ECONOMETRICS

Introductory Notes for Machine Learning

Michael Hoon

January 29, 2024

Contents

1 Introduction

These notes are heavily inspired by "Introductory Econometrics: A Modern Approach" by Jeffrey Wooldridge, and UC Berkeley's ENVECON C118 course.

2 Linear Models

$$y = \theta_0 + \theta_1 x_1 + \epsilon \tag{1}$$

The variable ϵ is called the error term in the relationship, representing factors other than x that affect y.

2.1 Unbiased Estimators

An estimator, $\hat{\theta}$ of θ is an **unbiased estimator** if

$$\mathbb{E}(\hat{\theta}) = \theta \tag{2}$$

for all possible values of θ. This means that its probability distribution has an expected value equal to the parameter it is supposed to be estimating.

2.2 Efficient Estimators

Definition 2.1. The Cramér-Rao bound states that the precision of any unbiased estimator is at most the Fisher Information. Equivalently, the reciprocal of the Fisher Information is a lower bound on its variance.

An unbiased estimator that achieves the **Cramér-Rao bound (CRB)** is said to be **fully efficient**. Such a solution achieves the lowest possible mean squared error among all unbiased methods, and is therefore the minimum variance unbiased estimator (MVUE).

In statistics, efficiency is a measure of quality of an estimator, of an experimental design, or of a hypothesis testing procedure. Essentially, a more efficient estimator needs fewer input data or observations than a less efficient one to achieve the Cramér-Rao bound.

In the scalar unbiased case, suppose θ is an unknown deterministic parameter that is to be estimated from n independent observations of x, each from a distribution according to some probability density function $f(x|\theta)$. The variance of any unbiased estimator $\hat{\theta}$ of θ is then bounded by the reciprocal of the Fisher Information $I(\theta)$:

$$\text{var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

Definition 2.2. The Fisher Information is defined by

$$I(\theta) = n\mathbb{E}(X|\theta)\left[\left(\frac{\partial \ln[f(x|\theta)]}{\partial \theta}\right)^2\right]$$

The efficiency of an unbiased estimator $\hat{\theta}$ measures how close this estimator's variance comes to this lower bound. Estimator efficiency is defined as

$$e(\hat{\theta}) = \frac{I(\theta)^{-1}}{\text{var}(\hat{\theta})}$$

or the **minimum possible variance** for an unbiased estimator divided by its **actual variance**. The Cramér-Rao lower bound thus gives

$$e(\hat{\theta}) \leq 1$$

3 Ordinary Least Squares

3.1 OLS Scalar Form Derivation

Let $\{(x_i, y_i) : i = 1, \ldots, n\}$ denote a random sample of size n from the population. We start by defining the simple linear regression model:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{3}$$

where ϵ_i denotes the error term for observation i, which can be calculated by $\epsilon_i = y_i - \hat{y}_i$. To derive OLS estimators for $\hat{\alpha}$ and $\hat{\beta}$, we seek to minimise the sum of squared residuals (errors) (SSE). We can then treat this as a linear optimization problem:

$$\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{N} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \tag{4}$$

Taking the partial derivative of equation (4) with respect to $\hat{\alpha}$ and $\hat{\beta}$, and setting them both to 0 yields

$$\frac{\partial(\text{SSE})}{\partial \hat{\alpha}} = \sum_{i=1}^{N} -2(y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \tag{5}$$

$$\frac{\partial(\text{SSE})}{\partial \hat{\beta}} = \sum_{i=1}^{N} -2x_i(y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \tag{6}$$

From equation (5), we can remove the -2 and simplify the summation terms where $\frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}$. From here, we obtain

$$N\hat{\alpha} = N\bar{y} - N\hat{\beta}\bar{x} \tag{7}$$

and dividing by N gives us:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{8}$$

Now, solving for $\hat{\beta}$. From equation (6), eliminating the -2 and distributing the x_i term element-wise gives us:

$$\sum_{i=1}^{N} \left(y_i x_i - \hat{\alpha} x_i - \hat{\beta} x_i^2 \right) = 0 \tag{9}$$

Substituting equation (8) into equation (9) gives us:

$$\sum_{i=1}^{N} \left(y_i x_i - (\bar{y} - \hat{\beta}\bar{x}) x_i - \hat{\beta} x_i^2 \right) = 0 \tag{10}$$

We can distribute the sum to each term:

$$\sum_{i=1}^{N} y_i x_i - \bar{y} \sum_{i=1}^{N} x_i + \hat{\beta}\bar{x} \sum_{i=1}^{N} x_i - \hat{\beta} \sum_{i=1}^{N} x_i^2 = 0 \tag{11}$$

rearranging and algebraically solving for $\hat{\beta}$ gives us:

$$\hat{\beta} = \frac{\sum_{i=1}^{N} y_i x_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2} \tag{12}$$

Using the properties of the summation operator from equation (43) and (44) in Appendix: Properties for Scalar Form OLS Derivation, we substitute this into equation (12) to obtain the final expression for $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{13}$$

and substituting (13) into (8), $\hat{\alpha}$ can now be formally expressed as:

$$\hat{\alpha} = \bar{y} - \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\bar{x} \tag{14}$$

3.2 OLS Normal Equation Derivation

Let $\mathbf{X}$ be a $n \times (k+1)$ matrix with observations on $(k+1)$ independent variables for n observations. Since our model usually contains a constant term, one of the columns in the $\mathbf{X}$ matrix will contain only ones. We define the following:

- Let $\mathbf{y}$ be an $n \times 1$ vector of observations on the dependent variable.
- Let $\boldsymbol{\epsilon}$ be an $n \times 1$ vector of errors.
- Let $\boldsymbol{\beta}$ be a $(k+1) \times 1$ vector of unknown population parameters that we want to estimate.

Our equation in matrix form will be the following:

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

which we can rewrite as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{15}$$

To obtain the estimates of the population parameter $\hat{\boldsymbol{\beta}}$, we must minimize the sum of squared errors (SSE). The vector of residuals is given by:

$$\boldsymbol{\epsilon} = \mathbf{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{16}$$

and the SSE is given by:

$$\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 & \epsilon_2 \dots \epsilon_n \end{bmatrix}_{1\times n} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1} \tag{17}$$

a little matrix algebra using the transpose property $(AB)^\top = B^\top A^\top$ gives us:

$$\begin{aligned} \boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}^\top\boldsymbol{y} - \hat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{y} - \boldsymbol{y}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{y}^\top\boldsymbol{y} - \mathbf{2}\hat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{y} + \hat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} \end{aligned} \tag{18}$$

To minimise the SSE, we take the derivative of equation **??** with respect to $\hat{\boldsymbol{\beta}}$.

$$\frac{\partial(\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon})}{\partial\hat{\boldsymbol{\beta}}} = -2\boldsymbol{X}^\top\boldsymbol{y} + 2\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} = 0 \tag{19}$$

Taking the second derivative with respect to $\hat{\boldsymbol{\beta}}$, we get $2\boldsymbol{X}^\top\boldsymbol{X}$, where if $\boldsymbol{X}$ has full rank, then the second derivative has a positive value and $\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}$ is indeed a minimum.

From equation **??**, manipulating the terms gives us:

$$(\boldsymbol{X}^\top\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\mathbf{y} \tag{20}$$

If $(\boldsymbol{X}^\top\boldsymbol{X})$ is of full rank (no multicollinearity), then its inverse exists and we can pre-multiply both sides by the inverse:

$$(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{X})\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y} \tag{21}$$

By definition, $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}(\boldsymbol{X}^\top\boldsymbol{X}) = I$ which is the identity matrix, so

$$\begin{aligned} I\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y} \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\mathbf{y} \end{aligned} \tag{22}$$

3.3 Orthogonal Projection

3.4 Derivation of R^2 in Matrix Form

Recall that the formula for R^2 Coefficient of Determination is:

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{23}$$

where SSR is the Sum of Squares due to Regression, and the SST is the Sum of Squares Total, and SSE is the Sum of Squared Errors. The expressions are listed below, with matrix notation on the right hand side.

$$\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = (\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^{\mathbf{2}} = (\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^\top(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}}) \tag{24}$$

$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = (\boldsymbol{y} - \bar{\boldsymbol{y}})^{\mathbf{2}} = (\boldsymbol{y} - \bar{\boldsymbol{y}})^\top(\boldsymbol{y} - \bar{\boldsymbol{y}}) \tag{25}$$

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 = \boldsymbol{\epsilon}^{\mathbf{2}} = \boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} \tag{26}$$

Considering that the assumptions of the Classical Linear Regression Model holds, we have mean centering where $\bar{\boldsymbol{y}} = 0$. This simplifies the quantity $\boldsymbol{y} - \bar{\boldsymbol{y}} = \boldsymbol{y} - \mathbf{0} = \boldsymbol{y}$, and gives us the following results:

$$\text{SSR} = \hat{\boldsymbol{y}}^\top\hat{\boldsymbol{y}} \tag{27}$$

$$\text{SST} = \boldsymbol{y}^\top\boldsymbol{y} \tag{28}$$

$$\text{SSE} = \boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} \tag{29}$$

With equations (27) and (28), we can express the R^2 coefficient of determination in matrix form as:

$$R^2 = \frac{\hat{\boldsymbol{y}}^\top\hat{\boldsymbol{y}}}{\boldsymbol{y}^\top\boldsymbol{y}} \tag{30}$$

To find the SSR expression $\hat{\boldsymbol{y}}^\top\hat{\boldsymbol{y}}$, consider the given equation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. To calculate the SST expressed as $\boldsymbol{y}^\top\boldsymbol{y}$ in equation (28):

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{\epsilon} \\
\boldsymbol{y}^\top \boldsymbol{y} &= (\boldsymbol{XB} + \boldsymbol{\epsilon})^\top (\boldsymbol{XB} + \boldsymbol{\epsilon}) \\
&= (\boldsymbol{B}^\top \boldsymbol{X}^\top + \boldsymbol{\epsilon}^\top)(\boldsymbol{XB} + \boldsymbol{\epsilon}) \\
&= (\boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{XB} + \boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\top \boldsymbol{XB} + \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}) \\
&= \boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{XB} + \boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{\epsilon} + (\boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{\epsilon})^\top + \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}
\end{aligned} \tag{31}
$$

From (31), we see that for the term $\boldsymbol{X}^\top \boldsymbol{\epsilon}$, the residual $\boldsymbol{\epsilon}$ is **orthogonal to all columns of** $\boldsymbol{X}$. This can be seen from:

$$
\begin{aligned}
\boldsymbol{X}^\top \boldsymbol{\epsilon} &= \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{X} \hat{\boldsymbol{\beta}} \\
&= \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\
&= \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{y} \\
&= \boldsymbol{0}
\end{aligned} \tag{32}
$$

and also geometrically from Figure 1, where the vector ϵ is orthogonal to the component of $\boldsymbol{y}$, being $\boldsymbol{X}\hat{\boldsymbol{\beta}}$. The dot product $(\boldsymbol{X}^\top \boldsymbol{\epsilon})$ of two orthogonal vectors is 0.

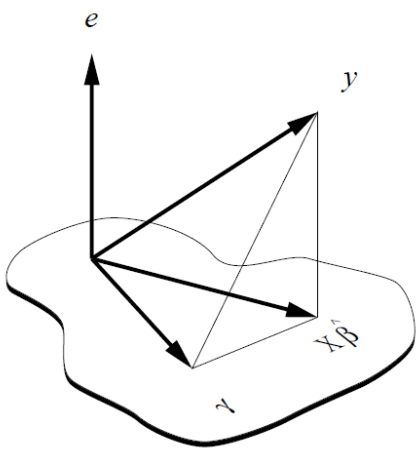

Figure 1: Orthogonal Decomposition of the Sum of Squares

Given the result from equation (32), we plug back into equation (31) to obtain the expression for SST:

$$
\begin{aligned}
\boldsymbol{y}^\top \boldsymbol{y} &= \boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{XB} + \boldsymbol{0} + \boldsymbol{0} + \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} \\
&= \boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{XB} + \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}
\end{aligned} \tag{33}
$$

Now, we know that SST = SSR + SSE, we can rearrange to find: SSR = SST − SSE. Expressing this in matrix form, we have:

$$
\begin{aligned}
\hat{\boldsymbol{y}}^\top \hat{\boldsymbol{y}} &= \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} \\
&= \boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{XB} + \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} \\
&= \boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{XB}
\end{aligned} \tag{34}
$$

Finally, plugging equation (34) back into equation (30), we get the expression for R^2 coefficient of determination

in matrix form:

$$R^2 = \frac{\boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{B}}{\boldsymbol{y}^\top \boldsymbol{y}} \tag{35}$$

3.5 Derivation of OLS Error term ϵ

From equation 22, we can substitute the expression back into equation 16, and factoring out the y vector we obtain:

$$\begin{aligned} \epsilon &= \mathrm{y} - \boldsymbol{X}\boldsymbol{\beta} \\ &= \mathrm{y} - \boldsymbol{X}[(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathrm{y}] \\ &= \left(I - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top\right) \mathrm{y} = M\mathrm{y} \end{aligned} \tag{36}$$

where I is the identity matrix, and M is the residual maker matrix.

3.6 The Gauss-Markov Assumptions

4 Maximum Likelihood Estimation

There are many methods of constructing an estimator for an unknown parameter θ, such as the Generalised Method of Moments (GMM), Maximum Likelihood Estimation (MLE), etc. In particular, MLE has a strong intuitive appeal, and often yields reasonable estimates of θ. If the **sample is large**, it will yield an excellent estimator of θ, and thus is the most widely use method of estimation in statistics.

Suppose that the random variables $X_1, \ldots, X_n$ form a random sample from a distribution $f(x|\theta)$; if X is a **continuous random variable**, $f(x|\theta)$ is a pdf, and if X is **discrete**, then $f(x|\theta)$ is a point mass function. The parameter θ could be a real-valued **unknown** parameter of a vector of parameters.

Definition 4.1. For every observed random sample $x_1, \ldots, x_n$ we define:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) = f(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \ldots f(x_n|\theta) \tag{37}$$

where $L(\theta)$ is the **likelihood function**.

The meaning of maximum likelihood is as follows: we choose the parameter that maximises the likelihood of obtaining the data at hand. For discrete distributions, the likelihood is the same as the probability.

If the maximum of L occurs at a point $\hat{\theta}$, then $\hat{\theta}$ is the parameter value for which the observed data is most likely to have been generated, so we use it for our estimate for θ. We call $\hat{\theta}$ the MLE for θ.

To simplify the product terms, we can equivalently maximise the log of the likelihood function:

$$\ell = \ln[L(\theta)] = \ln\left[\prod_{i=1}^{n} f(X_i|\theta)\right] = \sum_{i=1}^{n} \ln[f(X_i|\theta)] \tag{38}$$

4.1 Preliminary Intuition for MLE

Suppose we have $y_i \sim N(\mu, \sigma^2)$, and thus $\mathbb{E}[y] = \mu$, $\mathrm{var}(y) = \sigma^2$. In general, we have some observations on Y and we want to estimate the parameters μ, σ^2 from the data. The idea of maximum likelihood however, is to

find the estimate of the parameters that maximises the probability of observing the data that we have. This problem that we face is now the opposite of the typical probability problem: we have the data, but want to learn about the model (specifically the parameters).

Mathematically, in a traditional probability problem we would like to know $P(Data|Model)$, i.e. $P(y|\theta)$, but now we want to know $P(Model|Data)$, i.e. $P(\theta|y)$ - This is the inverse probability problem.

4.1.1 Bayes' Theorem

Recall that

$$P(\theta|y) = \frac{P(\theta)P(y|\theta)}{P(y)} \tag{39}$$

where $P(y)$ is just a function of the data, and we can ignore it. Thus,

$$P(\theta|y) \propto P(\theta)P(y|\theta) \tag{40}$$

$P(\theta)$ is the prior density of θ, $P(y|\theta)$ is the likelihood, and $P(\theta|y)$ is the posterior density of θ. We can see that the likelihood is the sample information that transforms a 'prior' into a 'posterior' density of θ.

Note that the prior, θ, is fixed before our observations and so can be treated as invariant to our problem. We can then rewrite 40 to

$$P(\theta|y) = k(y)P(y|\theta) \tag{41}$$

where $k(y) = \frac{P(\theta)}{P(y)}$ is an unknown function of the data. Since $k(y)$ is not a function of θ, it is treated as unknown positive constant (for any given data, $k(y)$ remains same over all hypothetical values of θ).

4.1.2 Likelihood

Definition 4.2. Fisher defined the notion of likelihood and the Likelihood Axiom:

$$\begin{aligned}\ell(\theta|y) &= k(y)P(y|\theta)\\ &\propto P(y|\theta)\end{aligned}$$

The best estimator, $\hat{\theta}$ is whatever value of $\hat{\theta}$ that **maximises**

$$\ell(\theta|y) = P(y|\theta)$$

We are looking for the $\hat{\theta}$ that maximises the likelihood of observing our sample. Because of proportionality, the $\hat{\theta}$ that maximises $\ell(\theta|y)$ will also maximise $P(y|\theta)$, i.e. probability of observing the data.

4.2 Analytical Method

If the y_i are all independent, then the likelihood of the whole sample is the product of the individual likelihoods over all the observations.

$$
\begin{aligned}
\ell &= \prod_{i=1}^{n} \ell_i \\
&= \prod_{i=1}^{n} P(y_i|\hat{\theta}) \\
\therefore \ln[\ell] &= \sum_{i=1}^{n} \ln[P(y_i|\hat{\theta})]
\end{aligned}
$$

Now to find $\hat{\theta}$, we first differentiate the likelihood function with respect to the parameter vector and set the resulting gradient vector to zero. Solve the system of equations to find the extrema. Then, take the second derivative to ensure maxima.

Appendix A Properties for Scalar Form OLS Derivation

To derive the two properties used to obtain equation (13), we use three basic properties of the summation operator:

$$\begin{aligned}\sum_{i=1}^{N}(x_i-\bar{x}) &= \sum_{i=1}^{N}x_i-\sum_{i=1}^{N}\bar{x} \\ &= \sum_{i=1}^{N}x_i-N\bar{x} \\ &= \sum_{i=1}^{N}x_i-N\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right) \\ &= \sum_{i=1}^{N}x_i-\sum_{i=1}^{N}x_i \\ &= 0\end{aligned} \tag{42}$$

$$\begin{aligned}\sum_{i=1}^{N}(x_i-\bar{x})^2 &= \sum_{i=1}^{N}(x_i^2-2x_i\bar{x}+\bar{x}^2) \\ &= \sum_{i=1}^{N}x_i^2-2\bar{x}\sum_{i=1}^{N}x_i+N\bar{x}^2 \\ &= \sum_{i=1}^{N}x_i^2-2N\bar{x}^2+N\bar{x}^2 \\ &= \sum_{i=1}^{N}x_i^2-N\bar{x}^2\end{aligned} \tag{43}$$

It can also be shown by using property (42) that:

$$\begin{aligned}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{N}(x_iy_i-x_i\bar{y}-\bar{x}y_i+\bar{x}\bar{y}) \\ &= \sum_{i=1}^{N}x_i(y_i-\bar{y})-\sum_{i=1}^{N}\bar{x}(y_i-\bar{y}) \\ &= \sum_{i=1}^{N}x_i(y_i-\bar{y})-0 \\ &= \sum_{i=1}^{N}x_i(y_i-\bar{y}) \\ &= \sum_{i=1}^{N}(x_i-\bar{x})y_i \\ &= \sum_{i=1}^{N}x_iy_i-N\bar{x}\bar{y}\end{aligned} \tag{44}$$

which is a generalization of property (43).