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.

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

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

Assumption 2. *The matrix $X'X = \sum_{i=1}^n x_i x_i'$ is of full rank.*

Under Assumption 2, $X'X$ is invertible and we can solve the first order condition (FOC)

$X'(Y - X\beta) = 0$ for β to get

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

We get the predicted values $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = P_X Y$ and residuals $\hat{U} = Y - \hat{Y} = (I - P_X)Y = M_X Y$, where P_X and M_X are projection matrices. They are idempotent – i.e. $P_x = P_x'$ and $P_x P_x = P_x$, and same for M_x – and 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:

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

Proposition 19 (Frisch-Waugh-Lovell (FWL) Theorem)**.**
Let $Y = X_1\beta_1 + X_2\beta_2 + U$. 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 that regressing Y on X_1 and X_2 and taking the OLS estimator for β_2 gives the same as first regressing Y on X_1 only as well as X_2 on X_1 and then regressing the residuals from the first regression on the residuals from the second regression. This theorem is useful when we are interested in 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.[3] Note that the FWL-theorem 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 (to see this, take X_1 to be a vector of ones and therefore β_1 to be the intercept).

[3] Or more generally if we are interested in a sub-vector of $\hat{\beta}$.

Assumption 3 (Conditional Independence). $\mathbb{E}[u_i|x_i] = \mathbb{E}[u_i] = 0$ *for* $u_i = y_i - x_i'\beta$.

Assumption 3 states that regressors x_i and errors u_i are independent. Note that $\mathbb{E}[u_i] = 0$ is guaranteed as long as an intercept is included (or demeaned variables are used; see Proposition 19 above). So the crucial part of this assumption lies in equating $\mathbb{E}[u_i|x_i] = \mathbb{E}[u_i]$. Because we can always write $y_i = \mathbb{E}[y_i|x_i] + u_i$ with $\mathbb{E}[u_i|x_i] = 0$, Assumption 3 also means that the conditional expectation function of y_i given x_i is linear: $\mathbb{E}[y_i|x_i] = x_i'\beta$, or $\mathbb{E}[Y|X] = X\beta$ in matrix notation.

Under Assumption 3, $\hat{\beta}_{OLS}$ 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 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 2.2).

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

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

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

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.

3.2 Asymptotic Properties

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 2.3 as well as to form the related confidence sets. 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.[4] 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 exact, but only asymptotically valid, 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 2.3, 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

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

$\hat{\beta}$.

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 3 to hold, both consistency and asymptotic Normality go through even under the weaker exogeneity assumption $\mathbb{E}[u_i x_i] = 0$.[5] 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

[5] 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.

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 (variables) – i.e. $rank(X) < k$ – and so $rank(X'X) < k$.[6] 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 $rank(X'X) \leq n < k$. 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 2 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.[7] 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).[8]

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

[6] For example, if $x = [x_1, x_2]'$, then $E(xx') = \begin{bmatrix} E(x_1^2) & E(x_1 x_2) \\ E(x_1 x_2) & E(x_2^2) \end{bmatrix}$ has determinant $|E(xx')| = E(x_1^2)E(x_2^2) - [E(x_1 x_2)]^2$, 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.

[7] 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.

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

the relation between Y and X (the relation will be less disturbed by the error term). Letting $\mathbb{V}[U|X] = \Sigma = diag(\sigma_1^2, ..., \sigma_n^2)$, 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 as $\hat{\Sigma} = diag(\hat{u}_1^2, \ldots, \hat{u}_n^2)$. However, it is not clear whether this feasible GLS performs better than OLS since the estimate of Σ introduces additional variation, thereby increasing the variance of $\hat{\beta}_{GLS}$.[9]

3.4.3 Endogeneity

While only the stronger conditional independence assumption $\mathbb{E}[u_i|x_i] = 0$ ensures unbiasedness, the essential assumption under OLS is exogeneity – $\mathbb{E}[x_i u_i] = 0$ – as it suffices to obtain consistency. Endogeneity refers to the case when $\mathbb{E}[x_i u_i] \neq 0$, which leads to inconsistency

[9] 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.

of the OLS estimator. In the following, two cases are discussed that can lead to endogeneity. A possible remedy is IV estimation, discussed in Section 6.4.

Omitted Variables Endogeneity can arise if there are omitted variables. Suppose the true model is

$$y_i = x_i'\beta + z_i'\delta + \varepsilon_i \ , \quad \text{with} \quad \mathbb{E}[x_i\varepsilon_i] = 0 \ ,$$

i.e. exogeneity holds in this true model, whereas the researcher estimates $y_i = x_i'\beta + u_i$ instead. In this misspecified model, exogeneity is only given if x_i and z_i are uncorrelated, since

$$\mathbb{E}[x_i u_i] = \mathbb{E}[x_i(z_i'\delta + \varepsilon_i)] = \mathbb{E}[x_i z_i']\delta \ .$$

The size and sign of the (asymptotic) bias can be assessed based on $\mathbb{E}[x_i z_i']$ and δ.

Measurement Errors Endogeneity is also 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.

Suppose the measurement error v_i is uncorrelated with the true x_i^* as well as ε_i (i.e. all factors other than x_i^* that affect y_i), i.e. the measurement error is completely random. 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.[10]

Suppose instead that v_i is correlated with the true regressor x_i^*, and only with x_i^*, i.e. it is uncorrelated with the actually measured x_i as well as ε_i. Then exogeneity holds:

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

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.

[10] i.e. if $\beta > 0$, then $\hat{\beta}_{OLS}$ has a downward bias, while if $\beta < 0$, then $\hat{\beta}_{OLS}$ has an upward bias.

Note that measurement errors in the outcome variable y_i are absorbed in u_i.[11] As long as x_i is uncorrelated with these measurement errors, we have $\mathbb{E}[x_i u_i] = 0$ and OLS is consistent.

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 $M_1 Y = M_1 X_2 b + V$. Then $\hat{b}_{OLS} = \hat{\beta}_{2,OLS}$ and $\hat{U} = \hat{V}$.*

Proof: We can write

$$Y = P_X Y + M_X Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + M_X Y \ .$$

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

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).*

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

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

Thereby,

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

[11] i.e. if $y_i = y_i^* + v_i$ and the true model is $y_i^* = x_i'\beta + \varepsilon_i$, we estimate $y_i = x_i'\beta + u_i$ with $u_i = \varepsilon_i + v_i$.

because i) $\mathbb{E}[CU] = \mathbb{E}[C\mathbb{E}[U|X]] = 0$ by Assumption 3, 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, $\mathbb{V}[\tilde{\beta}] \geqslant \mathbb{V}[\hat{\beta}]$ 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

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

with

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

because i) $\mathbb{E}[CU] = \mathbb{E}[C\mathbb{E}[U|X]] = 0$ by Assumption 3, 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, $\mathbb{V}[\tilde{\beta}] \geqslant \mathbb{V}[\hat{\beta}]$ as $\mathbb{V}[CU] \geqslant 0$. ■

Wald Test The asymptotic distribution of the Wald-test-statistic, T_W, follows from asymptotic Normality of $\hat{\beta}$, the Delta method (Proposition 11) and the fact that $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2_{dim(X)}$ for $X \sim N(\mu, \Sigma)$.

Using $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V)$ and the Delta method, we get

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

Therefore,

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

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