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

MŰEGYETEM 1782

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

Gerald J. Hahn [3] published an article on fitting models with no intercept term. In their paper, they focused on appropriateness of no-intercept linear regression model in different scenarios. For example, while physical nature of sample data might advocate for the use of no-intercept model, it would often happen that true nature of the data is not linear, and extrapolation of the data to the region that falls outside of the range of the sample makes it clear that other models should be used. They also provide a procedure for checking the significance of the intercept: one must fit the intercept model and then construct a confidence interval that contains true intercept, and then check, whether 0 is included in the interval. If it is included, then there is no evidence that the intercept model provides significantly better fit to the data than no intercept model. Hahn also raised an issue that was related the coefficient of determination that was erroneously computed in popular computing packages of that time.

Gordon [2] has also contributed to the dispute of computation of R^2 and suggested that this parameter has to be computed for both models by the same formula

$$R^2 = 1 - \frac{\sum_i \varepsilon_i^2}{\sum_i (y_i - \bar{y})^2} \tag{1.1}$$

However, as Eisenhauer [1] noted, the equation 1.1 might be inconsistent for no-intercept model and provide uninterpretable negative result, so a different formula for no-intercept model is commonly used,

$$R^2 = \frac{\sum \hat{y_i}^2}{\sum y_i^2} \tag{1.2}$$

This new formula becomes not consistent with the prior one, thus R^2 calculated for intercept and no-intercept models are incomparable.

1.3 Research question

The main goal of this paper is to familiarize reader with linear regression through the origin, derive properties of its estimators, and determine when RTO might be a better model than linear regression with intercept.

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 $(x_1,\ldots,x_n)$, where $x_i \in S$ for $i = 1,\ldots,n$. 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 $\mathbf{X} = (X_1,X_2,\ldots,X_n)$ 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 $\{P_\theta \,|\, \theta \in \Theta\}$ 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 $(X_1,\ldots,X_n)$ 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 $X_1,X_2,\ldots,X_n$.

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

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

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:

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

3.2 Assumptions

The objective of simple linear regression is, under some assumption [4, 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 it is crucial to know them.

Assumption 3.2.1 (linearity). $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$

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

Assumption 3.2.2 (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 $\mathbb{E}(\varepsilon_i) = \mathbb{E}(\mathbb{E}(\varepsilon_i|\mathbf{X})) = 0$ $\square$

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.

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

$\square$

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

Assumption 3.2.3 (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 3.2.4 (homoskedasticity). $\mathbb{E}(\varepsilon_i^2|\mathbf{X}) = \sigma^2 > 0 \quad (i = 1, 2, ..., n)$

Assumption 3.2.5 (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, conditioned on their observed values, in reality, $x_1, ..., x_n$ are realizations of random variables, so it is generally accepted to treat them as fixed. Although this might be bad technically, it allows us to shorten the notation, and make it more readable, while still being concise.

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.

3.3 Properties

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 $\operatorname{Var} y = \sigma^2$. Then, likelihood L can be computed as a product of marginal distributions:

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

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

Log-likelihood function

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

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 3.3.1. *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. □

Remark: Equations 3.8 and 3.9 are called normal equations [5] (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 3.3.2. $\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SSE}$

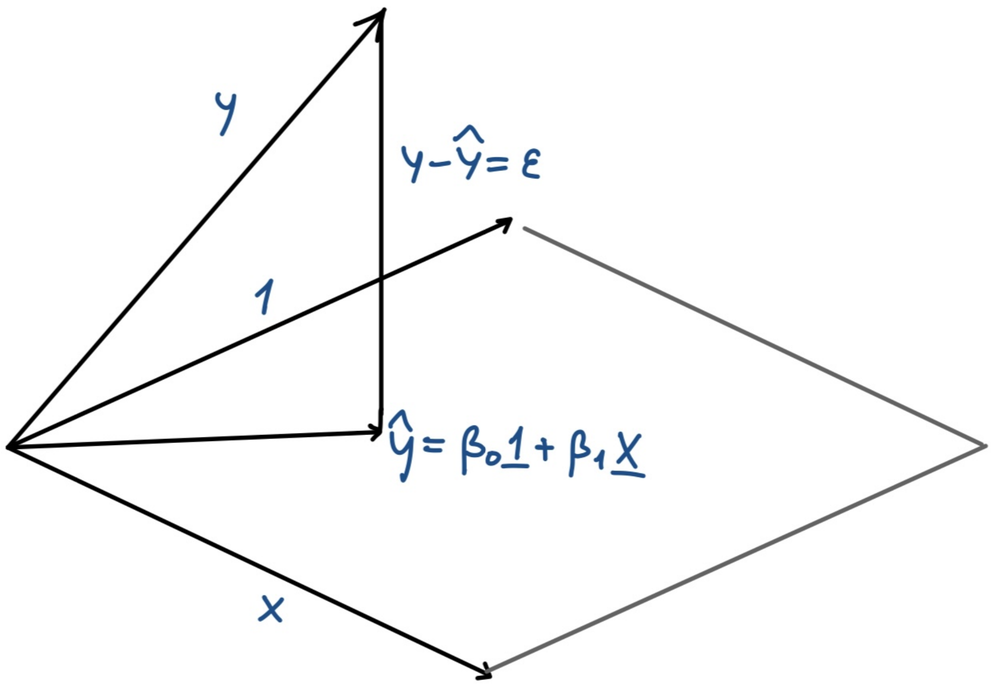


Figure 3.1: Orthogonal projection of y to

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:

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

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

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 $\hat{\beta}_1$:**

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

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

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

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

□

Proposition 3.3.4. $\mathrm{Var}[\hat{\beta}_0] = \frac{\sigma^2 S_{xx}^o}{n S_{xx}}$

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 assumed to be zero, when there is no displacement. In other words, physical nature of our data often bears information about the nature of the true model.

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:

$$\begin{aligned}
\frac{\partial l}{\partial \beta_1^o} &= -\frac{1}{2\sigma^2}\sum 2(y_i - \beta_1^o x_i)(-x_i) = 0 \\
\frac{1}{\sigma^2}\sum(y_i x_i - \beta_1^o x_i^2) &= 0 \\
\sum x_i y_i &= \sum x_i^2 \beta_1^o \\
\hat{\beta}_1^o &= \frac{\sum x_i y_i}{\sum x_i^2}
\end{aligned}$$

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

Proof.

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1^o] = \mathbb{E}[\frac{\sum x_i y_i}{\sum x_i^2}] = \sum \frac{1}{x_i^2}\mathbb{E}[\sum x_i y_i] = \\
= \frac{\sum x_i \mathbb{E}[y_i]}{\sum x_i^2} = \frac{\sum x_i^2 \beta_1^o}{\sum x_i^2} = \beta_1^o
\end{aligned}$$

□

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.

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

□

This tells us that when the true intercept is zero, no-intercept model provides better fit, as the two estimators $(\hat{\beta}_1, \hat{\beta}_1^o)$ are unbiased, but the latter has smaller variance. 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:*

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

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

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

Proof.

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

□

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

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

Remark: This is only true when true intercept is equal to one, and both of the estimators are MLE.

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.

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

□

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, $\dot{y}_i = \beta_1 x_i + \beta_0$, one might simulate the data points by adding normally-distributed error terms:

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

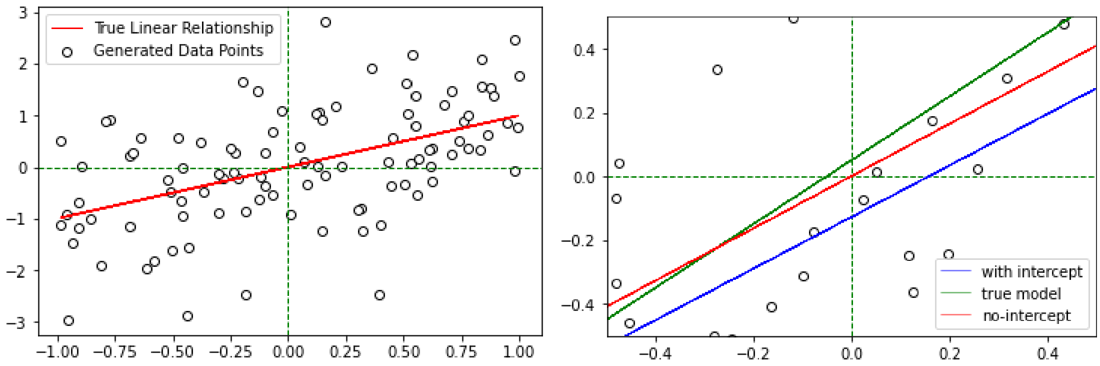
One might be interested now in comparing sum of squared deviations (SSD) of fitted points $\sum(\dot{y}_i - \hat{y}_i)^2$.

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:



(a) Plot of simulated points (b) Plot of three models

Figure 4.1: Linear Regression model on simultated datapoints

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

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

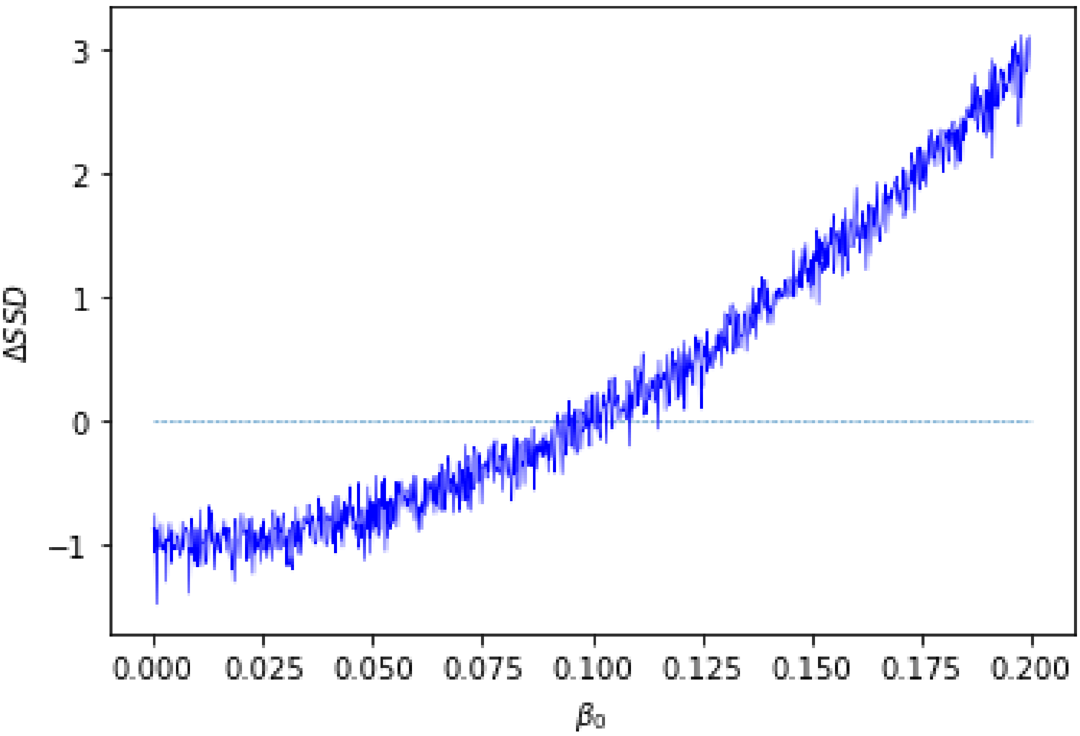


Figure 4.2: Plot of the difference of SSD to the values of intercept

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

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^o_1 is now biased.

Proposition 4.1.9. $\mathbb{E}[\hat{\beta}^o_1] = \frac{\sum \beta_0 x}{\sum x^2} + \beta_1$

Proof.

$$\begin{aligned}\mathbb{E}[\hat{\beta}^o_1] &= \frac{\mathbb{E}\sum xy}{\sum x^2} = \frac{\sum x \mathbb{E}[y]}{\sum x^2} = \frac{\sum x(\beta_1 x + \beta_0)}{\sum x^2}\\ &= \frac{\sum \beta_0 x + \sum \beta_1 x^2}{\sum x^2} = \frac{\sum \beta_0 x}{\sum x^2} + \beta_1\end{aligned}$$

□

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

Proof. Comes from the fact that $\mathbb{E}[\hat{\beta}^{o2}_1] = \mathrm{Var}[\hat{\beta}^o_1] + \mathbb{E}[\hat{\beta}^o_1]^2$ 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:

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

□

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

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 $\mathbf{b}$:

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

$$\begin{aligned}(*) &= n\beta_0^2 + 2\beta_0\beta_1 n\bar{x} + \beta_1^2 S_{xx}^o - \textstyle\sum 2(\beta_1 x + \beta_0)^2 + \frac{\sigma^2 S_{xx}^o}{S_{xx}} + n\beta_0^2 + \frac{\sigma^2 S_{xx}^o}{S_{xx}} + \\ &\quad + \beta_1^2 S_{xx}^o - \frac{2n\bar{x}^2\sigma^2}{S_{xx}} + 2\beta_0\beta_1 x \\ &= n\beta_0^2 + 2\beta_0\beta_1 n\bar{x} + \beta_1^2 S_{xx}^o - 2\beta_1^2 S_{xx}^o - 4\beta_0\beta_1 n\bar{x} - 2n\beta_0^2 + \frac{\sigma^2 S_{xx}^o}{S_{xx}} + n\beta_0^2 + \frac{\sigma^2 S_{xx}^o}{S_{xx}} + \\ &\quad + \beta_1^2 S_{xx}^o - \frac{2n\bar{x}^2\sigma^2}{S_{xx}} + 2\beta_0\beta_1 x \\ &= \frac{2\sigma^2}{S_{xx}}(S_{xx}^o - n\bar{x}^2)\end{aligned}$$

□

Looking at the expected value of SSD for both models we see that expected SSD for full model is only explained by variance of the error terms, whereas no-intercept model's SSD also depends on the value of true intercept. Therefore for small values of intercept no-intercept model performs better in terms of sum of squared deviations of fitted values from true values.

5. Summary and closing words

5.1 Conclusion

In this thesis, we have compared several properties of RTO and OLS models, found variances of estimators of slope, created confidence interval for difference of two slope estimators, and found found average values of sum of squared deviations of fitted values from true values, showing that on average, for small intercept values, linear regression through the origin provides better fit.

Bibliography

[1] Joseph G Eisenhauer. Regression through the origin. *Teaching statistics*, 25(3):76–80, 2003.

[2] HA Gordon. Errors in computer packages. least squares regression through the origin. *Journal of the Royal Statistical Society Series D: The Statistician*, 30(1):23–29, 1981.

[3] Gerald J Hahn. Fitting regression models with no intercept term. *Journal of Quality Technology*, 9(2):56–61, 1977.

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

[5] 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

```python

import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression

def linear_regression(x, y, intercept=True):
"""
facade function that implements LinearRegression from sklearn.linear_model
"""
model = LinearRegression(fit_intercept=intercept)
x = x.reshape(-1,1)
y = y.reshape(-1,1)
model.fit(x,y)
return model.predict(x)

def bayesian_information_criterion(y, y_fit, n , sigma, k):
"""
BIC
"""
max_log_likelihood = (
-n/2 * np.log(2 * np.pi * sigma**2)-np.sum((y+y_fit)**2)/sigma**2
)
BIC = k*np.log(n)-2*max_log_likelihood
return BIC

```

```python
def SSD(y, y_fit):
"""
function calculates sum of squared deviations of fitted
values and true values
"""
return np.sum((y-y_fit)**2)

```

A.2 Script 1

```python
import numpy as np
import matplotlib.pyplot as plt

np.random.seed(42)

num_samples = 100  # number of sample points

#X = np.linspace(0, 1, num_samples) * 2 - 1
X = np.random.rand(num_samples) * 2 - 1

beta =  1 # true slope
beta_0 = 0.05
noise = np.random.normal(0, 0.5, num_samples)  # standard normal noise term

y = beta * X + noise + beta_0


plt.plot(
X, beta * X, color='red', linewidth=1, label='True Linear Relationship'
)
plt.scatter(
X, y, color='white', edgecolor='black', marker='o', label='Generated Data Points'
)
plt.legend()
```

```python

plt.axhline(0, color='green', linewidth=1, linestyle='--')
plt.axvline(0, color='green', linewidth=1, linestyle='--')

plt.title('')
plt.show()

```