3 Least Squares Estimation of the Linear Regression Model

Most work in applied econometrics is interested in relating a RV Y to a RV X, e.g. income to years of schooling, or inflation to unemployment. Y can always be decomposed as $Y = \mathbb{E}[Y|X] + U$ with $\mathbb{E}[U|X] = 0$, i.e. a part that is "explained" by X and an error U that is unrelated to X.[1] The linear regression model supposes that the relationship between Y and X is linear: $\mathbb{E}[Y|X,\theta] = X\theta$.[2]

This chapter discusses Least Squares (LS) estimation of the linear regression model. Section 3.1 presents the mechanics behind and finite sample properties of LS estimation, before Section 3.2 analyzes asymptotic properties and Section 3.3 illustrates hypothesis tesing. Throughout these sections, ideal conditions are assumed. Possible violations thereof are treated in Section 3.4. Other estimation methods of the linear regression model as well as departures from linearity are discussed in Chapters 4 to 6.

While the first two chapters distinguished RVs from their realizations using upper- and lower-case letters, the present and following chapters will use letter cases in various ways to help distinguish vectors and matrices. Also, unless otherwise stated, we treat θ as a fixed parameter and condition all moments and distributions on it, i.e. the frequentist paradigm applies.

[1] In fact, under a quadratic loss function, the conditional expectation function $\mathbb{E}[Y|X]$ is the best (unrestricted) predictor of Y:

$$\mathbb{E}[Y|X] = \arg\min_{f(X)} \mathbb{E}[(Y - f(X))^2] \ ,$$

where $f(X)$ is any (i.e. potentially nonlinear) function of X.

[2] As discussed in Section 3.4, the assumption embodied in writing out the conditional expectation in this way – the conditional indepenence assumption – can be relaxed, preserving most good properties of the LS estimator in the linear regression model. That is, the linear regression model only needs to assume $Y = X\theta + U$ with $\mathbb{E}[XU] = 0$ rather than with the stronger assumption $\mathbb{E}[U|X] = \mathbb{E}[U] = 0$. The statement in the text is useful for pedagogical purposes.

3.1 Mechanics & Finite Sample Properties

Suppose a scalar y_i is related linearly to a k-dimensional vector x_i:

$$y_i = x_i'\beta + u_i \ ,$$

where β is a k-dimensional vector of parameters. y_i is called the regressand, outcome variable or dependent variable, x_i the vector of covariates, regressors, independent variables or explanatory variables, and u_i is the error term. Throughout this and several following chapters (unless specified otherwise), we assume that we have n independent observations available.

Assumption 1 (Independent Sampling)**.** *Observations $\{z_i\}_{i=1:n}$, with $z_i = \{y_i, x_i\}$ are independent across i (i.e. they are realizations of independent RVs).*

In matrix notation, we have

$$Y = X\beta + U \ ,$$

where

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{(n\times 1)} \ , \quad X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}_{(n\times k)} \quad \text{and} \quad U = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}_{(n\times 1)} \ .$$

Usually, the vector x_i is composed of a one as the first element along with actual explanatory variables $\tilde{x}_i$: $x_i = (1, \tilde{x}_i')'$. In this case, we say the regression includes an intercept, as we can write it as

$$y_i = x_i'\beta + u_i = \beta_0 + \tilde{x}_i'\tilde{\beta} + u_i \ ,$$

where we separated out the first element of $\beta = (\beta_0, \tilde{\beta}')'$. Note that this leads to the first column of X being all ones.

Assumption 2 (Conditional Independence)**.** *Regressors x_i and errors u_i are independent: $\mathbb{E}[u_i|x_i] = \mathbb{E}[u_i] = 0$ for $u_i = y_i - x_i'\beta$.*

The crucial part of this assumption lies in equating $\mathbb{E}[u_i|x_i] = \mathbb{E}[u_i]$. $\mathbb{E}[u_i] = 0$ is guaranteed as long as an intercept is included or demeaned variables are used (see Proposition 19 below). Importantly, $\mathbb{E}[u_i|x_i] = 0$ implies $\mathbb{E}[y_i|x_i] = x_i'\beta$, or $\mathbb{E}[Y|X] = X\beta$ in matrix notation. Hence, the only real difference to the running example in Chapter 2 is that the supposed expectation of our data y_i is not simply given by a parameter, θ, but by a parameter β multiplied by some regressors x_i.[3]

Assumption 3. *The matrix $X'X = \sum_{i=1}^n x_i x_i'$ is of full rank.*

The ordinary least squares (OLS) estimator minimizes the sum of squared errors $u_i = y_i - x_i'\beta$:

$$\hat{\beta}_{OLS} = \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n u_i^2 = \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n U'U = \arg\min_{\beta \in \mathbb{R}^k} (Y - X\beta)'(Y - X\beta) \; .$$

Under Assumption 3, $X'X$ is invertible and we can solve the first order condition (FOC) $X'(Y - X\beta) = 0$ for β to get

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y \; .$$

The Appendix uses linear algebra without matrix notation to derive $\hat{\beta}$ for the case of a single regressor x_i along with an intercept. These derivations are not applicable to more general cases.

We get the predicted values $\hat{Y}$ and residuals $\hat{U} = Y - \hat{Y}$:

$$\hat{Y} \equiv X\hat{\beta} = X(X'X)^{-1}X'Y \equiv P_X Y \; , \quad \hat{U} \equiv Y - \hat{Y} = (I - P_X)Y \equiv M_X Y \; ,$$

where $P_X = X(X'X)^{-1}X'$ and $M_X = I - P_X$ are projection matrices with two important properties. First, they are idempotent: $P_X = P_X'$ and $P_X P_X = P_X$, and the same for M_X. Second, they are orthogonal to each other: $M_X P_X = (I - P_X)P_X = P_X - P_X P_X = 0$. In mathematical terms, P_X projects onto the span of X, M_X on the orthogonal complement of the span of X, i.e. it projects onto the span of X and computes the residual: $P_X X = X$ and $M_X X = 0$. A linear regression is also called a linear projection of Y on X.

The total sum of squares (SST) is given by $\sum_{i=1}^n y_i^2 = Y'Y$. It measures the variability in y_i across observations i. We can decompose it into the explained sum of squares (SSE) $\hat{Y}'\hat{Y}$ and the residual sum of squares (SSR) $\hat{U}'\hat{U}$:

$$Y'Y = (P_X Y + M_X Y)'(P_X Y + M_X Y) = Y'P_X P_X Y + Y'M_X M_X Y = \hat{Y}'\hat{Y} + \hat{U}'\hat{U} \; .$$

Based on that, we get the R^2-statistic as a measure of how well X accounts for the variation in Y in the linear regression model:

$$R^2 = \frac{\hat{Y}'\hat{Y}}{Y'Y} = 1 - \frac{\hat{U}'\hat{U}}{Y'Y} \in [0, 1] \; .$$

[3] Besides that, we let β be a k-dimensional vector, whereas θ in Chapter 2 is a scalar. However, this has no conceptual implications; it only complicates the math (slightly) by requiring us to use matrix algebra.

Proposition 19 (Frisch-Waugh-Lovell (FWL) Theorem).
Consider the linear regression of Y on two sets of covariates, X_1 and X_2:

$$Y = X_1\beta_1 + X_2\beta_2 + U \; .$$

Take $P_1 = X_1(X_1'X_1)^{-1}X_1'$ and $M_1 = I - P_1$ to write

$$M_1 Y = M_1 X_2 b + V \; .$$

Then $\hat{b}_{OLS} = \hat{\beta}_{2,OLS}$ and $\hat{U} = \hat{V}$.

See proof in Appendix. The FWL theorem says the following. Suppose we are interested in the OLS estimator of β_2 in the above regression of Y on X_1 and X_2. This very same OLS estimator can be obtained by i) first regressing Y only on X_1 and also regressing X_2 on X_1, and then ii) regressing the residuals from the first regression, M_1Y, on the residuals from the second regression, M_1X_2. In short, the estimator of β_2 in a regression of Y on X_1 and X_2 can also be obtained in a regression of Y only on X_2 after X_1 is "partialled-out".

The FWL theorem establishes that regressions – disregarding any assumptions except Assumption 3 that ensures the existence of OLS estimators – measure the partial correlation between two variables. Concretely, in the above case, $\hat{\beta}_{2,OLS}$ measures the correlation between Y and X_2 after removing the correlation between Y and X_1 and between X_2 and X_1, respectively. As such, a linear regression can tell you, for example, whether a (significant) income difference between two racial groups persists even after educational differences between those groups are taken into account. The FWL theorem also implies that regressing Y on X_2 with an intercept gives the same results as regressing the demeaned Y on the demeaned X_2 without including an intercept (take X_1 to be a vector of ones, implying that β_1 is the intercept). This case is explicitly worked out in the Appendix. [4]

So far, we only used Assumption 3, required to derive $\hat{\beta}_{OLS}$. Under Assumption 2, $\hat{\beta}_{OLS}$ is unbiased conditionally on X, i.e.

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{OLS}|X] &= \mathbb{E}[(X'X)^{-1}X'Y|X] \\ &= \mathbb{E}[(X'X)^{-1}X'(X\beta + U)|X] = \beta + (X'X)^{-1}X'\mathbb{E}[U|X] = \beta \; .\end{aligned}$$

By LIE, then, it is also unconditionally unbiased: $\mathbb{E}[\hat{\beta}_{OLS}] = \mathbb{E}[\mathbb{E}[\hat{\beta}_{OLS}|X]] = \beta$. Note that

[4] Furthermore, the FWL theorem is useful when we are interested in theoretically analyzing the properties of a single $\hat{\beta}_m$ out of a vector $\hat{\beta} \in \mathbb{R}^k$ as it allows us to obtain it using a univariate regression, i.e. a regression with a single covariate. This holds more generally if we are interested in a sub-vector of $\hat{\beta}$.

unbiasedness is not necessarily a desirable property; a biased estimator with a lower variance might be preferred, as it might lead to a lower frequentist risk under, say, the quadratic loss function (see Section 2.1.1).

Assumption 4 (Homoskedasticity)**.** *$\mathbb{V}[u_i|x_i] = \sigma^2$ is the same for all i.*

Under Assumption 1 and Assumption 4, the conditional variance of $\hat{\beta}_{OLS}$ is

$$\begin{aligned}\mathbb{V}[\hat{\beta}_{OLS}|X] = \mathbb{E}[(\hat{\beta}_{OLS} - \beta)(\hat{\beta}_{OLS} - \beta)'|X] &= \mathbb{E}[(X'X)^{-1}X'UU'X(X'X)^{-1}|X] \\ &= (X'X)^{-1}X'\mathbb{E}[UU'|X]X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \ ,\end{aligned}$$

because $\mathbb{E}[UU'|X] = \sigma^2 I$.[5] By LIE again, the unconditional variance of $\hat{\beta}_{OLS}$ is

$$\mathbb{V}[\hat{\beta}_{OLS}] = \mathbb{E}[(\hat{\beta}_{OLS} - \beta)(\hat{\beta}_{OLS} - \beta)'] = \mathbb{E}\left[\mathbb{E}[(\hat{\beta}_{OLS} - \beta)(\hat{\beta}_{OLS} - \beta)' \mid X]\right] = \sigma^2\mathbb{E}[(X'X)^{-1}] \ .$$

These expressions show that a higher σ^2 increases the variance of $\hat{\beta}_{OLS}$, i.e. with a more noisy outcome variable y_i, it is harder to precisely measure the effect of x_i on it. On the ohter hand, more variation in the regressors x_i help us to estimate β precisely.

Proposition 20 (Gauss-Markov Theorem)**.**
If Assumptions 1 to 4 hold, then $\hat{\beta}_{OLS}$ has the smallest variance among the class of linear unbiased estimators, i.e. the OLS estimator is BLUE (best linear unbiased estimator).

See proof in Appendix.

For now, we found $\hat{\beta}_{OLS}$ and its first and second moment, but not its whole distribution. This (finite sample) distribution is needed to conduct hypothesis tests like the t-test discussed in Section 2.1.2 as well as to form the related confidence sets. Following the discussion in Section 2.1.1, we could assume that $u_i|\beta, x_i$ is Normally distributed, which in turn would imply that $\hat{\beta}_{OLS}|\beta, X$ is Normal with the above mean and variance, a case treated in Section 4.1. However, oftentimes we do not want to make such a strong assumption. Besides, it would give us the distribution of $\hat{\beta}_{OLS}$ only conditional on X, and we typically do not want to treat X as fixed, but consider the distribution of $\hat{\beta}_{OLS}$ as we vary both the observations of our outcome variable Y as well as the observations of our regressors X. As in Section 2.1.1, we can approximate the (finite sample) distribution of $\hat{\beta}_{OLS}$ by its asymptotic distribution.

[5] Concretely, Assumption 1 ensures that the off-diagonal elements of $\mathbb{E}[UU'|X]$ are zero, while Assumption 4 ensures that the diagonal elements of $\mathbb{E}[UU'|X]$ are all equal to σ^2.

3.2 Asymptotic Properties

The asymptotic analysis of $\hat{\beta}_{OLS}$ aims at establishing its properties as our sample grows to infinity: $n \to \infty$. The resulting asymptotic distribution of $\hat{\beta}_{OLS}$ is commonly used to approximate its finite sample distribution when the latter is not available, hence enabling asymptotically valid hypothesis testing and confidence set construction.

Under Assumptions 1 to 3, $\hat{\beta}_{OLS}$ is consistent, i.e. $\hat{\beta}_{OLS} \xrightarrow{p} \beta$. We have

$$\hat{\beta} - \beta = (X'X)^{-1}X'U = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i u_i \xrightarrow{p} 0 \ .$$

By WLLN, the denominator $\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \xrightarrow{p} \mathbb{E}[x_i x_i'] \equiv Q$ and the numerator

$$\frac{1}{n} \sum_{i=1}^{n} x_i u_i \xrightarrow{p} \mathbb{E}[x_i u_i] = \mathbb{E}[x_i \mathbb{E}[u_i | x_i]] = 0 \ .$$

By Slutsky's theorem, $\left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \xrightarrow{p} Q^{-1}$. Finally, putting the two pieces together, and again using Slutsky's theorem, we get $\hat{\beta} - \beta \xrightarrow{p} Q^{-1} \cdot 0 = 0$.

If in addition Assumption 4 holds, then $\hat{\beta}_{OLS}$ is asymptotically Normal with

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i u_i \xrightarrow{d} N(0, \sigma^2 Q^{-1}) \ .$$

As before, by WLLN and Slutsky, $\left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \xrightarrow{p} Q^{-1}$. By CLT,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i u_i = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} x_i u_i - \mathbb{E}[x_i u_i] \right) \xrightarrow{d} N\left(0, \mathbb{V}[x_i u_i]\right) \ ,$$

because we know that $\mathbb{E}[x_i u_i] = \mathbb{E}[x_i \mathbb{E}[u_i | x_i]] = 0$. Thereby,

$$\mathbb{V}[x_i u_i] = \mathbb{E}[(x_i u_i)(x_i u_i)'] = \mathbb{E}[x_i x_i' \mathbb{E}[u_i^2 | x_i]] = Q \sigma^2 \ .$$

Putting the two pieces together by Slutsky's theorem gives

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} Q^{-1} N(0, \sigma^2 Q) = N(0, \sigma^2 Q^{-1}) \ .$$

Again, loosely speaking, $\hat{\beta} \xrightarrow{d} N(\beta, \frac{\sigma^2}{n} Q^{-1})$ for $n \to \infty$.

Often, the asymptotic distribution is used as an approximation of the finite sample distribution, reasoning that $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{n} Q^{-1})$ approximately for large n. Thereby, we do not know Q and σ^2, but we can estimate them using the consistent estimators

$$\hat{Q} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \xrightarrow{p} Q \ , \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2 \xrightarrow{p} \sigma^2 \ .$$

Consistency of $\hat{Q}$ follows by WLLN, while consistency of $\hat{\sigma}^2$ follows by consistency of $\hat{\beta}$ and the plug-in property.[6] Estimating an expectation by replacing it with a sample mean (and replacing any unknown objects therein with consistent estimators) is referred to as the analogy principle. Note that our resulting approximation of the finite sample distribution, $N(\beta, \frac{\hat{\sigma}^2}{n} \hat{Q}^{-1})$, deviates from the actual ones both because we are using the asymptotic distribution as an approximation and because we are estimating the objects that appear in the asymptotic distribution.

3.3 Hypothesis Testing

Based on $\hat{\beta} \overset{approx.}{\sim} N(\beta, \frac{\hat{\sigma}^2}{n} \hat{Q}^{-1})$, we know $\hat{\beta}_j \overset{approx.}{\sim} N(\beta_j, \frac{\hat{\sigma}^2}{n} [\hat{Q}^{-1}]_{jj})$ for a single parameter $\beta_j \in \beta$, whereby $[\hat{Q}^{-1}]_{jj}$ is element (j, j) in the matrix $\hat{Q}^{-1}$. This enables us to test a point hypothesis $\mathscr{H}_0 : \beta_j = \beta_{j,0}$ using the (two-sided) t-test:

$$\varphi_t(x) = \mathbf{1}\{T_t < c\} \ , \quad \text{with} \quad T_t = \left| \frac{\hat{\beta}_{j,n} - \beta_j}{\hat{\sigma}_{\beta_{j,0}}} \right| ; .$$

Because the distribution of $\hat{\beta}_j$ is not exactly, but only asymptotically Normal, so too does the resulting test-statistic only asymptotically converge to a standard Normal distribution:

$$\frac{\hat{\beta}_{j,n} - \beta_j}{\hat{\sigma}_{\beta_{j,0}}} \xrightarrow{d} N(0, 1) \ .$$

Thereby, $\hat{\sigma}_{\beta j} = \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\hat{Q}_{jj}^{-1}}$ is the estimate of the standard deviation of $\hat{\beta}_j$. Contrast this with the discussion in Section 2.1.2, where estimator and hence the t-statistic were exactly Normally distributed. As a result, this hypothesis testing procedure is only asymptotically valid. In finite samples, it is only approximate and can be more or less accurate depending on how close our approximation $\hat{\beta} \overset{approx.}{\sim} N(\beta, \frac{\hat{\sigma}^2}{n} \hat{Q}^{-1})$ is to the actual finite sample distribution of $\hat{\beta}$.

[6] $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i' \hat{\beta})^2 \rightarrow \mathbb{E}[(y_i - x_i' \beta)^2] = \mathbb{E}[u_i^2] = \sigma^2$.

More general hypotheses $\mathscr{H}_0 : g(\beta) = 0$ vs. $\mathscr{H}_1 : g(\beta) \neq 0$ for some function $g : \mathbb{R}^k \to \mathbb{R}^m$ (i.e. $m \leq k$ restrictions) can be tested using the Wald test. It uses the following statistic:

$$T_W = n\, g(\hat{\beta}_{OLS})' \left[G(\hat{\beta}_{OLS}) \hat{V} G(\hat{\beta}_{OLS})' \right]^{-1} g(\hat{\beta}_{OLS}) \xrightarrow{d} \chi^2_m \ ,$$

where $\hat{V} = \hat{\sigma}^2 \hat{Q}^{-1}$ and where $G(\hat{\beta}_{OLS}) = \partial g(\beta)/\partial \beta \mid_{\beta = \hat{\beta}_{OLS}}$ is the $m \times k$ matrix of derivatives of g with respect to β evaluated at $\hat{\beta}_{OLS}$. The short derivation in the Appendix illustrates that the Wald test-statistic is based on the idea that if $\mathscr{H}_0$ is true, then the difference between $g(\hat{\beta}_{OLS})$ and $g(\beta) = 0$ should be small. Suppose we are interested in testing $\mathscr{H}_0 : \{\beta_2 + \beta_3 = 5, \beta_4 = 0\}$ under a five-dimensional vector β. Then we would take

$$g(\beta) = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \beta - \begin{bmatrix} 5 \\ 0 \end{bmatrix} \ , \quad \text{with} \quad G(\beta) = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \ .$$

If $g(\beta) = 0$ is s.t. it tests only $\beta_j = \beta_{j,0}$ for a single β_j, then the Wald test is equivalent to the t-test: $\varphi_W = \varphi_t$.

3.4 Violations of Ideal Conditions

Throughout the previous sections, we assumed a bunch of things in order to obtain $\hat{\beta}$ and analyze its properties. Now we will investigate how these properties change when we relaax these conditions.

First of all, note that while unbiasedness requires the conditional independence assumption 2 to hold, both consistency and asymptotic Normality go through even under the weaker exogeneity assumption $\mathbb{E}[u_i x_i] = 0$.[7] In the following, more substantial deviations from the ideal conditions in Assumptions 1 to 4 are discussed.

3.4.1 Singular $X'X$

If $X'X$ is not of full rank, then the OLS estimator is not even defined. There are two reasons that lead to this case.

First, consider the standard case $n > k$, i.e. we have more observations than explanatory variables in x_i (and hence parameters in β to estimate). Then $X'X$ can be singular because of perfect multicollinearity, i.e. one variable $x_{i,m}$ is a linear combination of the other variables $\{x_{i,j}\}_{j=1:k, j\neq m}$ (for all i). As a result, X does not contain k linearly independent columns

[7] It is weaker because it is implied by conditional independence: if $\mathbb{E}[u_i|x_i] = 0$, then $\mathbb{E}[u_i x_i] = \mathbb{E}\left[\mathbb{E}[u_i x_i | x_i]\right] = \mathbb{E}[\mathbb{E}[u_i|x_i] x_i] = 0$ by LIE.

(variables) – i.e. $rank(X) < k$ – and so $rank(X'X) < k$.[8] In case of high but not perfect multicollinearity, $X'X$ is close to singular, and we get noisy estimates in finite samples.

Second, we could have $k > n$, i.e. more variables than observations available. Then, even without perfect multicollinearity, $rank(X) \leq n < k$ and so $rank(X'X) < k$. Bayesian inference (or regularization), discussed in Section 4.5, is a way to deal with this case.

3.4.2 Heteroskedasticity

Suppose we replace the homoskedasticity-assumption 4 with the following one:

Assumption 5 (Heteroskedasticity)**.** $\mathbb{V}[u_i|x_i] = \sigma_i^2$.

As can be verified easily, this has no bearing on unbiasedness – Assumptions 1 to 3 needed – nor on consistency and asymptotic Normality – Assumptions 1 and 3 and exogeneity needed. However, it changes the asymptotic variance of $\hat{\beta}_{OLS}$:

$$\sqrt{n}(\hat{\beta}_{OLS} - \beta) \xrightarrow{d} N\left(0\ ,\ Q^{-1}\mathbb{E}[x_i x_i' u_i^2]Q^{-1}\right)\ ,$$

because $\mathbb{V}[x_i u_i] = \mathbb{E}[x_i x_i' u_i^2]$ does not simplify to $\sigma^2 Q$ as under homoskedasticity.[9] The asymptotic variance can again be estimated by replacing $\mathbb{E}[x_i x_i' u_i^2]$ with its sample analogue as a consistent estmator: $\frac{1}{n}\sum_{i=1}^n x_i x_i' \hat{u}_i^2$. The resulting standard errors are commonly referred to as White-standard errors in after White (1980).[10]

Note that if the variances $\{\sigma_i^2\}_{i=1}^n$ were known, we could transform the heteroskedastic model into a homoskedastic one by writing the regression as

$$y_i/\sigma_i = (x_i/\sigma_i)'\beta + u_i/\sigma_i\ .$$

In this model, observations are weighted by the inverses of their standard deviations and, as a result, less noisy observations are given more weight as they are more informative about the relation between Y and X (the relation will be less disturbed by the error term). Letting

[8] For example, if $x = [x_1, x_2]'$, then $E(xx') = \begin{bmatrix} E(x_1^2) & E(x_1x_2) \\ E(x_1x_2) & E(x_2^2) \end{bmatrix}$ has determinant $|E(xx')| = E(x_1^2)E(x_2^2) - [E(x_1x_2)]^2$, which has to be non-zero for $X'X$ to have full rank (in population). If $x_1 = 1$ is a constant, then $|E(xx')| = E(x_2^2) - E(x_2)^2 = Var(x_2) \neq 0$ has to hold, i.e. we need variation in x_2 to avoid perfect multicollinearity.

[9] We can write $\mathbb{E}[x_i x_i' u_i^2] = \mathbb{E}[x_i x_i' \mathbb{E}[u_i^2|x_i]] = \mathbb{E}[x_i x_i' \sigma_i^2]$, but this is not very helpful.

[10] Note that the presence of heteroskedasticity also changes the finite sample variance to $\mathbb{V}[\hat{\beta}_{OLS}] = \mathbb{E}[(X'X)^{-1}X'\Sigma X(X'X)^{-1}]$.

$\mathbb{V}[U|X] = \Sigma = diag(\sigma_1^2, ..., \sigma_n^2)$, we can write this in matrix notation as

$$\Sigma^{-\frac{1}{2}}Y = \Sigma^{-\frac{1}{2}}X\beta + \Sigma^{-\frac{1}{2}}U \ , \quad \text{with} \quad \mathbb{V}[\Sigma^{-\frac{1}{2}}U|X] = I \ .$$

The OLS estimator from this transformed model is referred to as the Generalized Least Squares (GLS) estimator:

$$\begin{aligned}\hat{\beta}_{GLS} &= \left(\left(\Sigma^{-\frac{1}{2}}X\right)'\Sigma^{-\frac{1}{2}}X\right)^{-1}\left(\Sigma^{-\frac{1}{2}}X\right)'\Sigma^{-\frac{1}{2}}Y \\ &= \left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}Y \ .\end{aligned}$$

Under otherwise the same conditions as for OLS, this estimator is unbiased and consistent and has variance

$$\mathbb{V}(\hat{\beta}_{GLS}) = \mathbb{E}[(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}UU'\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}] = \mathbb{E}\left[(X'\Sigma^{-1}X)^{-1}\right] \ .$$

Proposition 21 (Revised Gauss Markov Theorem)**.**
Under Assumptions 1 to 3 and 5, $\hat{\beta}_{GLS}$ is BLUE.

See Appendix for proof. Based on this result, under heteroskedasticity with known variances in Σ, we should use $\hat{\beta}_{GLS}$. However, the GLS estimator is not feasible, because (usually) we do not know Σ. A feasible version replaces Σ by an estimate $\hat{\Sigma}$ obtained from a preliminary ("first stage") OLS estimation:

$$\hat{\beta}_{GLS,f.} = \left(X'\hat{\Sigma}^{-1}X\right)^{-1}X'\hat{\Sigma}^{-1}Y \ , \quad \hat{\Sigma} = diag(\hat{u}_1^2, \ldots, \hat{u}_n^2) \ , \ \ \hat{u}_i = y_i - x_i'\hat{\beta}_{OLS} \ .$$

However, it is not clear whether this feasible GLS estimator performs better than the OLS estimator since the estimate of Σ introduces additional noise, thereby increasing the variance of $\hat{\beta}_{GLS,f.}$ relative to $\hat{\beta}_{GLS}$.[11] [12]

[11] The asymptotic analysis of such a two-step estimator requires special care. It is discussed in Section 5.2.1. See Section 7.1 for a brief discussion of one-step GLS estimators.

[12] Sometimes (in particular for small n), a functional form for $\sigma_i^2 = g(x_i; \vartheta)$ is specified, with the goal to reduce the variance of $\hat{\Sigma}$. Then ϑ is estimated so that $\hat{u}_i^2$ from the first stage OLS regression is close to $g(x_i; \vartheta)$. A possible approach is then to try both GLS and OLS and trust the GLS estimates only if they are close to the OLS estimates, since even under heteroskedasticity, OLS is consistent (and unbiased), while GLS with $\sigma_i^2 = g(x_i; \vartheta)$ estimated is not if $g(x_i; \vartheta)$ is misspecified, as some observations can wrongly receive too high or too low weights.

3.4.3 Endogeneity

The essential assumption under oLS is exogeneity, $\mathbb{E}[x_i u_i] = 0$, as it ensures consistency. Only the stronger conditional independence assumption $\mathbb{E}[u_i|x_i] = 0$ ensures unbiasedness, but we typically can live with a (finite-sample) bias provided that it vanishes asymptotically, i.e. provided that we have consistency.

Endogeneity refers to the case when $\mathbb{E}[x_i u_i] \neq 0$, implying inconsistency of the OLS estimator. In the following, three causes of endogeneity are discussed: measurement errors, simultaneity, and omitted variables. A possible remedy for each case is IV estimation, discussed in Section 6.4.

These issues have to be dealt with if the goal is to obtain a consistent estimator. This is oftentimes the case, as we stipulate that some outcome variable y_i is indeed generated by a linear model (e.g. as postulated within some economic theory) and move on to estimate the coefficients in that linear model. Furthermore, numerous empirical studies require consistency as they are interested in the "causal" effect of some covariate x_k on the outcome y_i. However, in these cases it is a better idea to apply causal inference methods (see Chapter 12), which construct consistent estimators without assuming any particular model that generated the data, let alone the rather restrictive linear regression model. In contrast, any estimator of an effect of some x_k on y_i derived based on a linear regression is consistent only if that linear model is correctly specified.

Regardless of endogeneity, linear regressions can always be used as devices to uncover partial correlations. See the FWL theorem (Proposition 19) above and recall the example of determining the partial correlation between race and income accounting for educational differences between racial groups.

Measurement Errors in Regressors Endogeneity is violated if the regressors are measured with error. Suppose the true model is

$$y_i = x_i^{*\prime}\beta + \varepsilon_i \ , \quad \text{with} \quad \mathbb{E}[x_i^* \varepsilon_i] = 0 \ ,$$

but the researcher estimates $y_i = x_i'\beta + u_i$ using $x_i = x_i^* + v_i$, where v_i is the measurement error. Then we know $u_i = \varepsilon_i - v_i'\beta$.

Now, suppose the measurement error v_i is completely random: it is uncorrelated with the true x_i^* as well as ε_i (i.e. all factors other than x_i^* that affect y_i). Even then exogeneity fails:

$$\mathbb{E}[x_i u_i] = \mathbb{E}[(x_i^* + v_i)(-v_i'\beta + \varepsilon_i)] = -\mathbb{E}[x_i^* v_i']\beta - \mathbb{E}[v_i v_i']\beta + \mathbb{E}[v_i \varepsilon_i] = -\mathbb{E}[v_i v_i']\beta \ .$$

If we consider a scalar regressor, then $\mathbb{E}[v_i v_i'] = \mathbb{E}[v_i^2] > 0$ and so $\hat{\beta}_{OLS}$ is biased towards zero.[13] [14] The bias and inconsistency can be mitigated with IV methods, provided that we have a variable z_i that is correlated with x_i but uncorrelated with u_i. This could be a second, noisy measure of x_i^* whose measurement error is uncorrelated with that of x_i, v_i.

In contrast, measurement errors in the outcome variable y_i are absorbed in u_i. Concretely, suppose the true model is

$$y_i^* = x_i'\beta + \varepsilon_i \ ,$$

but we estimate $y_i = x_i'\beta + u_i$ for $y_i = y_i^* + v_i$. Then, $u_i = \varepsilon_i + v_i$. As long as x_i is uncorrelated with measurement errors v_i, exogeneity is satisfied and OLS is consistent: $\mathbb{E}[x_i u_i] = 0$. In that case, the only negative consequence of measurement errors is that they increase the variance of the error term u_i (relative to the variance of the true error term ε_i), leading to more noisy estimators.

Simultaneity Suppose we are interested in estimating β in the linear regression

$$y_i = x_i'\beta + u_i \ ,$$

whereby the mth covariate in x_i,

$$x_{i,m} = z_i'\gamma + y_i\delta + v_i \ ,$$

is a function of y_i. In other words, simultaneously y_i is determined by $x_{i,m}$ (among other variables) and $x_{i,m}$ is determined by y_i. In the context of the estimation of causal effects, we also speak of "reverse causality". Inserting the first equation for y_i in the second equation, it is easy to see that $\mathbb{E}[x_{i,m}u_i] \neq 0$, i.e. exogeneity is violated. Assuming that all other covariates in x_i are exogenous, i.e. $\mathbb{E}[x_{i,l}u_i] = 0$ for $l \neq m$, we can obtain a consistent estimator by using z_i as an IV for $x_{i,m}$.[15]

If we have a set of variables $y_{i1}, ..., y_{ik}$ all of which are (possibly) simultaneously determined,

[13] This means that if $\beta > 0$, then $\hat{\beta}_{OLS}$ has a downward bias, while if $\beta < 0$, then $\hat{\beta}_{OLS}$ has an upward bias.

[14] Exogeneity would be satisfied under the odd case that v_i is correlated with the true regressor x_i^* but uncorrelated with the actually measured x_i and with ε_i. We would get

$$\mathbb{E}[x_i u_i] = -\mathbb{E}[x_i v_i'\beta] + \mathbb{E}[x_i \varepsilon_i] = -\mathbb{E}[x_i v_i'\beta] + \mathbb{E}[(x_i^* + v_i)\varepsilon_i] = 0 \ .$$

In reality, one likely encounters a case in-between these two extremes. The bottom line is that OLS is inconsistent under measurement errors in the explanatory variables.

[15] That is, we would use $(x_{i,-m}', z_i')'$ as an IV for x_i, where $x_{i,-m} = (x_{i,1}, ..., x_{i,m-1}, x_{i,m+1}, ..., x_{ik})'$ encompasses all covariates in x_i other than $x_{i,m}$.

we speak of a simultaneous equation system. For simplicity, let $k = 2$. We have

$$\begin{aligned} y_{i1} &= x_i'\gamma_1 + y_{i2}\delta_1 + v_{i1} \\ y_{i2} &= x_i'\gamma_2 + y_{i1}\delta_2 + v_{i2} \ , \end{aligned}$$

where v_{i1} and v_{i2} are independent errors. The canonical example is supply and demand, which are simultaneously determined, as suppliers set the price based on the (supposed) demand function, and consumers possibly adjust their demand based on their perceived relation between price and supplied quantity. For $y_i = (y_{i1}, y_{i2})'$, we can write this in matrix-notation as

$$Ay_i = Cx_i + v_i \ , \quad \mathbb{V}[v_i] = D$$

where D is diagonal and

$$A = \begin{bmatrix} 1 & -\delta_1 \\ -\delta_2 & 1 \end{bmatrix} \ , \quad C = \begin{bmatrix} \gamma_1' \\ \gamma_2' \end{bmatrix} \ .$$

This case arises naturally in the context of causal analysis for multivariate time series (see Section 9.2), and analogous methods can be applied to estimate A, C and D.[16]

Omitted Variables The most prominent cause of endogeneity are omitted variables (OVs). Suppose the true model is

$$y_i = x_i'\beta + w_i'\delta + \varepsilon_i \ ,$$

where exogeneity holds: $\mathbb{E}[x_i\varepsilon_i] = 0$. Suppose further that the researcher omits w_i and instead estimates

$$y_i = x_i'\beta + u_i \ , \quad \text{where} \quad u_i = \varepsilon_i + w_i'\delta \ .$$

In this misspecified model, exogeneity is only given if x_i and w_i are uncorrelated, since

$$\mathbb{E}[x_iu_i] = \mathbb{E}[x_i(w_i'\delta + \varepsilon_i)] = \mathbb{E}[x_iw_i']\delta \ .$$

Since $\hat{\beta} - \beta \xrightarrow{p} \mathbb{E}[x_ix_i']^{-1}\mathbb{E}[x_iu_i]$, we can assess the sign (and size) of the asymptotic bias based on the signs (and sizes) of the correlation between x_i and w_i, $\mathbb{E}[x_iw_i']$, and the "effect" of w_i on y_i, δ. Once again, the inconsistency can be mitigated with IV methods, provided that we have a variable z_i that is correlated with x_i but uncorrelated with u_i and in particular with the omitted variable(s) w_i.

[16] Section 9.2 deals with vector-autoregressions (VARs). For $p = 1$ lag, a VAR(p) can be written in structural form as $Ay_t = By_{t-1} + \varepsilon_t$. It is assumed that we have T observations of the n-dimensional variable y_t. To deal with cross-sectional simultaneous equation systems, take $p = 0$ lags, add covariates x_t multiplied by C to the equation, denote the dimension of y_t by k, change the notation from t-subscripts to i-subscripts and replace T by n. Furthermore, the analysis is simplified by the fact that we typically assume independence of cross-sectional observations, whereas this does not hold for data observed over time.

3.5 Covariates as "Control Variables"

Suppose you are interested in relating a variable x_i^* linearly to an outcome variable y_i,

$$y_i = \beta_0 + \beta_1 x_i^* + u_i \ ,$$

and your goal is to consistently estimate β_1, which you interpret through the lens of this linear model as the effect of x_i^* on y_i. Suppose further that you have a whole bunch of other covariates x_i available. Such variables, whose effect on y_i is not of direct interest, are referred to as (potential) control variables, which could be added to the linear regression above.

The discussion in the previous sections provides guidance on which variables must and which must not be added if the goal is to obtain a consistent estimator of β_1. In the above model, we can think of x_i as being contained in the error term u_i. For simplicity, assume for the moment that these variables enter linearly: $u_i = x_i'\delta$. Thereby, some of the elements of δ may be zero, meaning that the corresponding variables in x_i are irrelevant for y_i. The question is ultimately which of these variables we want to "control for", i.e. bring out of the error term and add as additional regressor ("control variables"), leading to

$$y_i = \beta_0 + \beta_1 x_i^* + \tilde{x}_i'\tilde{\delta} + \varepsilon_i \ ,$$

where $\tilde{x}_i$ is a subset of the variables in x_i and $\tilde{\delta}$ is the corresponding subset of δ.

The discussion on the OV bias in Section 3.4.3 suggests that we must include any variable $x_{i,m}$ that

1. affects y_i, i.e. is indeed present in u_i ($\delta_m \neq 0$); and
2. is correlated with x_i^* ($\mathbb{E}[x_{i,m}x_i^*] \neq 0$).

Such variables are generally seen as "good controls" (see caveat below). In fact, they are essential controls, as omitting them implies a biased and inconsistent estimator of β_1.

Variables $x_{i,m}$ that do not satisfy the first of these two conditions are irrelevant to the analysis. They are bad controls. Asymptotically, their inclusion or omission has no effect on the properties of our estimator of β_1, as the estimator of their corresponding coefficient converges to its true value of zero. However, in finite samples their inclusion can lead to a more noisy estimator for β_1 as they might be correlated with x_i^*. To see this, recall that by the FWL theorem, the estimator $\hat{\beta}_1$ in a regression of y_i on x_i^* and such irrelevant variables $x_{i,m}$ is the same as the estimator obtained when regressing y_i on x_i^* with $x_{i,m}$ "partialled-out".[17] Relative to the case of estimating β_1 in a regression of y_i only on x_i^*, including $x_{i,m}$ may reduce the variance left in x_i^* after partialling-out $x_{i,m}$ more than it reduces the variance left in y_i, leading to a more noisy estimator of β_1.

By analogous reasoning, variables $x_{i,m}$ that satisfy the first condition but not the second are good controls. As before, their inclusion or omission has asymptotically no effect on our estimator of β_1, as they are uncorrelated with x_i^*. However, they are likely to lead to a more precise estimator in finite samples as they remove more variation in y_i than in x_i^*.

The discussion on simultaneity in Section 3.4.3 presents a caveat to the preceding arguments. It demonstrates that we must not add a variable $x_{i,m}$ that both affects y_i and is itself affected by y_i. By the above, if such a $x_{i,m}$ is uncorrelated with x_i^*, then it is a bad control that we can comfortably omit from the analysis. However, if it is correlated with x_i^*, we either need to include it into the analysis and estimate a simultaneous equations system (see Section 3.4.3), or we omit it and try to arrive at a consistent estimator of β_1 by using an IV that is correlated with x_i^* but uncorrelated with such a $x_{i,m}$.

Furthermore, the discussion on measurement errors in Section 3.4.3 suggests that any variable we add should be measured without error. Again, if such a mis-measured variable is uncorrelated with x_i^*, it is a bad control that we should ignore. If instead it is correlated with x_i^*, we ignore it and construct a consistent estimator of β_1 using an IV that is correlated with x_i^* but uncorrelated with such a $x_{i,m}$.

A more fundamental word of caution is that adding OVs – i.e. variables $x_{i,m}$ that affect y_i and are correlated with x_i^* – only renders the estimator of β_1 consistent if indeed the resulting model is correctly specified. Concretely, estimating

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 x_{i,m} + v_i$$

instead of the above $y_i = \beta_0 + \beta_1 x_i^* + u_i$ will only solve inconsistency due to OVs if indeed $x_{i,m}$ linearly affects y_i. If instead $u_i = \beta_2 x_{i,m} + \beta_3 x_{i,m}^2 + \varepsilon_i$, say, then the OV issue persists, as $\beta_3 x_{i,m}^2$ is left in the error term v_i and, if x_i^* is correlated with $x_{i,m}$, then it typically is with its square, too. In this example, it would be easy to add the square term as well, but typically we do not know the exact functional form by which $x_{i,m}$ – or any other variable for that matter, including x_i^* – affects y_i. The following chapters present more flexible approaches to model the relation between y_i and x_i^* and other covariates x_i. However, ultimately, consistency of any estimator derived within a structural model requires that model to be correctly specified. As such, structural models are useful either as statistical devices (e.g.

[17] Recall the terminology from Section 3.1; this means regressing the residual from a regression of y_i on $x_{i,m}$ on the residual from a regression of x_i^* on $x_{i,m}$.

the linear regression as a partial correlation analysis; see discussion in Section 3.1) or in cases where we have a good reason to condition our analysis on a structural model, e.g. if the latter is derived in some economic theory and we want to estimate a parameter within that theoretical framework. If instead the interest lies in establishing the causal effect of some x_i^* on some y_i in a particular real-world setting, such as in the context of policy evaluation, then the model-free causal inference methods (see Chapter 12) or the non-parametric local regressions (see Section 13.2) are required. The discussion in Section 12.3.3 is particularly relevant from an OV-perspective.

Appendix

Claim. *(FWL Theorem) Let $Y = X_1\beta_1 + X_2\beta_2 + U$. Take $M_1 = I - P_1$ and $P_1 = X_1(X_1'X_1)^{-1}X_1'$ and write $M_1Y = M_1X_2b + V$. Then $\hat{b}_{OLS} = \hat{\beta}_{2,OLS}$ and $\hat{U} = \hat{V}$.*

Proof: We can write

$$Y = P_XY + M_XY = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + M_XY \ .$$

Left-multiplying this expression by $X_2'M_1$ gives

$$X_2'M_1Y = X_2'M_1X_1\hat{\beta}_1 + X_2'M_1X_2\hat{\beta}_2 + X_2'M_1M_XY \ .$$

We know $M_1X_1 = 0$. Also, $M_XX_1 = 0$ and $M_XX_2 = 0$, which implies $M_XM_1X_2 = 0$. As a result, solving for $\hat{\beta}_2$ yields

$$\hat{\beta}_2 = (X_2'M_1X_2)^{-1}X_2'M_1Y = ((M_1X_2)'M_1X_2)^{-1}(M_1X_2)'Y = \hat{b}_{OLS} \ .$$

Multiplying the above expression for Y by M_1 instead gives

$$M_1Y = M_1X_2\hat{\beta}_2 + M_1M_XY \ ,$$

i.e. $\hat{V} = M_1M_XY = M_XY = \hat{U}$. ■

As an example, consider

$$y_i = \beta_0 + \beta_1x_i + u_i = [1, x_i]\beta + u_i \ ,$$

where $\beta = (\beta_0, \beta_1)'$. In matrix notation, this is

$$Y = \tilde{X}\beta + U \ ,$$

where $\tilde{X} = [\iota, X]$ and ι is an $n \times 1$ vector of ones. The following establishes using linear algebra that the FWL theorem holds in this case, i.e. $\hat{\beta}_1$ coincides with $\hat{b}$, the OLS estimator of b from the regression

$$M_\iota Y = M_\iota X b + V \ .$$

Note that the latter is equivalently written as

$$(y_i - \bar{y}) = (x_i - \bar{x})b + v_i \ ,$$

as

$$M_\iota Y = (I - \iota(\iota'\iota)^{-1}\iota')Y = \left(I - \frac{1}{n}\iota\iota'\right)Y = Y - \iota\bar{y} \ ,$$

due to the facts that $\iota'\iota = n$ and $\iota' Y = \sum_{i=1}^n y_i$. The analogous holds for $M_\iota X$.

With two regressors, we can find $\hat{\beta}_0$ and $\hat{\beta}_1$ separately using linear algebra but avoiding matrix notation:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

leads to the FOCs

$$-2\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$$
$$-2\sum_{i=1}^n x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \ .$$

Solving the former for β_0, we get $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. Plugging this expression into the latter FOC and solving for β_1, we get

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{\sum_{i=1}^n x_i(x_i - \bar{x})} \ .$$

Under matrix notation, we know the expression for $\hat{\beta}$, but we need to separate out $\hat{\beta}_1$, its second element:

$$
\begin{aligned}
\hat{\beta} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y \\
&= \left(\sum_{i=1}^n \tilde{x}_i\tilde{x}_i'\right)^{-1}\sum_{i=1}^n \tilde{x}_i y_i \\
&= \left(\sum_{i=1}^n \begin{bmatrix}1\\ x_i\end{bmatrix}\begin{bmatrix}1 & x_i\end{bmatrix}\right)^{-1}\sum_{i=1}^n \begin{bmatrix}1\\ x_i\end{bmatrix} y_i \\
&= \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1}\begin{bmatrix}\sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i\end{bmatrix} \\
&= \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\begin{bmatrix}\sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n\end{bmatrix}\begin{bmatrix}\sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i\end{bmatrix} \\
&= \frac{1}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}\begin{bmatrix}\frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1\end{bmatrix}\begin{bmatrix}\bar{y} \\ \frac{1}{n}\sum_{i=1}^n x_i y_i\end{bmatrix}\ .
\end{aligned}
$$

This shows that

$$
\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2} = \frac{\frac{1}{n}\sum_{i=1}^n x_i(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n x_i(x_i - \bar{x})}\ .
$$
[18]

For the OLS estimator of b, we have

$$
\begin{aligned}
\hat{b} &= ((M_\iota X)'(M_\iota X))^{-1}(M_\iota X)'(M_\iota Y) \\
&= (X'M_\iota X)^{-1} X'M_\iota Y \\
&= \left(X'X - X'\iota(\iota'\iota)^{-1}\iota X\right)^{-1}\left(X'Y - X'\iota(\iota'\iota)^{-1}\iota Y\right) \\
&= \left(\sum_{i=1}^n x_i^2 - \frac{1}{n}(\sum_{i=1}^n x_i)^2\right)^{-1}\left(\sum_{i=1}^n x_i y_i - \frac{1}{n}(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}\right)\ ,
\end{aligned}
$$

which establishes that $\hat{b} = \hat{\beta}_1$.

Claim. *(Gauss-Markov Theorem) If Assumptions 1 to 4 hold, then $\hat{\beta}_{OLS}$ has the smallest variance among the class of linear unbiased estimators, i.e. the OLS estimator is BLUE (best linear unbiased estimator).*

[18] For the last equality, note that $\bar{x}\bar{y} = \frac{1}{n}\sum_{i=1}^n x_i\bar{y} = \frac{1}{n}\sum_{i=1}^n \bar{x}y_i$.

Proof: Take an alternative linear estimator $\tilde{\beta} = A \cdot Y$ for some A. We can write $\tilde{\beta} = \hat{\beta} + CY = \hat{\beta} + CX\beta + CU$, with $C = A - (X'X)^{-1}X$. By unbiasedness, $CX = 0$. We get

$$\mathbb{V}[\tilde{\beta}] = \mathbb{V}[\hat{\beta}] + \mathbb{V}[CU] + \mathrm{Cov}(\hat{\beta}, CU) + \mathrm{Cov}(CU, \hat{\beta}) \ .$$

Thereby,

$$\mathrm{Cov}(\hat{\beta}, CU) = \mathbb{E}[\hat{\beta}(CU)'] - \mathbb{E}[\hat{\beta}]\mathbb{E}[CU]' = 0$$

because i) $\mathbb{E}[CU] = \mathbb{E}[C\mathbb{E}[U|X]] = 0$ by Assumption 2, and because ii)

$$\mathbb{E}[\hat{\beta}(CU)'] = \mathbb{E}[\beta U'C' + (X'X)^{-1}X'UU'C'] = \sigma^2\mathbb{E}[(X'X)^{-1}X'C'] = 0$$

by LIE, Assumption 4 and using $CX = 0$. Overall, $\mathbb{V}[\hat{\beta}] \geqslant \mathbb{V}[\hat{\beta}]$ as $\mathbb{V}[CU] \geqslant 0$. ■

Claim. *(Revised Gauss Markov Theorem) Under Assumptions 1 to 3 and 5, $\hat{\beta}_{GLS}$ is BLUE.*

Proof: Proof is analogous to the proof of the Gauss Markov Theorem. Take an alternative linear estimator

$$\tilde{\beta} = A \cdot Y = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y + CY = \hat{\beta} + CX\beta + CU \ .$$

Unbiasedness implies $CX = 0$. We get

$$\mathbb{V}[\tilde{\beta}] = \mathbb{V}[\hat{\beta}] + \mathbb{V}[CU] + \mathrm{Cov}\left(\hat{\beta}, CU\right) + \mathrm{Cov}\left(CU, \hat{\beta}\right) \ ,$$

with

$$\mathrm{Cov}\left(\hat{\beta}, CU\right) = \mathbb{E}[\hat{\beta}\,(CU)'] - \mathbb{E}[\hat{\beta}]\mathbb{E}[CU]' = 0 \ ,$$

because i) $\mathbb{E}[CU] = \mathbb{E}[C\mathbb{E}[U|X]] = 0$ by Assumption 2, and because ii)

$$\mathbb{E}[\hat{\beta}(CU)'] = \mathbb{E}[\beta U'C' + (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}UU'C'] = \mathbb{E}[(X'\Sigma^{-1}X)^{-1}X'C'] = 0$$

by LIE, Assumption 5 and using $CX = 0$. Overall, $\mathbb{V}[\hat{\beta}] \geqslant \mathbb{V}[\hat{\beta}]$ as $\mathbb{V}[CU] \geqslant 0$. ■

Wald Test The asymptotic distribution of the Wald-test-statistic T_W is derived as follows. The starting point is the fact that $\sqrt{n}(\hat{\beta}-\beta_0) \overset{d}{\to} N(0,V)$. The Delta method (Proposition 11) then tells us that

$$\sqrt{n}\left(g(\hat{\beta}) - g(\beta_0)\right) \overset{d}{\to} G(\beta_0) \cdot N(0,V) = N(0, G(\beta_0)VG(\beta_0)') \ , \quad \text{with} \quad G(\beta_0) = \left. \frac{\partial g(\beta)}{\partial \beta} \right|_{\beta=\beta_0} \ .$$

Finally, using the fact that $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2_{dim(X)}$ for $X \sim N(\mu, \Sigma)$, we get

$$\sqrt{n}\left(g(\hat{\beta}) - g(\beta_0)\right)' \left[G(\beta_0)VG(\beta_0)'\right]^{-1} \sqrt{n}\left(g(\hat{\beta}) - g(\beta_0)\right) \overset{d}{\to} \chi^2_m \ .$$

Under $\mathscr{H}_0$, $g(\beta_0) = 0$. Also, because we do not know β_0, we replace $G(\beta_0)$ with $G(\hat{\beta})$, as $\hat{\beta}$ is our estimator of β_0.