The Least Squares Linear Regression Model

Henrique Veras

PIMES/UFPE

Introduction

Model builders are oftern interested in understanding the *conditional variation* of one variable relative to others rather than their *joint probability*

Question: What feature of the conditional probability distribution are we interested in?

Usually, the expected value $E[y|x]$, but sometimes might be:

Conditional median or other quantiles of the distribution (20th percentile, 5th percentile, etc), variance

Linear regression deals with **conditional mean**

The Linear Regression Model

$\mathbf{y} = f(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k) + \varepsilon$, where ε is called the **disturbance** term.

Our **theory** will specify the population regression equation $f(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k)$, which encompasses its format and the variables that matter.

Assumptions of the Linear Regression Model

The linear regression model consists of a set of assumptions about how a data set will be produced by an underlying ”data generating process.”

Assumption A1: The model specifies a linear relationship between y and $\mathbf{x}_1, \cdots, \mathbf{x}_k$:

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \cdots + \mathbf{x}_k\beta_k + \varepsilon$$

Notice that the assumption is about the linearity in the parameters rather than in the $\mathbf{x}$’s.

Linearity of the Regression Model

Each observation of a given data set looks like

$$y_1 = \beta_1 x_{11} + \beta_2 x_{21} + \cdots \beta_k x_{k1} + \varepsilon_1$$

$$y_2 = \beta_1 x_{12} + \beta_2 x_{22} + \cdots \beta_k x_{k2} + \varepsilon_1$$

$$\vdots$$

$$y_n = \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots \beta_k x_{kn} + \varepsilon_1$$

Linearity of the Regression Model

In Matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \ldots & X_{k1} \\ 1 & X_{12} & X_{22} & \ldots & X_{k2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{1n} & X_{2n} & \ldots & X_{kn} \end{bmatrix}_{n\times k} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_n \end{bmatrix}_{k\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

Ful Rank

Assumption A2: The columns of X are linearly independent and there are at least k observations.

Assumption A2 states that there are no linear relationships among the variables.

Here's an example of a model that cannot be estimated, although we might be interested in quantifying each of the coefficients: the determinants of Monet's prices:

$$\ln \text{Price} = \beta_1 \ln \text{Size} + \beta_2 \ln \text{Aspect Ratio} + \beta_3 \ln \text{Height} + \varepsilon$$

where Size = Width × Height and Aspect Ratio = $Width/Height$

Regression

Assumption A3: The disturbance is assumed to have conditional expected value zero at every observation: $E(\varepsilon|\mathbf{X}) = 0$

No value of **X** conveys any information about ε. We assume that ε_i's are purely random draws from a population.

Moreover, we assume $E[\varepsilon_i|\varepsilon_1, \cdots, \varepsilon_{i-1}, \varepsilon_{i+1}, \cdots, \varepsilon_n] = 0$.

Notice that by the **Law of Iterated Expectations**:

$$E[\varepsilon_i] = E_X[E[\varepsilon_i|\mathbf{X}]] = E_X[0] = 0$$

Regression

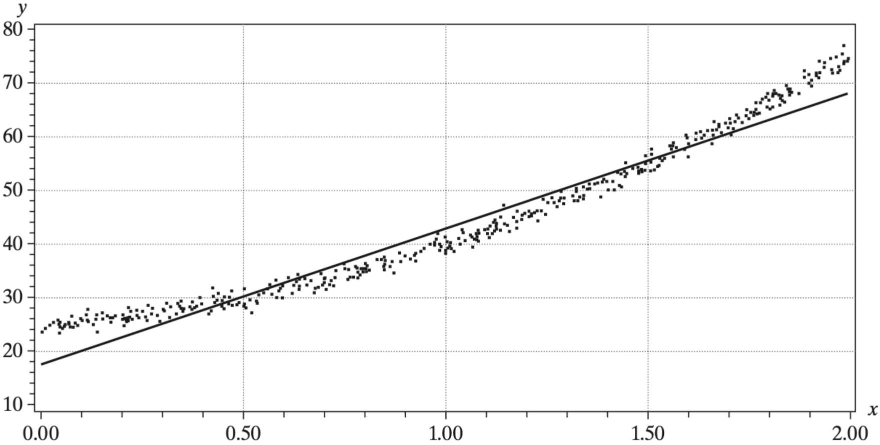
Point to note: $E[\varepsilon|\mathbf{X}] = 0 \Rightarrow Cov(\mathbf{X}, \varepsilon) = 0$. But the converse is not true: $E[\varepsilon] = 0$ **does not** imply that $E[\varepsilon|\mathbf{X}] = 0$.

Accordingly, $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta$.

Assumptions **A1** and **A3** comprise the *linear regression model.*

What if $E[\varepsilon] \neq 0$?

FIGURE 2.2 Disturbances with Nonzero Conditional Mean and Zero Unconditional Mean.



Regression

Assumption **A3** is called the **exogeneity** assumption and it yields $E[\mathbf{y}] = \mathbf{X}\beta$.

Whenever $E(\varepsilon|x) \neq 0$, we say that x is **endogenous** to the model. One way that this can happen is when we leave out a variable that matters for the relationship.

Suppose the DGP of a given relationship is given by

$$Income = \gamma_1 + \gamma_2 educ + \gamma_3 age + u$$

but we estimate the model

$$Income = \gamma_1 + \gamma_2 educ + \varepsilon$$

How do we show that **A3** is not satisfied?

Homoskedasticity and Nonautocorrelated Disturbances

Assumption A4: $E[\varepsilon\varepsilon'|\mathbf{X}] = \sigma^2\mathbf{I}$

Also, notice that $Var[\varepsilon] = E[Var(\varepsilon|\mathbf{X})] + Var[E(\varepsilon|\mathbf{X})] = \sigma^2\mathbf{I}$

Data Generating Process for the Regressors

Assumption A5: $\mathbf{X}$ may be fixed or random.

Fixed $\mathbf{X}$: Experimental designs, whereby the researcher fixes the values of $\mathbf{X}$ to find $\mathbf{y}$.

Random $\mathbf{X}$: Observational studies. However, some columns of the $\mathbf{X}$ can be fixed, such as indicator variables for a given time period or time trends.

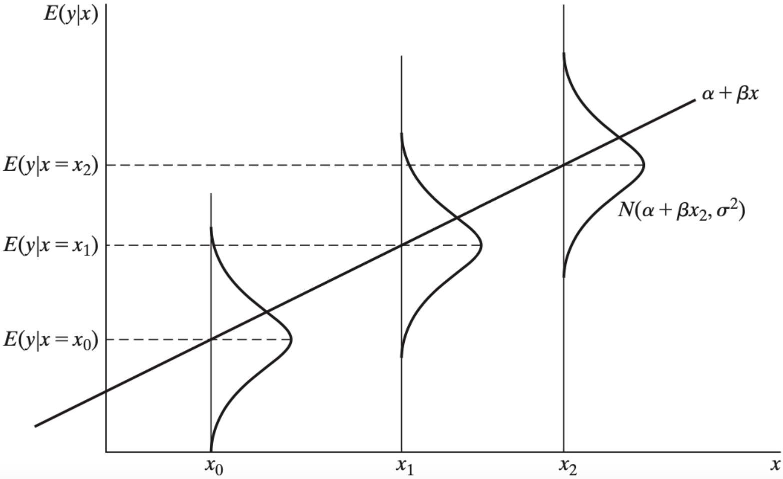
Normality

Assumption A6: $\varepsilon|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

This assumption is useful for hypothesis testing and constructing confidence intervals but might not be needed as the Central Limit Theorem applies to sufficiently large data.

Visual Summary of the Assumptions

FIGURE 2.3 The Normal Linear Regression Model.



Computational Aspects of the Least Squares Regression

Let's now consider the algebraic problem of choosing a vector $\mathbf{b}$ so that the fitted line $\mathbf{x}_i'\mathbf{b}$ is *close* to the data.

We need to specify what do we mean by *close* to the data (the fitting criterion).

Usually, the fitting criterion is the *Least Squares* method: minimizing the sum of the squared deviations from the mean.

Crucial feature: LS regression provides us a device for "holding other things constant".

The LS Population and Sample Models

Recall the population regression model: $E[y_i|\mathbf{x}_i] = \mathbf{x}_i'\beta$

We aim to find an estimate $\hat{y}_i = \mathbf{x}_i'\mathbf{b}$

Define the *residuals* from the estimated regression as

$$e_i = y_i - \mathbf{x}_i'b$$

Notice that $y_i = \mathbf{x}_i'\beta + \varepsilon_i = \mathbf{x}_i'b + e_i$

The LS Coefficient Vector

The Least Squares criterion requires us to minimize

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i' b)^2$$

In matrix terms, we minimize

$$S(\mathbf{b}) = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

Expanding, we have

$$S(\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

The LS Coefficient Vector

The necessary condition for a minimum is

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0}$$

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

From **A2**, we know that $\mathbf{X}$ has full rank, which guarantees the existence of its inverse. Then, pre-multiplying both sides by $(\mathbf{X}'\mathbf{X})^{-1}$:

$$b_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

For the solution b_0 to minimize the sum of the squared residuals, the matrix $\frac{\partial^2 S(\mathbf{b})}{\partial \mathbf{b}^2} = 2\mathbf{X}'\mathbf{X}$ must be positive definite.

Algebraic Aspects of the LS Solution

We have

$$\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y} = -\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = -\mathbf{X}'\mathbf{e} = \mathbf{0}$$

Hence, for every column of $\mathbf{X}$, $\mathbf{x}_k'\mathbf{e} = 0$.

Denote the first row $\mathbf{X}$ as $\mathbf{x}_1 \equiv \mathbf{i}$, two implications follow:

1. The LS residuals sum to zero.
2. The regression hyperplane passes through the point of means of the data.

Projection

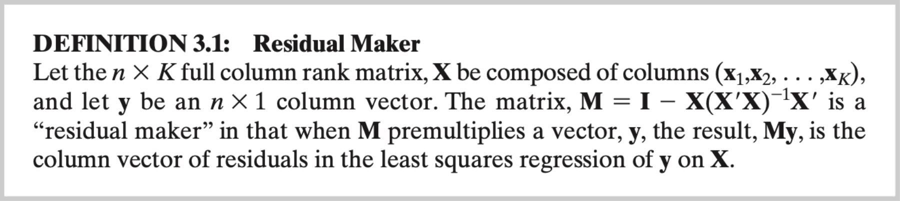
Recall the LS residuals:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

Inserting $\mathbf{b}_0$, we have

$$\mathbf{e} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{M}\mathbf{y}$$

The matrix **M** is called the "*residual maker*":



DEFINITION 3.1: Residual Maker
Let the $n \times K$ full column rank matrix, **X** be composed of columns $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K)$, and let **y** be an $n \times 1$ column vector. The matrix, $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a "residual maker" in that when **M** premultiplies a vector, **y**, the result, **My**, is the column vector of residuals in the least squares regression of **y** on **X**.

The Residual Maker

Properties of the matrix M:

1. M is symmetric ($\mathbf{M} = \mathbf{M}'$)
2. M is idempotent ($\mathbf{M} = \mathbf{M}^2$)
3. $\mathbf{MX} = \mathbf{0}$ (why?)

The Projection Matrix

Now let

$$\hat{\mathbf{y}} = \mathbf{y} - \mathbf{e} = \mathbf{I}\mathbf{y} - \mathbf{M}\mathbf{y} = (\mathbf{I} - \mathbf{M})\mathbf{y}$$

Thus,

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}$$

P is called a *projection* matrix: If a vector **y** is pre-multiplied by **P**, the result is the fitted values in the LS regression of **y** on **X**.

The Projection Matrix

Properties of $\mathbf{P}$:

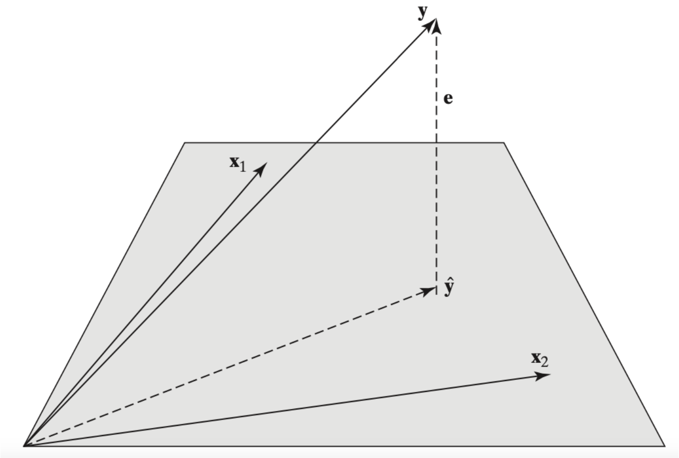
1. $\mathbf{P}$ is symmetric
2. $\mathbf{P}$ is idempotent
3. $\mathbf{PX} = \mathbf{X}$

Moreover, notice that $\mathbf{P}$ and $\mathbf{M}$ are orthogonal: $\mathbf{PM} = \mathbf{MP} = \mathbf{0}$

Therefore, the LS regression partitions the vector $\mathbf{y}$ into two **orthogonal** parts:

$$\mathbf{y} = \mathbf{Py} + \mathbf{My} = \text{Projection} + \text{Residuals}$$

FIGURE 3.2 Projection of **y** into the Column Space of **X.**



Partitioning and Partial Regressions

In some situations, we are only interested in a subset of the full set of variables in $\mathbf{X}$. The remaining variables are added to the model as “controls”.

Recall the returns to education example.

Suppose we have

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon = \mathbf{X}_1\beta_{\mathbf{1}} + \mathbf{X}_2\beta_2 + \varepsilon$$

How can we find the algebraic solution for $\mathbf{b_2}$? That is, what is the LS estimator of a given subset of parameters in β?

Partial Regressions

Set up the **normal** equations:

$$\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{b}_1 \\ \mathbf{b}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y}\end{bmatrix}$$

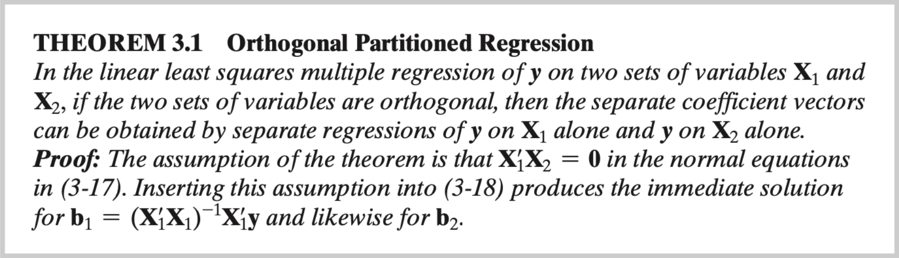
Solving the system above for $\mathbf{b}_1$ yields

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')(\mathbf{y} - \mathbf{X}_2\mathbf{b}_2)$$

Partial Regressions

Suppose that $\mathbf{X}_1'\mathbf{X}_2 = 0$. (what does this mean?)

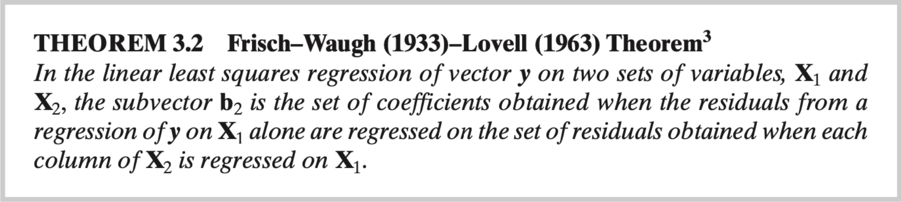
For this special case, the theorem below states that $\mathbf{b}_1$ can be obtained by regressing $\mathbf{y}$ on $\mathbf{X}_1$ only. Likewise, $\mathbf{b}_2$ can be obtained by regressing $\mathbf{y}$ on $\mathbf{X}_2$ only.



THEOREM 3.1 Orthogonal Partitioned Regression
In the linear least squares multiple regression of $\mathbf{y}$ *on two sets of variables* $\mathbf{X}_1$ *and* $\mathbf{X}_2$*, if the two sets of variables are orthogonal, then the separate coefficient vectors can be obtained by separate regressions of* $\mathbf{y}$ *on* $\mathbf{X}_1$ *alone and* $\mathbf{y}$ *on* $\mathbf{X}_2$ *alone.*
Proof: *The assumption of the theorem is that* $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ *in the normal equations in (3-17). Inserting this assumption into (3-18) produces the immediate solution for* $\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$ *and likewise for* $\mathbf{b}_2$.

The FWL Theorem

For the general case, in which $\mathbf{X}_1$ and $\mathbf{X}_2$ might not be orthogonal, the following theorem provides the more general solution:



THEOREM 3.2 Frisch–Waugh (1933)–Lovell (1963) Theorem[3]
In the linear least squares regression of vector $\mathbf{y}$ on two sets of variables, $\mathbf{X}_1$ and $\mathbf{X}_2$, the subvector $\mathbf{b}_2$ is the set of coefficients obtained when the residuals from a regression of $\mathbf{y}$ on $\mathbf{X}_1$ alone are regressed on the set of residuals obtained when each column of $\mathbf{X}_2$ is regressed on $\mathbf{X}_1$.

The FWL Theorem

We can represent $\mathbf{b}_2$ as

$$\mathbf{b}_2 = (\mathbf{X}_2^{*'} \mathbf{X}_2^*)^{-1} \mathbf{X}_2^{*'} \mathbf{y}^*$$

where $\mathbf{X}_2^* = \mathbf{M}_1 \mathbf{X}_2$ and $\mathbf{y}^* = \mathbf{M}_1 \mathbf{y}$.

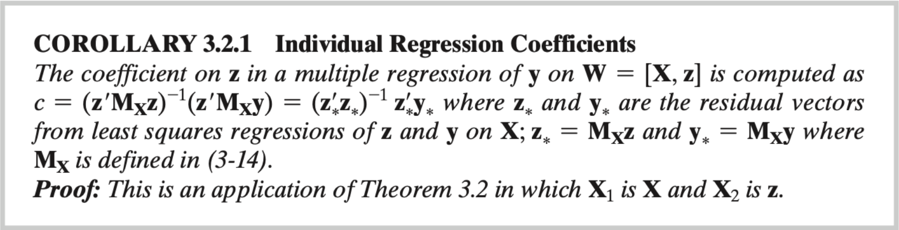
Two questions:

1. What is $\mathbf{M}_1 \mathbf{X}_2$?
2. What is $\mathbf{M}_1 \mathbf{y}$?

The FWL Theorem

A special case of the FWL theorem is when we are interested in the computation of a single coefficient.

Consider the regression of $\mathbf{y}$ on a set of variables $\mathbf{X}$ and an additional variable z. Denote the coefficients $\mathbf{b}$ and c, respectively.



COROLLARY 3.2.1 Individual Regression Coefficients
The coefficient on $\mathbf{z}$ *in a multiple regression of* $\mathbf{y}$ *on* $\mathbf{W} = [\mathbf{X}, \mathbf{z}]$ *is computed as* $c = (\mathbf{z}'\mathbf{M_X}\mathbf{z})^{-1}(\mathbf{z}'\mathbf{M_X}\mathbf{y}) = (\mathbf{z}_*'\mathbf{z}_*)^{-1}\,\mathbf{z}_*'\mathbf{y}_*$ *where* $\mathbf{z}_*$ *and* $\mathbf{y}_*$ *are the residual vectors from least squares regressions of* $\mathbf{z}$ *and* $\mathbf{y}$ *on* $\mathbf{X}$; $\mathbf{z}_* = \mathbf{M_X}\mathbf{z}$ *and* $\mathbf{y}_* = \mathbf{M_X}\mathbf{y}$ *where* $\mathbf{M_X}$ *is defined in (3-14).*
Proof: *This is an application of Theorem 3.2 in which* $\mathbf{X}_1$ *is* $\mathbf{X}$ *and* $\mathbf{X}_2$ *is* $\mathbf{z}$.

The FWL Theorem

Example: Suppose we are interested again in the returns to education equation

$$Income = \beta_1 + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \varepsilon$$

To find b_1:

1. Regress $Income$ on age and age^2 and obtain residuals r_1
2. Regress $educ$ on age and age^2 and obtain residuals r_2
3. Regress r_1 on r_2 and find slope coefficient b_1.

Regression with a constant term

Consider now the partition in which $\mathbf{X}_1 = \mathbf{i}$ and $\mathbf{X}_2$ is the set of variables in the regression.

Take a given column $\mathbf{x}$ of $\mathbf{X}_2$. According to the FWL theorem,

$$\mathbf{x}^* = \mathbf{M}_1\mathbf{x}$$

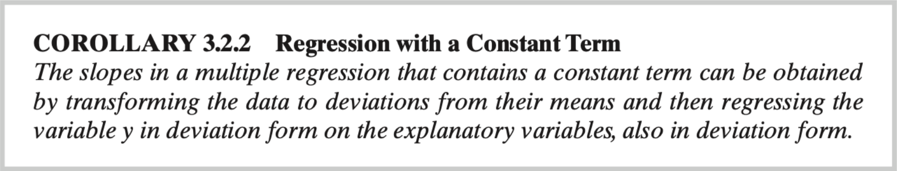
This yields

$$\mathbf{x}^* = \mathbf{x} - \mathbf{i}\bar{\mathbf{x}}$$

Regression with a constant term

The result above says that the residuals in the regression of the columns of $\mathbf{X}_2$ on a constant term are deviations from the sample mean.

Therefore, each column of $\mathbf{M}_1\mathbf{X}_2$ is the original variable, now in the form of deviations from the mean. This general result is summarized in the following corollary.



COROLLARY 3.2.2 Regression with a Constant Term
The slopes in a multiple regression that contains a constant term can be obtained by transforming the data to deviations from their means and then regressing the variable y in deviation form on the explanatory variables, also in deviation form.

Table of Contents

Econometrics

- Intro
- The Linear Regression Model
- Assumptions of the Linear Regression Model
- Computation of the Least Squares Regression
- Partitioning and Partial Regressions
- The FWL Theorem