BUDAPESTI UNIVERSITY OF TECHNOLOGY AND ECONOMICS

INSTITUTE OF MATHEMATICS

BACHELOR THESIS

Linear Regression through Origin

Author:
Dyussenov Nuraly

Supervisor:
Dr. Jozsef Mala
Associate Professor, BME Fac. of Nat. Sci.

Budapest, December 6, 2023

Contents

List of Figures

1. Introduction

1.1 Background

In the world of regression analysis, choosing the right model is a constant challenge, balancing simplicity and accuracy. This thesis focuses on a specific aspect—linear regression through the origin (RTO) —examining its statistical properties when dealing with just one explanatory variable. Our goal is to identify situations where this approach might be more suitable than the commonly used simple linear regression. Through this study, we aim to shed light on the conditions that make regression through the origin a preferable choice, offering insights that bridge mathematical rigor with real-world applicability. Join us on this journey as we navigate the complexities of statistical modeling, striving to understand when and why regression through the origin might outperform its more conventional counterpart.

1.2 Literature review

1.3 Research question

1.4 Outline

2. Preliminaries

2.1 Statistics Basics

In this part we will cover most of the statistics that we are going to need later and also some of the notations that we are going to use.

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 any abstract set. However, for our purposes, S (the sample space) is usually taken as $\mathbb{R}$.

Definition (Sample): In statistics, our data are often represented by a vector $\mathbf{X} = (X_1, X_2, \dots, X_n)$ of i.i.d. (independent, identically distributed) random variables, denoted as the sample (of size n). The random variables X_i take values in $\mathbb{Z}$ or $\mathbb{R}$. The common distribution of the X_i is referred to as the parent distribution, and the sample is said to be drawn from that parent distribution.

Definition (Model): A statistical model is a family $\{P_\theta \mid \theta \in \Theta\}$ of distributions on the sample space. In the case where $\Theta \subset \mathbb{R}^d$, it is a parametric model, with Θ being the parameter set (space).

Definition (Statistic): A statistic $T = T(x_1,...,x_n)$ is any function of the sample data, often used to summarize or draw inferences about the underlying population.

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 referred to as 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. Denoted as $\hat{\theta}$, an estimator for the parameter θ is any measurable function of the random variables $X_1, X_2, \dots, X_n$.

Definition (Unbiased Estimator): If $\hat{\theta}$ is an estimator of θ, we can define the

quantity $Bias(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta$. The estimator $\hat{\theta}$ is termed unbiased if its bias is 0.

Definition (MSE of an Estimator): Consider the model $\{P_\theta \mid \theta \in \Theta\}$ and the sample $(X_1, \ldots, X_n)$ from it. The mean square error (or quadratic risk) of an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ for the parameter θ is defined as

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

when θ is the true parameter.

Definition (Likelihood): We say that $L(\theta)$ is a likelihood of an estimator θ if $L(\theta) = p(x|\theta)$, where x is a realization $(x_1, ..., x_n)$ of a sample X. Thus, likelihood can be obtained by taking the product of marginal probability mass functions for fixed x, and expressing the result as a function of θ.

$$p(x|\theta) = \Pi_{i=1}^{n} p(x_i|\theta) \tag{2.1}$$

We will call estimator $\hat{\theta}$ a Maximum Likelihood Estimator of θ if it maximizes 2.1. One common approach to find this estimator is to take partial derivative with respect to θ of log-likelihood function $l(\theta) = \log L(\theta)$. Since logarithm is a monotonically increasing function, the maximum of $l(\theta)$ will be the same as maximum of $L(\theta)$

3. Simple Linear Regression

Before we start delving into RTO, it's best to get familiar with a more general case - Simple Linear Regression.

In Simple Linear Regression, we are given a random sample of data points $(x_1, y_1), ..., (x_n, y_n)$ from a population, and our goal is to find a linear function that describes the relationship between x (often called, explanatory variable, regressor) and y (often called dependent variable, regressand) as good as possible:

$$y_i = \beta_1 x_i + \beta_0 \quad (i = 1, 2, ..., n) \tag{3.1}$$

or in matrix notation

$$\mathbf{Y} = \beta \mathbf{X}$$

where $\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}$

Since, sample is random, the equation (3.1) is not true in general, so we take into account error term $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$, where $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are i.i.d random variables:

$$\mathbf{Y} = \beta \mathbf{X} + \varepsilon \tag{3.2}$$

The objective of simple linear regression is, under some assumption [1, p.4-12], to estimate the parameters β_0 and β_1, so that they will provide best fit. The assumptions are important since they serve as a foundation for later theory, so we will take some to analyze them:

Assumption 3.0.1 (linearity). $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$

Linearity means that the underlying relationship between regressand and regressor is linear.

Assumption 3.0.2 (**strict exogeneity**). $\mathbb{E}(\varepsilon_i|\mathbf{X}) = 0 \quad (i = 1,2,...,n)$

From this assumption it follows that:

$$\mathbb{E}(\varepsilon_i) = 0 \quad (i = 1,2,...,n) \tag{3.3}$$

Proof. From law of total expectation it also follows that $\mathbb{E}(\varepsilon_i) = \mathbb{E}(\mathbb{E}(\varepsilon_i|\mathbf{X})) = 0$ □

The cross moments are orthogonal to all observations:

$$\mathbb{E}(x_i\varepsilon_j) = 0 \quad (i,j = 1,2,...,n) \tag{3.4}$$

Proof.

$$\begin{aligned} \mathbb{E}(x_i\varepsilon_j) &= \mathbb{E}(\mathbb{E}(x_i\varepsilon_j|x_j)) && \text{(by law of total expectation)} \\ &= \mathbb{E}(x_i\mathbb{E}(\varepsilon_j|x_j)) && \text{(by linearity of expectation)} \\ &= \mathbb{E}(x_i 0) = 0 \end{aligned}$$

□

Because the mean of error terms is zero, the orthogonality condition, implies that error terms have zero correlation with observations.

Assumption 3.0.3 (**no multicollinearity**). *Data vector* $\mathbf{X}$ *has no duplicates with probability 1.*

The more general variant of this assumption, in the case when we have classical linear regression with multiple regressors, the multicollinearity would mean that the rank of the $n \times K$ data matrix $\mathbf{X}$ is exactly K with probability 1. In other words, columns of the data matrix have to be linearly independent with probability 1.

Assumption 3.0.4 (**homoskedasticity**). $\mathbb{E}(\varepsilon_i^2|\mathbf{X}) = \sigma^2 > 0 \quad (i = 1,2,...,n)$

Assumption 3.0.5 (no correlation between observatoins)**.**

$$\mathbb{E}(\varepsilon_i\varepsilon_j) = 0 \quad (i,j = 1,2,...,n; i \neq j)$$

This is equivalent to saying that covariance of error terms is zero, since means are zero.

Remark: Although in the assumptions we treat regressors as random variables, in reality, $x_1, ..., x_n$ are realizations of random variables, so it is generally accepted to tream them as fixed

ask professor

For the purposes of this paper, we will also assume that error terms in eq. 3.2 are normally distributed with mean 0 and variance σ^2.

To compute likelihood function, we can notice that y_i is a normally distributed random variable with mean $\mathbb{E}[y_i] = \beta_1 x_i + \beta_0$ and variance Var $y = \sigma^2$. Then, likelihood L can be computed as a product of marginal distributions:

$$L(y_1, \dots, y_n | \beta_0, \beta_1, \sigma) = \prod \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y_i - (\beta_1 x_i + \beta_0)^2)) \tag{3.5}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2\right) \tag{3.6}$$

Log-likelihood function

$$l(y_1, \dots, y_n | \beta_0, \beta_1, \sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 \tag{3.7}$$

By taking partial derivatives with respect to our parameters, we can find that estimates of β_0, β_1 and σ that maximize the likelihood:

$$\frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i)) = 0 \tag{3.8}$$

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))x_i = 0 \tag{3.9}$$

$$\frac{\partial l}{\partial \sigma} = \frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 = 0 \tag{3.10}$$

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 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, then...

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}; \tag{3.11}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}; \tag{3.12}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}SSE. \tag{3.13}$$

So $\hat{\beta}_0, \hat{\beta}_1$ are Maximum Likelikelihood Estimators of the model.

Proposition 3.0.6. *Finding values of β_0, β_1 that minimize MSE is same as finding MLE of β_0, β_1*

Proof. From equations 3.8 and 3.9 we can see that $y - \hat{y}$ is perpendicular to both $\mathbf{1}$ and $\mathbf{x}$, and we know that $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$, i.e. $\hat{y}$ lies in span$\{\mathbf{1}, \mathbf{x}\}$. Notice that $y - \hat{y} = \varepsilon$, so since $\hat{y}$ is perpendicular to ε, MLE from 3.11 and 3.12 are indedd MSE estimators too. □

include figure

Remark: Equations 3.8 and 3.9 are called normal equations [2] (from the fact that $\mathbf{y} - \hat{\mathbf{y}}$ are orthogonal to $\mathbf{1}$ and $\mathbf{x}$) and in matrix form it could be written as:

$$\mathbf{X^T X b} = \mathbf{X^T Y} \tag{3.14}$$

Where $\mathbf{b}$ is 2×1 vector of the regression MLE coefficients $(\hat{\beta}_0, \hat{\beta}_1)^T$

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.

Alternatively, it is enough to show that

$$\sum(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$$

,

since then:

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

From 3.8 we know that $\sum(y_i-\hat{y}_i)=0$. From 3.9 we know that $\sum(y_i-\hat{y}_i)x_i=0$, $\hat{y}_i=\beta_0+\beta_1x_i\Rightarrow x_i=\frac{1}{\beta_1}(\hat{y}_i-\beta_0)\Rightarrow\sum\hat{y}_i(y_i-\hat{y}_i)=0$

Finally,

$$\sum(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)=\sum\hat{y}_i(y_i-\hat{y}_i)-\bar{y}\sum(y_i-\hat{y}_i)=0$$

□

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:

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

$$\mathbb{E}[\hat{\beta}_1]=\mathbb{E}[\frac{S_{xy}}{S_{xx}}]=\mathbb{E}[\frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}]$$
$$=\mathbb{E}[\frac{\Sigma(x_i-\bar{x})y_i}{\Sigma(x_i-\bar{x})^2}]=\frac{\Sigma(x_i-\bar{x})\mathbb{E}[y_i]}{\Sigma(x_i-\bar{x})^2}$$
$$=\frac{\Sigma(x_i-\bar{x})(\beta_0+\beta_1x_i)}{\Sigma(x_i-\bar{x})^2}=\frac{\Sigma(x_i\beta_0-\bar{x}\beta_0+\beta_1x_i^2-\beta_1x_i\bar{x})}{\Sigma x_i^2-n\bar{x}^2}$$
$$=\frac{\cancel{n\bar{x}\beta_0}-\cancel{n\bar{x}\beta_0}+\Sigma\beta_1x_i^2-n\beta_1\bar{x}^2}{\Sigma x_i^2-n\bar{x}^2}=\frac{(\Sigma x_i^2-n\bar{x}^2)\beta_1}{\Sigma x_i^2-n\bar{x}^2}=\beta_1$$

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

$$\mathbb{E}(\hat{\beta}_0)=\mathbb{E}(\bar{y}-\hat{\beta}_1\bar{x})=\bar{y}-\bar{x}\mathbb{E}(\hat{\beta}_1)=\frac{1}{n}\mathbb{E}[\Sigma y_i]-\beta_1\bar{x}=$$

$$= \frac{1}{n}\mathbb{E}[\Sigma(\beta_0 + \beta_1 x_i)] - \beta_1 \bar{x} = \frac{1}{n} n\beta_0 + \frac{1}{n} n\beta_1 \bar{x} - \bar{x}\beta_1 = \beta_0$$

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

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

Proposition 3.0.7. $\mathbb{V}ar[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$

Proof. Assume, $Y_i \sim N(0, \sigma^2)$

$$\mathbb{V}ar[\hat{\beta}_1] = \mathbb{V}ar(\frac{1}{S_{xx}}\sum(x_i - \bar{x})Y_i) = \frac{1}{S_{xx}^2}\sum \mathbb{V}arY_i = \frac{\sigma^2}{S_{xx}}$$

□

Proposition 3.0.8. $\text{Var}[\hat{\beta}_0] = \frac{\sigma^2 S_{xx}^o}{nS_{xx}}$

Proof. add □

4. Linear Regression Regression with no intercept term

4.1 Simple Linear Regression with no intercept term

In certain statistical applications, the conventional assumption of a non-zero intercept term (β_0) in a simple linear regression model may not align with the nature of the data. For example, in economics the cost of production be assumed to be zero, when there is no production, or in physics, when we are describing the relationship between force and the displacement, forced is assumede to be zero, when there is no displacement.

In linear regression through the origin, regression equation takes the form:

$$\mathbf{y} = \beta_1^o \mathbf{x} + \varepsilon, \tag{4.1}$$

where $\mathbf{x}^T = (x_1, x_2, ..., x_n)$

Likelihood function L is:

$$\begin{aligned} L(y_1, ..., y_n | \beta_1^o, \sigma) &= \prod \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2 (y_i - \beta_1^o x_i)^2}) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2} \sum (y_i - \beta_1^o x_i)^2) \end{aligned}$$

Log-likelihood l is:

$$l(y_1, ..., y_n | \beta_1^o, \sigma) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - \beta_1^o x_i)^2$$

Thus we can compute the maximum likelihood estimator of β_1^o by taking partial derivative of log-likelihood with respect to β_1^o:

$$\frac{\partial l}{\partial \beta_1^o} = -\frac{1}{2\sigma^2}\sum 2(y_i - \beta_1^o x_i)(-x_i) = 0$$
$$\frac{1}{\sigma^2}\sum (y_i x_i - \beta_1^o x_i^2) = 0$$
$$\sum x_i y_i = \sum x_i^2 \beta_1^o$$
$$\hat{\beta_1^o} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Proposition 4.1.1. *$\hat{\beta}_1^o$ is unbiased*

Proof.

$$\mathbb{E}[\hat{\beta_1^o}] = \mathbb{E}[\frac{\sum x_i y_i}{\sum x_i^2}] = \sum \frac{1}{x_i^2}\mathbb{E}[\sum x_i y_i] =$$
$$= \frac{\sum x_i \mathbb{E}[y_i]}{\sum x_i^2} = \frac{\sum x_i^2 \beta_1^o}{\sum x_i^2} = \beta_1^o$$

□

Proposition 4.1.2. $\mathbb{V}ar[\hat{\beta}_1^0] = \frac{\sigma^2}{\sum x_i^2}$

Proof.

$$\mathbb{V}ar[\hat{\beta}_1^0] = \mathbb{V}ar[\frac{\sum x_i y_i}{\sum x_i^2}] = \frac{1}{(\sum x_i^2)^2}\sum x_i^2 \mathbb{V}ar[Y_i] = \frac{\sigma^2}{\sum x_i^2}$$

□

Proposition 4.1.3.

$$\mathbb{V}ar[\hat{\beta}_1^0] < \mathbb{V}ar[\hat{\beta}_1]$$

Proof.

$$\sum x_i^2 > \sum (x_i - \bar{x})^2$$
$$\frac{1}{\sum x_i^2} < \frac{1}{\sum (x_i - \bar{x})^2}$$
$$\frac{\sigma^2}{\sum x_i^2} < \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$
$$\mathbb{V}ar[\hat{\beta}_1^0] < \mathbb{V}ar[\hat{\beta}_1]$$

□

This might suggest that $\hat{\beta}_1^0$ might be more accurate estimator than $\hat{\beta}_1$ for the slope term. This gives us some motivation to compare the two estimators more closely.

It would be convenient for us to find confidence interval for $\beta_1^0 - \beta_1$, since if 0 lies in the CI, then we can statistically infer that two estimators are very close.

Proposition 4.1.4. *The difference of the two estimators is normally distributed as follows:*

rm

$$\hat{\beta}_1^0 - \hat{\beta}_1 \sim N(\beta_1^0 - \beta_1, \sigma^2(\frac{1}{S_{xx}} - \frac{1}{S_{xx}^0}))$$

Before proving proposition 4.1.4, let's first understand some properties of $\hat{\beta}_1^0 - \hat{\beta}_1$:

Proposition 4.1.5. $\text{Cov}(\hat{\beta}_1^0, \hat{\beta}_1^0 - \hat{\beta}_1) = 0$

Proof.

$$\begin{aligned}
\text{Cov}(\hat{\beta}_1^0, \hat{\beta}_1^0 - \hat{\beta}_1) &= \text{Cov}(\frac{\sum xy}{\sum x^2}, \frac{\sum xy}{\sum x^2} - \frac{\sum (x-\bar{x})y}{\sum (x-\bar{x})^2}) \\
&= \frac{\sum x^2}{(\sum x^2)^2}\text{Var } y - \frac{\sum x(x-\bar{x})}{\sum x^2 \sum (x-\bar{x})^2}\text{Var } y \\
&= \sigma^2(\frac{1}{\sum x^2} - \frac{\sum x^2 - \bar{x}\sum x}{\sum x^2 \sum (x^2 - 2x\bar{x} + \bar{x}^2)}) \\
&= \sigma^2(\frac{1}{\sum x^2} - \frac{\sum x^2 - n\bar{x}^2}{\sum x^2 (\sum x^2 - 2n\bar{x}^2 + n\bar{x}^2)}) \\
&= \sigma^2(\frac{1}{\sum x^2} - \frac{\sum x^2 - n\bar{x}^2}{\sum x^2 (\sum x^2 - n\bar{x}^2)}) \\
&= \sigma^2(\frac{1}{\sum x^2} - \frac{1}{\sum x^2}) = 0
\end{aligned}$$

□

Proposition 4.1.6. $\mathbb{E}[\hat{\beta}_1^0 - \hat{\beta}_1] = \beta_1^0 - \beta_1$

not true in general

Proof. By linearity of expected value, $\mathbb{E}[\hat{\beta}_1^0 - \hat{\beta}_1] = \mathbb{E}[\hat{\beta}_1^0] - \mathbb{E}[\hat{\beta}_1] = \beta_1^0 - \beta_1$ □

Now we are ready to prove proposition 4.1.4

Proof of prop. 4.1.4. $\hat{\beta}_1^0 - \hat{\beta}_1$ is a linear combination of mutually independent normally distributed r.v.-s $\Rightarrow$ it is normally distributed. you can show that

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1^0 - \hat{\beta}_1) &= \mathrm{Var}(\hat{\beta}_1^0) + \mathrm{Var}(\hat{\beta}_1) - 2\mathrm{Cov}(\hat{\beta}_1^0, \hat{\beta}_1) = \frac{\sigma^2}{S_{xx}^0} + \frac{\sigma^2}{S_{xx}} - 2\mathrm{Cov}(\hat{\beta}_1^0, \hat{\beta}_1^0 - \hat{\beta}_1) - 2\mathrm{Cov}(\hat{\beta}_1^0, \hat{\beta}_1^0) \\ &= \frac{\sigma^2}{S_{xx}^0} + \frac{\sigma^2}{S_{xx}} - 0 - 2\mathrm{Var}\,\hat{\beta}_1^0 = \frac{\sigma^2}{S_{xx}} - \frac{\sigma^2}{S_{xx}^0}\end{aligned}$$

□

Using proposition 4.1.4, we can now construct confidence interval for $\beta_1^0 - \beta_1$:

$$\hat{\beta}_1^0 - \hat{\beta}_1 \sim N(\beta_1^0 - \beta_1, \sigma^2(\frac{1}{S_{xx}} - \frac{1}{S_{xx}^0}))$$

$$Z := \frac{\hat{\beta}_1^0 - \hat{\beta}_1 - \mathbb{E}[\hat{\beta}_1^0 - \hat{\beta}_1]}{\mathrm{Var}(\hat{\beta}_1^0 - \hat{\beta}_1)} = \frac{\hat{\beta}_1^0 - \hat{\beta}_1 - (\beta_1^0 - \beta_1)}{\sigma^2(\frac{1}{S_{xx}} - \frac{1}{S_{xx}^0})})$$

$$\beta_1^0 - \beta_1 = \hat{\beta}_1^0 - \hat{\beta}_1 - \sigma^2 Z(\frac{1}{S_{xx}} - \frac{1}{S_{xx}^0})$$

CI is:

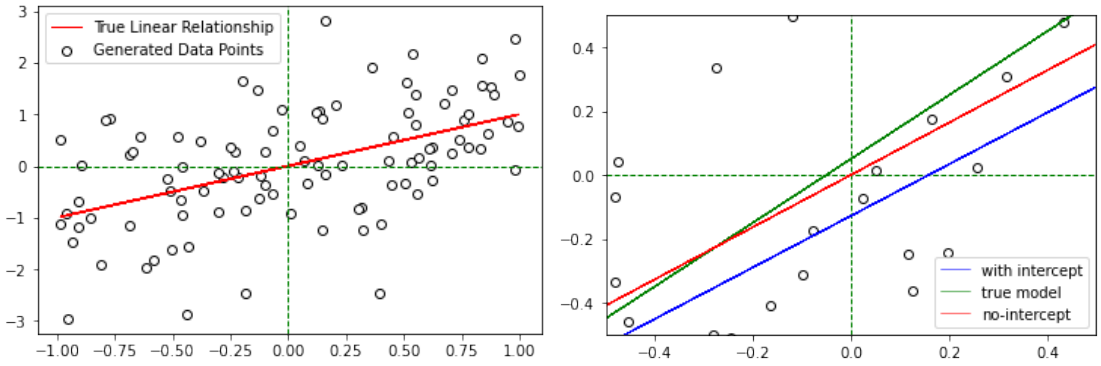
$$\hat{\beta}_1^0 - \hat{\beta}_1 - \sigma^2 Z_\alpha(\frac{1}{S_{xx}} - \frac{1}{S_{xx}^0}) \leq \beta_1^0 - \beta_1 \leq \hat{\beta}_1^0 - \hat{\beta}_1 + \sigma^2 Z_\alpha(\frac{1}{S_{xx}} - \frac{1}{S_{xx}^0}) \tag{4.2}$$

4.1.1 Motivational example

Even though RTO might be worse than performing full linear regression model in terms of SSE (since full model always gives unbiased estimators, they are also MLE estimators, so SSE is minimized), there are other statistical parameters that are better in RTO when intercept term is small enough.

Given that the true model is known, $y_i = \beta_1 x_i + \beta_0$, one might simulate the data points by adding normally-distributed error terms:

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



(a) Plot of simulated points

(b) Plot of three models

Figure 4.1: Linear Regression model on simultated datapoints

One might be interested now in comparing sum of squared deviations (SSD) of fitted points $\sum(y_i - \hat{y}_i)^2$.

Let's start by generating a sample of 100 points, and assume that parameters are known as $\beta = 1, \alpha = 0.05$ (see fig. 4.1a):

Now, let's fit two models: no-intercept model, and full model. For that purposes we will use LinearRegression function from `sklearn.linear_model` package. On fig. 4.1b you can see the two models as well as the true model plotted up close.

Now let's compare two statistics for these two models: BIC (Bayesian Information Criterion) and SSD. (SSD and BIC of no-intercept model will be denoted as SSD_{no} and BIC_{no})

The results are as follows:

$$BIC_{no} = 451.95123880008833$$
$$BIC = 457.6345669168381$$
$$SSD_{no} = 0.45764140442744106$$
$$SSD = 0.46305539852212607$$

It is clearly seen that no-intercept model provides better fit in terms of these parameters.

Moreover, when we perform 1000 Monte Carlo simulations for 1000 different in-

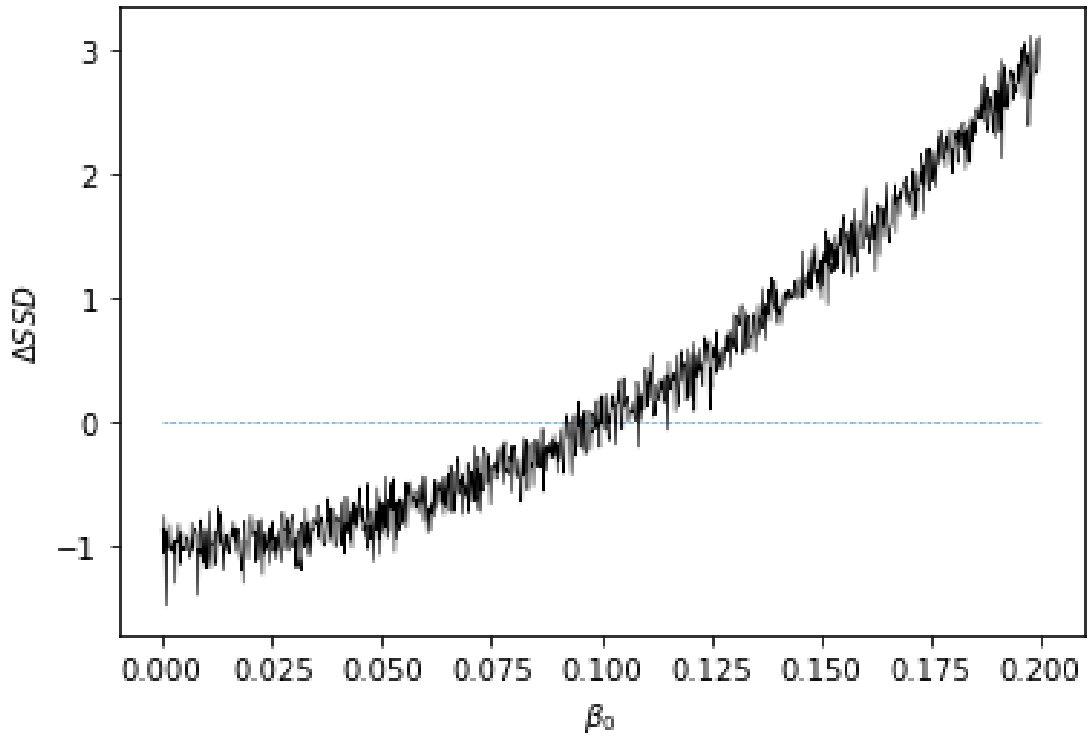


Figure 4.2: Plot of the difference of SSD to the values of intercept

tercept values β_0, ranging from 0 to 0.2, and $\beta_1 = 1, \sigma^2 = 1$, and we plot the graph of $\Delta SSD := SSD_{no} - SSD$ to β_0, we can see that ΔSSD takes negative values in the region that is bellow blue dashed line in fig. 4.2. This might suggest that no-intercept model is better in terms of SSD compared to the full model. However let's give this assumption a mathematical justification.

4.1.2 Expected value of SSD and SSD_{no}

Our goal here is to calculate $\mathbb{E}[SSD]$ and $\mathbb{E}[SSD_{no}]$ to justify the region below zero in fig. 4.2. In this section we are going to prove the following two propositions:

Proposition 4.1.7. $\mathbb{E}[SSD_{no}] = n\beta_0^2 - \beta_0^2 n^2 \frac{\bar{x}^2}{\sum x^2} + \sigma^2$

Proposition 4.1.8. $\mathbb{E}[SSD] = \frac{2\sigma^2}{S_{xx}}(S_{xx}^o - n\bar{x}^2)$

Assuming $\beta_0 > 0$. In order to prove proposition 4.1.7, we have to take into account that $\beta_0 \neq 0$, so the estimator β_1^o is now biased.

Proposition 4.1.9. $\mathbb{E}[\hat{\beta}_1^o] = \frac{\sum \beta_0 x}{\sum x^2} + \beta_1$

Proof.

$$\begin{aligned}\mathbb{E}[\hat{\beta}_1^o] &= \frac{\mathbb{E}\sum xy}{\sum x^2} = \frac{\sum x\mathbb{E}[y]}{\sum x^2} = \frac{\sum x(\beta_1 x + \beta_0)}{\sum x^2} \\ &= \frac{\sum \beta_0 x + \sum \beta_1 x^2}{\sum x^2} = \frac{\sum \beta_0 x}{\sum x^2} + \beta_1\end{aligned}$$

□

Proposition 4.1.10. $\mathbb{E}[\hat{\beta}_1^{o2}] = \frac{\sigma^2}{\sum x^2} + (\frac{\sum \beta_0 x}{\sum x^2} + \beta_1)^2$

Proof. Comes from the fact that $\mathbb{E}[\hat{\beta}_1^{o2}] = \mathrm{Var}[\hat{\beta}_1^o] + \mathbb{E}[\hat{\beta}_1^o]^2$ and then one can apply prop. 4.1.9 and prop. 4.1.2 □

Proof of proposition 4.1.7. Here, we assume that $\hat{y}$ are values fitted by no-intercept model:

$$\begin{aligned}\mathbb{E}[SSD_{no}] &= \sum(\dot{y} - \hat{y})^2 && \text{(df } SSD_{no}\text{)} \\ &= \sum \dot{y}^2 - \sum 2\dot{y}\mathbb{E}[\hat{y}] + \sum \mathbb{E}[\hat{y}^2] && \text{(linearity)} \\ &= \sum \dot{y}^2 - \sum 2\dot{y}x(\frac{\sum \beta_0 x}{\sum x^2} + \beta_1) + \sum x^2(\frac{\sigma^2}{\sum x^2} + (\frac{\sum \beta_0 x}{\sum x^2})^2 + \frac{2\beta_0\beta_1 \sum x}{\sum x^2} + \beta_1^2) && \text{(pp 4.1.9 \& 4.1.10)} \\ &= \sum(\beta_0^2 + 2\beta_0\beta_1 x + x^2\beta_1^2) - 2x(\beta_0 + \beta_1 x)(\frac{\sum \beta_0 x}{\sum x^2} + \beta_1) + \\ &\quad \sigma^2 + \beta_0^2\frac{n^2\bar{x}^2}{S_{xx}^o} + 2n\beta_0\beta_1\bar{x} + \beta_1^2 S_{xx}^0 \\ &= n\beta_0^2 + \cancel{2n\beta_0\beta_1\bar{x}} + \cancel{\beta_1^2 S_{xx}^o} - 2\beta_0^2 n^2\frac{\bar{x}^2}{S_{xx}^o} - \cancel{4n\beta_0\beta_1\bar{x}} - \cancel{2\beta_1^2 S_{xx}^o} + \\ &\quad \sigma^2 + \beta_0^2\frac{n^2\bar{x}^2}{S_{xx}^o} + \cancel{2n\beta_0\beta_1\bar{x}} + \cancel{\beta_1^2 S_{xx}^o} \\ &= n\beta_0^2 - \beta_0^2 n^2\frac{\bar{x}^2}{S_{xx}^o} + \sigma^2\end{aligned}$$

□

Before we begin the proof of proposition 4.1.8 let us change our assumptions again, and now let $\hat{y}$ define fitted values by full model.

Proposition 4.1.11. $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}\sigma^2}{\sum(x-\bar{x})^2}$

Proof. Let us adopt matrix notation from equation 3.2. $Var(\hat{\beta})$ will give us a matrix that has variances of estimators $\hat{\beta}_0, \hat{\beta}_1$ in diagonal elements and their covariances in off-diagonal elements. Define 2×1 vector $\hat{\beta}$ as $\mathbf{b}$:

$$\begin{aligned}\text{Var}[\mathbf{b}] &= \mathbb{E}[\mathbf{b}^2] - \mathbb{E}[\mathbf{b}]\mathbb{E}[\mathbf{b}^T] \\ &= \mathbb{E}[((\mathbf{X^T X})^{-1}\mathbf{X^T Y})^2] - \beta^2\end{aligned}$$

Replace $\mathbf{Y}$ by eq. 3.2

$$\begin{aligned}&= \mathbb{E}[((\mathbf{X^T X})^{-1}\mathbf{X^T}(\mathbf{X}\beta + \varepsilon))^2] - \beta^2 \\ &= \mathbb{E}[((\mathbf{X^T X})^{-1}\mathbf{X^T X}\beta + (\mathbf{X^T X})^{-1}\mathbf{X^T}\varepsilon))^2] - \beta^2\end{aligned}$$

The term $(\mathbf{X^T X})^{-1}\mathbf{X^T X}$ is equal to identity matrix $\mathbf{I_2}$

$$\begin{aligned}&= \mathbb{E}[(\beta + (\mathbf{X^T X})^{-1}\mathbf{X^T}\varepsilon))^2] - \beta^2 \\ &= \beta^2 + \mathbb{E}[((\mathbf{X^T X})^{-1}\mathbf{X^T}\varepsilon)^2] - \beta^2\end{aligned}$$

The cross term becomes zero since $\mathbb{E}[\varepsilon] = 0$

$$\begin{aligned}&= ((\mathbf{X^T X})^{-1}\mathbf{X^T})^2\mathbb{E}[\varepsilon^2] \\ &= (\mathbf{X^T X})^{-1}\mathbf{X^T X}(\mathbf{X^T X})^{\mathbf{T}\ -1}\sigma^2\end{aligned}$$

$(\mathbf{X^T X})^{\mathbf{T}\ -1} = (\mathbf{X^T X})^{-1}$ comes from the fact that $X^T X$ is a symmetric matrix

$$= \sigma^2(\mathbf{X^T X})^{-1}$$

Now the off-diagonal term then can be computed easily as $\frac{-\bar{x}\sigma^2}{\sum(x-\bar{x})^2}$

□

Proof of proposition 4.1.8.

$$\begin{aligned}\mathbb{E}[SSD] &= \mathbb{E}[\sum(\dot{y}-\hat{y})^2] = \mathbb{E}[\sum(\dot{y}^2 - 2\dot{y}\hat{y} + \hat{y}^2)] = \\ &= n\beta_0^2 + 2\beta_0\beta_1\sum x + \beta_1^2\sum x^2 - \sum 2\dot{y}\mathbb{E}[\hat{y}] + \sum \mathbb{E}[\hat{y}^2] = (*)\end{aligned}$$

$(\mathbb{E}[\hat{y}] = \mathbb{E}[\hat{\beta}_1]x + \mathbb{E}[\hat{\beta}_0] = \beta_1 x + \beta_0 = \dot{y})$

$\mathbb{E}[\dot{y}^2] = \mathrm{Var}[\hat{\beta}_0] + \mathbb{E}[\hat{\beta}_0] + x^2(\mathrm{Var}[\hat{\beta}_1 + \mathbb{E}[\hat{\beta}_1]^2]) + 2x\mathbb{E}[\hat{\beta}_0\hat{\beta}_1]$

and $\mathbb{E}[\hat{\beta}_0\hat{\beta}_1] = \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \mathbb{E}[\hat{\beta}_0]\mathbb{E}[\hat{\beta}_0]$, so substituting prop 4.1.11 we get

$$(*) = n\beta_0^2 + 2\beta_0\beta_1 n\bar{x} + \beta_1^2 S_{xx}^o - \sum 2(\beta_1 x + \beta_0)^2 + \frac{\sigma^2}{}$$

□

4.1.3 Relevant Literature

Refer to seminal works and research studies that have explored or utilized the simple linear regression model without an intercept term. A brief review of the literature provides additional context and allows for a synthesis of existing knowledge in this specialized domain.

4.2 Comparative Analysis

5. Applications to Linear Regression through Origin

5.1 Something to add 1

5.2 Something to add 1

6. Theoretical results

6.1 A theoretical resilt

6.2 Towards some advanced topic

7. Programming simulations

8. Summary and closing words

Bibliography

[1] Fumio Hayashi. *Econometrics*. Princeton University Press, 2011.

[2] John Neter, William Wasserman, and Michael H Kutner. *Applied linear regression models*. Richard D. Irwin, 1983.

A. Program Codes

A.1 linear_regression.py module

```
1
2  import numpy as np
3  import matplotlib.pyplot as plt
4  from sklearn.linear_model import LinearRegression
5
6  def linear_regression(x, y, intercept=True):
7  """
8  facade function that implements LinearRegression from sklearn.linear_model
9  """
10 model = LinearRegression(fit_intercept=intercept)
11 x = x.reshape(-1,1)
12 y = y.reshape(-1,1)
13 model.fit(x,y)
14 return model.predict(x)
15
16 def bayesian_information_criterion(y, y_fit, n , sigma, k):
17 """
18 BIC
19 """
20 max_log_likelihood = (
21 -n/2 * np.log(2 * np.pi * sigma**2)-np.sum((y+y_fit)**2)/sigma**2
22 )
23 BIC = k*np.log(n)-2*max_log_likelihood
24 return BIC
25
```

```
26 def SSD(y, y_fit):
27 """
28 function calculates sum of squared deviations of fitted
29 values and true values
30 """
31 return np.sum((y-y_fit)**2)
32
```

A.2 Script 1

```
1  import numpy as np
2  import matplotlib.pyplot as plt
3
4  np.random.seed(42)
5
6  num_samples = 100  # number of sample points
7
8  #X = np.linspace(0, 1, num_samples) * 2 - 1
9  X = np.random.rand(num_samples) * 2 - 1
10
11 beta =  1 # true slope
12 beta_0 = 0.05
13 noise = np.random.normal(0, 0.5, num_samples)  # standard normal noise term
14
15 y = beta * X + noise + beta_0
16
17
18 plt.plot(
19 X, beta * X, color='red', linewidth=1, label='True Linear Relationship'
20 )
21 plt.scatter(
22 X, y, color='white', edgecolor='black', marker='o', label='Generated Data Points'
23 )
24 plt.legend()
```

```
25
26  plt.axhline(0, color='green', linewidth=1, linestyle='--')
27  plt.axvline(0, color='green', linewidth=1, linestyle='--')
28
29  plt.title('')
30  plt.show()
31
```