

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, January 29, 2024



M Ű E G Y E T E M 1 7 8 2

Contents

1	Introduction	1
1.1	Background	1
1.2	Literature Review	1
1.3	Research Question	2
1.4	Outline	2
2	Preliminaries	3
2.1	Statistics Basics	3
3	Simple Linear Regression	5
3.1	Simple linear regression	5
3.2	Assumptions	5
3.3	Properties	7
4	Linear Regression Regression with no intercept term	13
4.1	Simple Linear Regression with no intercept term	13
4.2	Model Comparison	16
5	Summary and closing words	25
5.1	Conclusion	25
A	Appendix	27
A.1	Miscellaneous calculations	27
A.2	linear_regression.py module	27
A.3	Script 1	29
A.4	Script 2	30

CONTENTS

ii

A.5 Script 3	32
------------------------	----

List of Tables

4.1	AIC, BIC, and SSD for Intercept and No Intercept Models.	20
-----	--	----

List of Figures

3.1	Orthogonal projection of \mathbf{y} to $\text{span}\{\mathbf{1}, \mathbf{x}\}$	9
4.1	Linear Regression model on simulated datapoints	20
4.2	Plot of the difference of SSD to the values of intercept	21

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 [5] published an article on fitting models with no intercept term. In their paper, they focused on the appropriateness of the no-intercept linear regression model in different scenarios. For example, while the physical nature of sample data might advocate for the use of a no-intercept model, it often happens that the true nature of the data is not linear. Extrapolation of the data to the region that falls outside 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 the 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 a significantly better fit to the data than the no-intercept model. Hahn also raised an issue related to the coefficient of determination that was

erroneously computed in popular computing packages of that time.

Gordon [4] has also contributed to the dispute over the computation of R^2 and suggested that this parameter has to be computed for both models using the same formula

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

However, as Eisenhauer [3] noted, the equation 1.1 might be inconsistent for the no-intercept model and provide an uninterpretable negative result. So, a different formula for the no-intercept model is commonly used:

$$R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} \quad (1.2)$$

This new formula is 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 the reader with linear regression through the origin, derive properties of its estimators, compare RTO with the full linear regression model with an intercept, and determine when RTO might outperform the full OLS model.

1.4 Outline

We will start with a recap of some statistical notions that are fundamental to the topic of discussion. Then, we will go through the necessary conditions of OLS regression and connect the notion of least squares with maximum likelihood. After that, we will derive properties of linear regression for both intercept and no-intercept models, and lastly, provide an example of the case when the no-intercept model is better than the full model and try to explain why.

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 $T = T(x_1, \dots, x_n)$ 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

2. PRELIMINARIES

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, \dots, X_n) from it. The mean square error (or quadratic risk) of an estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ for the parameter θ is defined as

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

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, \dots, 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 .

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

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 $L(\theta)$

Remark: For the sake of convenience, the summation operator and summands will frequently be expressed without an explicit index. The index will only be specified when its omission might cause confusion.

For example:

$$\sum (x - \bar{x})$$

Is equivalent to:

$$\sum_{i=1} (x_i - \bar{x})$$

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 $(x_1, y_1), \dots, (x_n, y_n)$ 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:

$$y_i = \beta_1 x_i + \beta_0 \quad (i = 1, 2, \dots, n) \quad (3.1)$$

or in matrix notation

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

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

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

$$\mathbf{Y} = \beta \mathbf{X} + \varepsilon \quad (3.2)$$

3.2 Assumptions

The objective of simple linear regression is, under some assumption [6, p.4-12], to estimate the parameters β_0 and β_1 , so that they will provide best fit. The assumptions

3. SIMPLE LINEAR REGRESSION

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). $\mathbb{E}(\varepsilon_i|\mathbf{X}) = 0 \quad (i = 1, 2, \dots, n)$

From this assumption it follows that:

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

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

The cross moments are orthogonal to all observations:

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

Proof.

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

□

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 collinearity). *Vector (x_1, x_2, \dots, x_n) is not collinear with constant $\mathbf{1}$ vector with probability 1*

The more general variant of this assumption, in the case when we have classical linear regression with multiple regressors, is no multicollinearity. No 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.

3. SIMPLE LINEAR REGRESSION

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

Assumption 3.2.5 (no correlation between observatoins).

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

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, \dots, x_n are realizations of random variables, so it is generally accepted to treat them as fixed. Although this might be somewhat misleading, it allows us to shorten the notation, and make it more readable, while still being concise.

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

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

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

Log-likelihood function

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

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:

3. SIMPLE LINEAR REGRESSION

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

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

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

Let the solutions of the above equations be denoted as $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$ for $\beta_0, \beta_1, \sigma^2$. If $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $S_{xx} = \sum (x - \bar{x})^2, S_{xy} = \sum (x - \bar{x})y$, then...

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}; \quad (3.11)$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}; \quad (3.12)$$

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

So $\hat{\beta}_0, \hat{\beta}_1$ are Maximum Likelihood 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 $\text{span}\{\mathbf{1}, \mathbf{x}\}$. Notice that $y - \hat{y} = \varepsilon$, so since \hat{y} is perpendicular to ε , MLE from 3.11 and 3.12 are indeed MSE estimators too. \square

Remark: Equations 3.8 and 3.9 are called normal equations [7] (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:

$$\mathbf{X}^T \mathbf{X} \mathbf{b} = \mathbf{X}^T \mathbf{Y} \quad (3.14)$$

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

Proposition 3.3.2.
$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SSE}$$

3. SIMPLE LINEAR REGRESSION

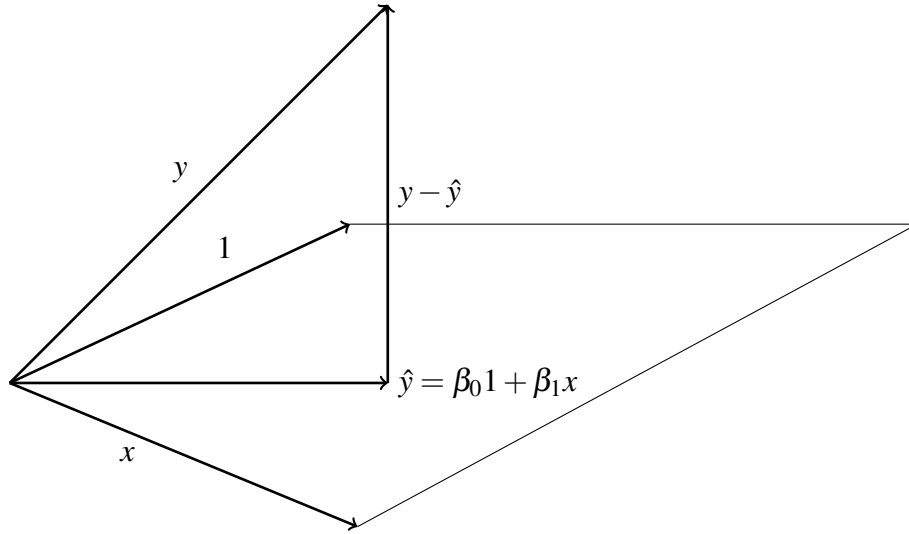


Figure 3.1: Orthogonal projection of \mathbf{y} to $\text{span}\{\mathbf{1}, \mathbf{x}\}$

Proof. The vector $\mathbf{y} - \hat{\mathbf{y}}$ is perpendicular to $\hat{\mathbf{y}} - \bar{\mathbf{y}}$, 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:

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

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,

3. SIMPLE LINEAR REGRESSION

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

□

Conventionally the terms $\sum_{i=1}^n (y_i - \bar{y})^2$, $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ are called Sum of Squares Total (SST), Sum of Squares due to Regression (SSR), and Sum of Squares Error (SSE), respectively.

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

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

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

$$\begin{aligned} \mathbb{E}(\hat{\beta}_0) &= \mathbb{E}(\bar{y} - \hat{\beta}_1 \bar{x}) = \bar{y} - \bar{x} \mathbb{E}(\hat{\beta}_1) = \frac{1}{n} \mathbb{E}[\sum y_i] - \beta_1 \bar{x} = \\ &= \frac{1}{n} \mathbb{E}[\sum (\beta_0 + \beta_1 x_i + \varepsilon)] - \beta_1 \bar{x} = \frac{1}{n} n \beta_0 + \frac{1}{n} n \beta_1 \bar{x} - \bar{x} \beta_1 = \beta_0 \end{aligned}$$

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

Define matrix

$$U = \begin{bmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{(x_1 - \bar{x})}{\sqrt{S_{xx}}} & \cdots & \frac{(x_n - \bar{x})}{\sqrt{S_{xx}}} \\ \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

3. SIMPLE LINEAR REGRESSION

Where all rows after the second one are deterministic such that U is orthogonal, then $UY = Z$, so that:

$$Z_1 = \sqrt{n}\bar{Y}, Z_2 = \frac{S_{xy}}{\sqrt{S_{xx}}}, \text{ and } Z_j \sim N(0, \sigma^2) \quad (j = 3, \dots, n)$$

Let $\sum_i Z_i^2 = \sum_i Y_i^2$. Notice that we can do this, since choice of U was arbitrary up to the orthogonality of the matrix. Then:

$$\begin{aligned} \sum y^2 &= \sum (y - \bar{y})^2 + n\bar{y}^2 \\ &= \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2 + n\bar{y}^2 \\ &= \sum (\hat{\beta}_0 + \beta_1 x - \hat{\beta}_0 - \hat{\beta}_1 \bar{x})^2 + \sum (y - \hat{y})^2 + n\bar{y}^2 \\ &= \sum (\hat{\beta}_1 x - \hat{\beta}_1 \bar{x})^2 + \sum (y - \hat{y})^2 + n\bar{y}^2 \\ &= \hat{\beta}_1^2 \sum (x - \bar{x})^2 + \sum (y - \hat{y})^2 + n\bar{y}^2 \\ &= \frac{S_{xy}^2}{S_{xx}} + n\bar{y}^2 + \sum (y - \hat{y})^2 \end{aligned}$$

Since $\sum y^2 \sim \sigma^2 \chi(n)$, it follows that $\sum (y - \bar{y})^2 \sim \sigma^2 \chi(n-2)$

Hence, $\mathbb{E}[\frac{SSE}{n-2}] = \frac{\sigma^2(n-2)}{n-2} = \sigma^2$

Proposition 3.3.3. $\text{Var}[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}$

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

$$\text{Var}[\hat{\beta}_1] = \text{Var}\left(\frac{1}{S_{xx}} \sum (x_i - \bar{x}) Y_i\right) = \frac{1}{S_{xx}^2} \sum \text{Var} Y_i = \frac{\sigma^2}{S_{xx}}$$

□

Proposition 3.3.4. $\text{Var}[\hat{\beta}_0] = \frac{\sigma^2 S_{xx}^o}{n S_{xx}}$, where $S_{xx}^o = \sum x^2$

Proof.

$$\text{Var}[\hat{\beta}_0] = \text{Var}[\bar{y} - \bar{x}\hat{\beta}_1] = \text{Var}[\bar{y}] + \bar{x}^2 \text{Var}[\hat{\beta}_1] - 2\bar{x} \text{Cov}(\bar{y}, \hat{\beta}_1) =$$

3. SIMPLE LINEAR REGRESSION

$(\text{Cov}(\bar{y}, \hat{\beta}_1) = 0$ see Appendix A.1.1)

$$\begin{aligned} &= \frac{1}{n^2} \text{Var}[\sum y] + \bar{x}^2 \frac{\sigma^2}{(x - \bar{x})^2} \\ &= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{(x - \bar{x})^2} \\ &= \frac{\sigma^2 \sum (x - \bar{x})^2}{n \sum (x - \bar{x})^2} + \frac{n \bar{x}^2 \sigma^2}{n \sum (x - \bar{x})^2} \\ &= \frac{\sigma^2}{n \sum (x - \bar{x})^2} (\sum (x - \bar{x})^2 + n \bar{x}^2) \\ &= \frac{\sigma^2 \sum x^2}{n \sum (x - \bar{x})^2} \end{aligned}$$

□

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:

$$\mathbf{y} = \beta_1^o \mathbf{x} + \varepsilon, \quad (4.1)$$

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

Likelihood function L is:

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

Log-likelihood l is:

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

Thus we can compute the maximum likelihood estimator of β_1^o by taking partial

4. LINEAR REGRESSION REGRESSION WITH NO INTERCEPT TERM

derivative of log-likelihood with respect to β_1^o :

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

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

Proof.

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

□

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

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

□

Proposition 4.1.3.

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

Proof.

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

□

4. LINEAR REGRESSION REGRESSION WITH NO INTERCEPT TERM

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

Proof.

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

□

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 zero, since both of the estimators are MLE in this case.

Now we are ready to prove proposition 4.1.4

4. LINEAR REGRESSION REGRESSION WITH NO INTERCEPT TERM

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.

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

□

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

$$\begin{aligned}\hat{\beta}_1^0 - \hat{\beta}_1 &\sim N(\beta_1^0 - \beta_1, \sigma^2(\frac{1}{S_{xx}} - \frac{1}{S_{xx}^0})) \\ Z &:= \frac{\hat{\beta}_1^0 - \hat{\beta}_1 - \mathbb{E}[\hat{\beta}_1^0 - \hat{\beta}_1]}{\text{Var}(\hat{\beta}_1^0 - \hat{\beta}_1)} = \frac{\hat{\beta}_1^0 - \hat{\beta}_1 - (\beta_1^0 - \beta_1)}{\sigma^2(\frac{1}{S_{xx}} - \frac{1}{S_{xx}^0})} \\ \beta_1^0 - \beta_1 &= \hat{\beta}_1^0 - \hat{\beta}_1 - \sigma^2 Z(\frac{1}{S_{xx}} - \frac{1}{S_{xx}^0})\end{aligned}$$

CI is:

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

4.2 Model Comparison

The goal of this section is to compare two models given that the observed data is generated from the function $y_i = \beta_1 x + \beta_0 + \varepsilon_i$ with our standard assumptions.

Coefficient of Determination

R^2 or a coefficient of determination is defined as

$$R^2 = 1 - \frac{SSE}{SST}$$

4. LINEAR REGRESSION REGRESSION WITH NO INTERCEPT TERM

Is is often used to measure how well a model predicts an outcome. It is a proportion of variation in the dependent variable that is explained by independent variables.

However, notion of SST and SSR for no-intercept model is different from the full model as defined in prop. 3.3.2, since $\sum(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$ will generally take a non-zero value, so the proof of prop. 3.3.2 fails. It is often argued [3] that defining SST as the sum of squared deviations from the mean is inappropriate when the regression line passes through the origin, but does not necessarily pass through (\bar{x}, \bar{y}) , therefore equation in prop. 3.3.2 is replaced by:

$$\underbrace{\sum_{i=1}^n y_i^2}_{SST} = \underbrace{\sum_{i=1}^n \hat{y}_i^2}_{SSR} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SSE}$$

Considering the new definition of *SST* for RTO, it now does not make much sense to use R^2 values of RTO and ordinary OLS linear regressions for model comparison. However it still important to mention this, because this measure is widely used and will result in erroneous inferences if one does not realize this subtle difference.

Akaike Information Criterion

Akaike Information Criterion is a statistical tool that was introduced by Akaike in 1974 [1] and defined as:

$$AIC = (-2) \log(\text{maximum likelihood}) + 2k$$

Where k is the number of independently adjusted parameters within the model.

AIC in a sense estimates the prediction error for a given set of data, and effectively takes into account the number of assumed parameters, thus punishing overfitting, which often occurs due to increasing number of parameters in the model also increases likelihood, even if the model may give a poorer fit.

AIC has its relation to another important notion in statistics, Kullback-Leibler divergence, which is defined as follows:

Given that x_1, x_2, \dots, x_n are obtained results after n independent observations of random variable with probability function $g(x)$, and a parametric family of density function is given by $f(x|\theta)$ then the average log-likelihood tends to this integral (assuming that

4. LINEAR REGRESSION REGRESSION WITH NO INTERCEPT TERM

integral exists) with probability 1

$$S(g; f(\cdot|\theta)) = \int g(x) \log f(x|\theta) dx \quad (4.3)$$

The idea behind using the log function here is that it is the most sensitive to small deviations of $f(x|\theta)$ from $g(x)$. Then the difference

$$I(g; f(\cdot|\theta)) = S(g; g) - S(g; f(\cdot|\theta))$$

is called Kullback-Leibner mean information, and also can be interpreted as "distance" between two distributions f and g , since $S(g; g) - S(g; f(\cdot|\theta))$ is also

$$\mathbb{E}_g[\log g(x)] - \mathbb{E}_g[\log f(x|\theta)]$$

with respect to the density function $g(x)$. Hence, low values of KL divergence indicate that the two probability density functions are relatively close to each other.

Akaike found that maximized log-likelihood is a biased estimate of $\mathbb{E}[\log f(x|\hat{\theta})]$, where $\hat{\theta}$ is a biased estimate of θ , where bias approximately equals k , the overall number of independent parameters of a given model. Since, $\mathbb{E}[\log g(x)]$ is a constant, minimizing AIC means minimizing the estimate of KL divergence.

Bayesian Information Criterion

Bayesian Information Criterion, Schwartz (1978) [8], is alternative estimate of goodness of fit of the model that penalizes the number of independent parameters and defined as:

$$BIC = -2 \log(\text{maximized likelihood}) + k \log n \quad (4.4)$$

The way how AIC and BIC penalize increasing number of parameters is different, and ultimately the choice between two criteria should depend on assumptions about reality and the intent of the model-based inference [2]. However, for the sake of practice, we shall use both AIC and BIC in our model comparison.

4.2.1 $\beta_0 = 0$:

When the true intercept term is equal to zero, for both models we have that their parameters are MLE of true parameters. However, from prop. 4.1.3 we know that the slope estimator of the no-intercept model is more accurate, hence no-intercept model is preferred.

4.2.2 $\beta_0 \neq 0$:

$$\sum (\hat{y} - \hat{y})^2$$

Where $\hat{y} = \beta_1 x + \beta_0$ and β_0, β_1 are true intercept and slope values, respectively.

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:

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

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

Let's start by generating a sample of 100 points, and assume that parameters are known as $\beta = 1, \alpha = 0.005, \sigma = 1$ (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 (The code for plotting all of the plots can be found in appendix).

Now let's compare three statistics for these two models: BIC, AIC and SSD. The results can be seen in the Table 4.1. It is clearly seen that no-intercept model provides

4. LINEAR REGRESSION REGRESSION WITH NO INTERCEPT TERM

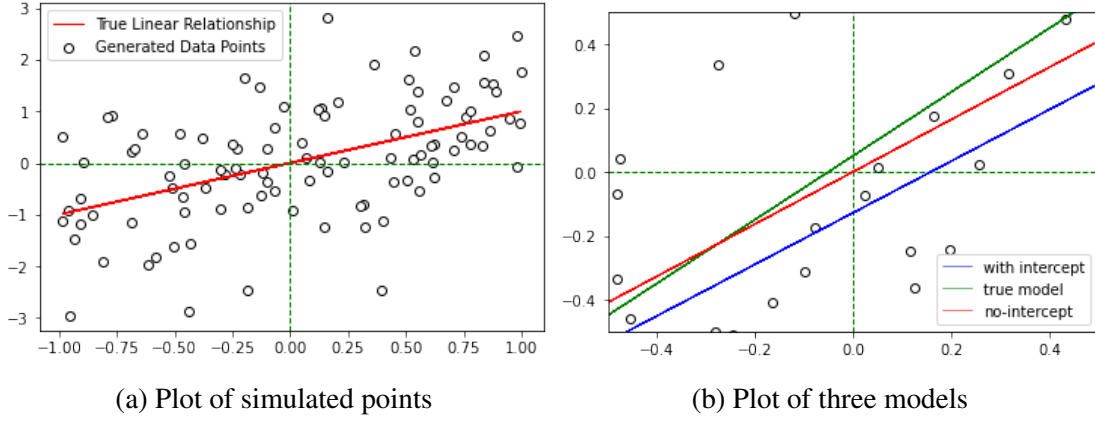


Figure 4.1: Linear Regression model on simulated datapoints

	AIC	BIC	SSD
Intercept Model	1102.35	1110.16	0.46305
No Intercept Model	1100.24	1105.45	0.45767

Table 4.1: AIC, BIC, and SSD for Intercept and No Intercept Models.

slightly 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 below 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.

4.2.3 Expected value of SSD and SSD_{no}

We will denote with SSD and SSD_{no} the sum of squared deviations of the fitted values from values on the true relationship line for full and no-intercept models, respectively.

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

Proposition 4.2.2. $\mathbb{E}[SSD] = \frac{2\sigma^2}{S_{xx}}(S_{xx}^o - n\bar{x}^2)$

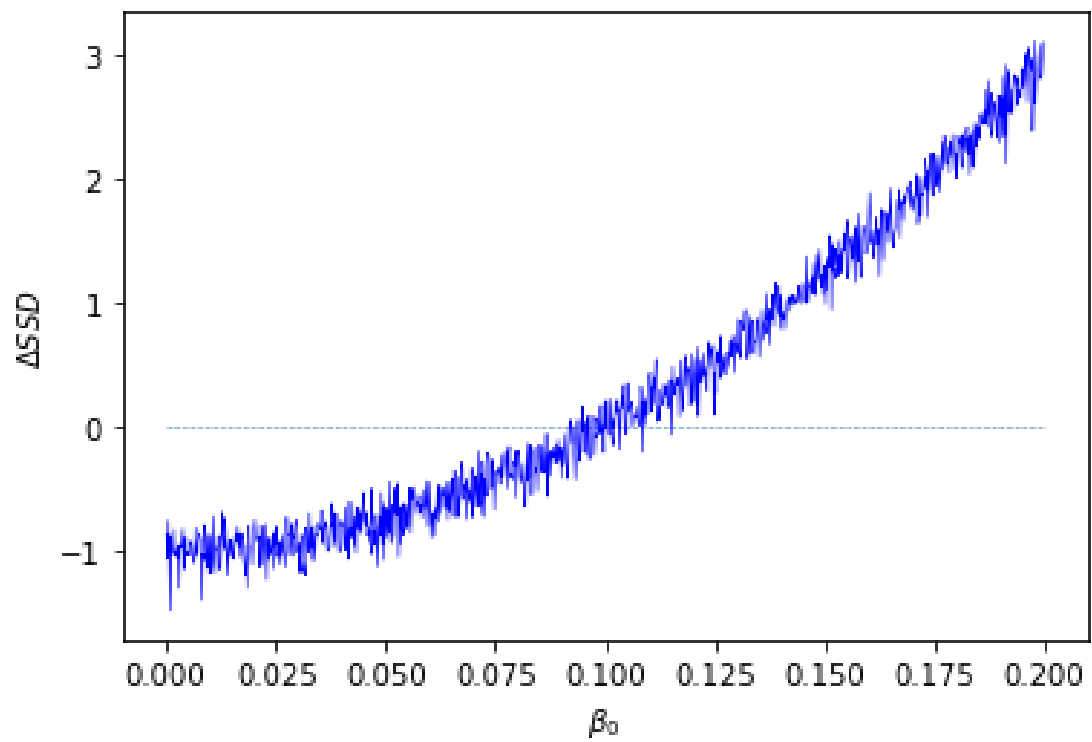


Figure 4.2: Plot of the difference of SSD to the values of intercept

4. LINEAR REGRESSION REGRESSION WITH NO INTERCEPT TERM

Assuming $\beta_0 > 0$. In order to prove proposition 4.2.1, we have to take into account that $\beta_0 \neq 0$, so the estimator β_1^o is now biased.

Proposition 4.2.3. $\mathbb{E}[\hat{\beta}_1^o] = \frac{\sum \beta_0 x}{\sum x^2} + \beta_1$

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

Proof. Comes from the fact that $\mathbb{E}[\hat{\beta}_1^{o2}] = \text{Var}[\hat{\beta}_1^o] + \mathbb{E}[\hat{\beta}_1^o]^2$ and then one can apply prop. 4.2.3 and prop. 4.1.2

□

Proof of proposition 4.2.1. Here, we assume that \hat{y} are values fitted by no-intercept model:

$$\begin{aligned}\mathbb{E}[SSD_{no}] &= \sum (\hat{y} - \hat{y})^2 && (\text{df } SSD_{no}) \\ &= \sum \hat{y}^2 - \sum 2\hat{y}\mathbb{E}[\hat{y}] + \sum \mathbb{E}[\hat{y}^2] && (\text{linearity}) \\ &= \sum \hat{y}^2 - \sum 2\hat{y}x(\frac{\sum \beta_0 x}{\sum x^2} + \beta_1) + \sum x^2(\frac{\sigma^2}{\sum x^2} + (\frac{\sum \beta_0 x}{\sum x^2})^2 + \frac{2\beta_0\beta_1 \sum x}{\sum x^2} + \beta_1^2) && (\text{pp 4.2.3 \& 4.2.4}) \\ &= \sum (\beta_0^2 + 2\beta_0\beta_1 x + x^2\beta_1^2) - 2x(\beta_0 + \beta_1 x)(\frac{\sum \beta_0 x}{\sum x^2} + \beta_1) + \\ &\quad \sigma^2 + \beta_0^2 \frac{n^2 \bar{x}^2}{S_{xx}^o} + 2n\beta_0\beta_1 \bar{x} + \beta_1^2 S_{xx}^o \\ &= n\beta_0^2 + \cancel{2n\beta_0\beta_1 \bar{x}} + \beta_1^2 S_{xx}^o - 2\beta_0^2 n^2 \frac{\bar{x}^2}{S_{xx}^o} - \cancel{4n\beta_0\beta_1 \bar{x}} - \cancel{2\beta_1^2 S_{xx}^o} + \\ &\quad \sigma^2 + \beta_0^2 \frac{n^2 \bar{x}^2}{S_{xx}^o} + \cancel{2n\beta_0\beta_1 \bar{x}} + \beta_1^2 S_{xx}^o \\ &= n\beta_0^2 - \beta_0^2 n^2 \frac{\bar{x}^2}{S_{xx}^o} + \sigma^2\end{aligned}$$

□

4. LINEAR REGRESSION REGRESSION WITH NO INTERCEPT TERM

Before we begin the proof of proposition 4.2.2 let us change our assumptions again, and now let \hat{y} define fitted values by full model.

Proposition 4.2.5. $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}\sigma^2}{\sum(x-\bar{x})^2}$

Proof. Let us adopt matrix notation from equation 3.2. $\text{Var}(\hat{\beta})$ 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} :

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

Replace \mathbf{Y} by eq. 3.2

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

The term $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}$ is equal to identity matrix \mathbf{I}_2

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

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

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

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

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

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

□

4. LINEAR REGRESSION REGRESSION WITH NO INTERCEPT TERM

Proof of proposition 4.2.2.

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

$$(\mathbb{E}[\hat{y}] = \mathbb{E}[\hat{\beta}_1]x + \mathbb{E}[\hat{\beta}_0] = \beta_1x + \beta_0 = \hat{y})$$

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

and $\mathbb{E}[\hat{\beta}_0\hat{\beta}_1] = \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \mathbb{E}[\hat{\beta}_0]\mathbb{E}[\hat{\beta}_1]$, so substituting prop 4.2.5 we get

$$\begin{aligned} (*) &= n\beta_0^2 + 2\beta_0\beta_1n\bar{x} + \beta_1^2S_{xx}^o - \sum 2(\beta_1x + \beta_0)^2 + \frac{\sigma^2S_{xx}^o}{S_{xx}} + n\beta_0^2 + \frac{\sigma^2S_{xx}^o}{S_{xx}} + \\ &\quad + \beta_1^2S_{xx}^o - \frac{2n\bar{x}^2\sigma^2}{S_{xx}} + 2\beta_0\beta_1x \\ &= n\beta_0^2 + 2\beta_0\beta_1n\bar{x} + \beta_1^2S_{xx}^o - 2\beta_1^2S_{xx}^o - 4\beta_0\beta_1n\bar{x} - 2n\beta_0^2 + \frac{\sigma^2S_{xx}^o}{S_{xx}} + n\beta_0^2 + \frac{\sigma^2S_{xx}^o}{S_{xx}} + \\ &\quad + \beta_1^2S_{xx}^o - \frac{2n\bar{x}^2\sigma^2}{S_{xx}} + 2\beta_0\beta_1x \\ &= \frac{2\sigma^2}{S_{xx}}(S_{xx}^o - n\bar{x}^2)\end{aligned}$$

□

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 is expected to perform 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 conducted a comparative analysis of several properties between RTO and OLS models. We explored the variances of estimators of the slope, created a confidence interval for the difference of two slope estimators, and determined the average values of the sum of squared deviations of fitted values from true values. Our findings indicate that, on average, for small intercept values, linear regression through the origin provides a better fit.

When we augment the normal OLS assumptions with the additional assumption that the true intercept term is zero, we can confidently assert that RTO is a valid choice between the two models. This is because RTO is unbiased and provides a smaller variance.

Fitting RTO to a dataset generated from a linear function that includes an intercept term revealed that, although RTO produced biased estimators, the bias for very small intercept values can be mitigated by its simplicity. The full model includes an estimate for the intercept term, increasing the overall variability of the model. On the other hand, the no-intercept model lacks this estimate, resulting in a more precise model, even though it is generally less accurate.

Bibliography

- [1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.
- [2] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.
- [3] Joseph G Eisenhauer. Regression through the origin. *Teaching statistics*, 25(3):76–80, 2003.
- [4] 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.
- [5] Gerald J Hahn. Fitting regression models with no intercept term. *Journal of Quality Technology*, 9(2):56–61, 1977.
- [6] Fumio Hayashi. *Econometrics*. Princeton University Press, 2011.
- [7] John Neter, William Wasserman, and Michael H Kutner. *Applied linear regression models*. Richard D. Irwin, 1983.
- [8] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

A. Appendix

A.1 Miscellaneous calculations

Proposition A.1.1. $\text{Cov}(\bar{y}, \hat{\beta}_1) = 0$

Proof.

$$\begin{aligned}\text{Cov}(\bar{y}, \hat{\beta}_1) &= \text{Cov}\left(\frac{1}{n} \sum_i y_i, \frac{\sum_j (x_j - \bar{x}) y_j}{\sum (x - \bar{x})^2}\right) \\ &= \frac{1}{n} \sum_i \sum_j \frac{x_j - \bar{x}}{\sum_i (x_i - \bar{x})^2} \text{Cov}(y_j, y_i) \\ &= \frac{1}{n} \sum_i \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \sigma^2 \\ &= 0\end{aligned}$$

since $\sum (x_i - \bar{x}) = 0$

□

A.2 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 calculation for OLS
19     """
20    max_log_likelihood = (
21    -n/2 * np.log(2 * np.pi * sigma**2)-np.sum((y-y_fit)**2)/(2*sigma**2)
22    )
23    BIC = k*np.log(n)-2*max_log_likelihood
24    return BIC
25
26 def akaike_information_criterion(y, y_fit, n , sigma, k):
27     """
28     AIC calculation for OLS
29     """
30    max_log_likelihood = (
31    -n/2 * np.log(2 * np.pi * sigma**2)-np.sum((y-y_fit)**2)/(2*sigma**2)
32    )
33    AIC = -2*max_log_likelihood+2*k
34    return AIC
35
36 def SSD(y, y_fit):
37     """
38     function calculates sum of squared deviations of fitted
39     values and true values
```

```
40  """
41  return np.sum((y-y_fit)**2)
42
```

A.3 Script 1

The following script plots generates data from given true linear function by adding normally distributed error terms. The result is plotted

```
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  sigma = 0.5
12  beta = 1 # true slope
13  beta_0 = 0.0005
14  noise = np.random.normal(0, sigma, num_samples) # standard normal noise term
15
16  y = beta * X + noise + beta_0
17
18  plt.plot(
19  X, beta * X, color='red',
20  linewidth=1, label='True Linear Relationship'
21  )
22  plt.scatter(
23  X, y, color='white', edgecolor='black',
24  marker='o', label='Generated Data Points'
25  )
26  plt.legend()
```

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

A.4 Script 2

The following code performs both OLS and RTO linear regression models, plots the data, and computes the values of SSD

```
1 mean_x = np.average(X)
2 mean_y = np.average(y)
3
4 Sxx = np.sum((X-mean_x)**2)
5 Sxy = np.sum((X-mean_x)*y)
6
7
8 beta_1_hat = Sxy/Sxx
9 beta_0_hat = mean_y - beta_1_hat * mean_x
10
11 beta_hat = np.sum(X*y)/np.sum(X**2)
12
13
14 plt.scatter(X, y, color='white', edgecolor='black', marker='o')
15
16 plt.axhline(0, color='green', linewidth=1, linestyle='--')
17 plt.axvline(0, color='green', linewidth=1, linestyle='--')
18
19
20 plt.plot(X, beta_1_hat * X + beta_0_hat, color='blue',
21          linewidth=0.5, label='with intercept')
22 plt.plot(X, beta * X + beta_0, color='green',
```

```
23 linewidth=0.5, label='true model')
24 plt.plot(X, beta_hat * X, color='red', linewidth=0.5, label='no-intercept')
25
26 plt.legend()
27
28 # zoomed plot
29
30 plt.scatter(X, y, color='white', edgecolor='black', marker='o')
31
32 plt.axhline(0, color='green', linewidth=1, linestyle='--')
33 plt.axvline(0, color='green', linewidth=1, linestyle='--')
34
35 plt.plot(X, beta_1_hat * X + beta_0_hat,
36 color='blue', linewidth=0.5, label='with intercept')
37 plt.plot(X, beta * X + beta_0,
38 color='green', linewidth=0.5, label='true model')
39 plt.plot(X, beta_hat * X, color='red', linewidth=0.5, label='no-intercept')
40
41 plt.legend()
42
43 plt.xlim(-0.5, 0.5)
44 plt.ylim(-0.5, 0.5)
45
46
47 y_hat = beta_1_hat * X + beta_0_hat # intercept model response variable
48 y_hat_hat = beta_hat * X # no-intercept response variable
49
50
51 y = beta * X + beta_0
52
53 from linear_regression import (bayesian_information_criterion,
54 akaike_information_criterion)
55
56 y = y.flatten()
```

```
57 X = X.flatten()
58 Y_pred = Y_pred.flatten()
59 Y = Y_pred_no = Y_pred_no.flatten()
60
61 BIC = bayesian_information_criterion(y, Y_pred,
62 num_samples, sigma, 3)
63 BIC_no = bayesian_information_criterion(y, Y_pred_no,
64 num_samples, sigma, 2)
65
66 AIC = akaike_information_criterion(y, Y_pred,
67 num_samples, sigma, 3)
68 AIC_no = akaike_information_criterion(y, Y_pred_no,
69 num_samples, sigma, 2)
70
71 print(f"BIC of intercept model is {BIC: >22}")
72 print(f"BIC of no-intercept model is {BIC_no: >19}")
73 print(f"AIC of intercept model is {AIC: >22}")
74 print(f"AIC of no-intercept model is {AIC_no: >19}")
75
76 SS_dev_intercept = np.sum((y-y_hat)**2)
77 SS_dev_no_intercept = np.sum((y-y_hat_hat)**2)
78
79 print(f"SSD of intercept model: {SS_dev_intercept}")
80 print(f"SSD of no-intercept model: {SS_dev_no_intercept}")
81
```

A.5 Script 3

This script is used to generate fig. 4.2

```
1 delta_SSD = np.zeros(1000)
2 delta_AIC = np.zeros(1000)
3 delta_BIC = np.zeros(1000)
4 alphas = np.zeros(1000)
5
```

```
6  np.random.seed(42)
7
8  for increment in range(1000):
9      SS_dev_intercept_mean = 0
10     SS_dev_no_intercept_mean = 0
11     AIC_intercept_mean = 0
12     AIC_no_intercept_mean = 0
13     BIC_intercept_mean = 0
14     BIC_no_intercept_mean = 0
15     num_samples = 100
16     for i in range(100):
17
18         beta = 1 # true slope
19         beta_0 = 0 + increment/5000
20         noise = np.random.normal(0, 1, num_samples) # standard normal noise term
21         X = np.random.rand(num_samples) * 2 - 1
22         y = beta * X + noise + beta_0
23
24         mean_x = np.average(X)
25         mean_y = np.average(y)
26
27         Sxx = np.sum((X-mean_x)**2)
28         Sxy = np.sum((X-mean_x)*(y))
29
30
31         beta_1_hat = Sxy/Sxx
32         beta_0_hat = mean_y - beta_1_hat * mean_x
33
34         beta_hat = np.sum(X*y)/np.sum(X**2)
35
36         y_hat = beta_1_hat * X + beta_0_hat # intercept model response variable
37         y_hat_hat = beta_hat * X # no-intercept response variable
38
39         y_true = beta * X + beta_0
```

```
40
41 SS_dev_intercept = np.sum((y_true-y_hat)**2)
42 SS_dev_no_intercept = np.sum((y_true-y_hat_hat)**2)
43
44 AIC_intercept = akaike_information_criterion(y, y_hat,
45 num_samples, sigma, 3)
46 AIC_no_intercept = akaike_information_criterion(y, y_hat_hat,
47 num_samples, sigma, 2)
48
49 BIC_intercept = bayesian_information_criterion(y, y_hat,
50 num_samples, sigma, 3)
51 BIC_no_intercept = bayesian_information_criterion(y, y_hat_hat,
52 num_samples, sigma, 2)
53
54 AIC_intercept_mean += AIC_intercept
55 AIC_no_intercept_mean += AIC_no_intercept
56
57 BIC_intercept_mean += BIC_intercept
58 BIC_no_intercept_mean += BIC_no_intercept
59
60 SS_dev_intercept_mean += SS_dev_intercept
61 SS_dev_no_intercept_mean += SS_dev_no_intercept
62
63 SS_dev_intercept_mean /= num_samples
64 SS_dev_no_intercept_mean /= num_samples
65
66 AIC_intercept_mean /= num_samples
67 AIC_no_intercept_mean /= num_samples
68
69 BIC_intercept_mean /= num_samples
70 BIC_no_intercept_mean /= num_samples
71
72 delta SSD[increment] = SS_dev_no_intercept_mean - SS_dev_intercept_mean
73 delta AIC[increment] = AIC_no_intercept_mean - AIC_intercept_mean
```

```
74 delta_BIC[increment] = BIC_no_intercept_mean - BIC_intercept_mean
75 alphas[increment] = beta_0
76
77 zeros = np.zeros_like(alphas)
78
79 plt.plot(alphas, zeros, '--', linewidth=0.5)
80 plt.plot(alphas, delta_SSD, color='blue', linewidth=0.5)
81
82 plt.xlabel(r"$\beta_0$")
83 plt.ylabel(r"$\Delta$ SSD")
84 plt.show()
85
86
```
