i = 1, \dots, n\}$. We can describe the underlying relationship between y_i and x_i involving this error term ε_i by:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

This relationship between the true (but unobserved) underlying parameters β_0 and β_1 and the data points is called a linear regression model.

The goal is to find estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$ for the parameters β_0 and β_1 which would provide the "best" fit in some sense for the data points. As mentioned in the introduction, in this article the "best" fit will be understood as in the least-squares approach: a line that minimizes the sum of squared residuals (see also Errors and residuals) $\hat{\varepsilon}_i$ (differences between actual and predicted values of the dependent variable y), each of which is given by, for any candidate parameter values β_0 and β_1:

$$\hat{\varepsilon}_i = y_i - \beta_0 - \beta_1 x_i.$$

In other words, $\hat{\beta}_0$ and $\hat{\beta}_1$ solve the following minimization problem:

$$\text{Find } \min_{\beta_0, \beta_1} Q(\beta_0, \beta_1), \quad \text{for } Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

We assume that the response variable is normally distributed as follows:

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \quad i = 1, \dots, n$$

such that the Y_i are independent and β_0, β_1, and σ are unknown parameters.

If $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)$, and we let $S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$, then it is easy to see that

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Likelihood function

$$L(y_1,\dots,y_n|\beta_0,\beta_1,\sigma)=\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_i))^2\right)$$

Log-likelihood function

$$l(y_1,\dots,y_n|\beta_0,\beta_1,\sigma)=c-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_i))^2$$

$$\begin{aligned}\frac{\partial l}{\partial\beta_0}&=\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_i))=0\\ \frac{\partial l}{\partial\beta_1}&=\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_i))x_i=0\\ \frac{\partial l}{\partial\sigma}&=\frac{n}{\sigma}-\frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1x_i))^2=0\end{aligned}$$

Let the solutions of the above equations be denoted as $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ for β_0, β_1, σ^2. If $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$, then...

$$\begin{aligned}\hat{\beta}_0&=\bar{y}-\hat{\beta}_1\bar{x};\\ \hat{\beta}_1&=\frac{S_{xy}}{S_{xx}};\\ \hat{\sigma}^2&=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\frac{1}{n}SSE.\end{aligned}$$

Proposition

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

Proof. The vector $\mathbf{y}-\hat{y}$ is perpendicular to $\hat{y}-\mathbf{1}\,\bar{y}$, thus the proposition is true by the Pythagorean theorem.

Proposition 1.4. The estimators $\hat{\beta}_0$, $\hat{\beta}_1$, $\frac{SSE}{n-2}$ are unbiased estimators of β_0, β_1, σ^2 respectively.

Proof. **TODO**

Proof:

1. **Unbiasedness of $\hat{\beta}_0$:**

$$\mathbb{E}(\hat{\beta}_0)=\mathbb{E}(\bar{y}-\hat{\beta}_1\bar{x})=\bar{y}-\bar{x}\mathbb{E}(\hat{\beta}_1)=\beta_0$$

2. **Unbiasedness of $\hat{\beta}_1$:**

$$\mathbb{E}(\hat{\beta}_1) = \frac{1}{S_{xx}}\mathbb{E}(S_{xy}) = \frac{1}{S_{xx}}(n\beta_1) = \beta_1$$

3. **Unbiasedness of $\frac{SSE}{n-2}$ as an estimator of σ^2:**

$$\mathbb{E}\left(\frac{SSE}{n-2}\right) = \frac{1}{n-2}(n-2)\sigma^2 = \sigma^2$$

2.3 Simple Linear Regression with no intercept term

In simple linear

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

Bibliography

A. Program Codes