LEAST SQUARES AND INSTRUMENTAL VARIABLE ESTIMATION IN ECONOMETRICS

ANDREAS PETROU-ZENIOU

Abstract. Econometricians are primarily concerned with the application of statistical methodologies to data with the goal of inferring causality. In many cases, classical least squares regression is not sufficient to infer causality. In even more cases, classical least squares is inconsistent with many standard econometric models. This paper will cover one of these cases, endogeneity, where regressors may be linked with the outcomes indirectly.

Contents

1. Introduction

This paper provides a detailed introduction to the linear algebraic foundations of econometric regression models, with an emphasis on building intuition through

Date: July 2021.

thorough treatment of mathematical derivations and examples. We introduce the linear regression of random vector X onto random variable Y with error term e:

$$Y = X'\beta + e \tag{1.1}$$

Here, X' denotes the transpose of column vector X. We also cover the least squares estimator for this regression:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{1.2}$$

We also introduce common assumptions in econometrics, and the models econometricians use when these assumptions do not hold, such as generalized least squares. Finally, we work towards solving the endogeneity problem, where, for example, regressors may be linked to the outcome variable through some unobservable factor. We introduce the notion of *instrumental variables* Z and introduce two models to handle endogeneity, the instrumental variable (IV) estimator

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y \tag{1.3}$$

and two-stage least squares (2SLS) estimator.

$$\hat{\beta}_{2SLS} = \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'Y \tag{1.4}$$

This paper assumes some statistics and linear algebra exposure, though proofs and computations will be completed in as much detail as possible. The notation for this paper is from both [2] and [6].

2. Least Squares

In least squares regression, we attempt to fit a line or plane to a set of observations. In other words, we ostensibly want to predict out-of-sample expectations for the outcome variable given out-of-sample regressor values. This section will cover the theoretical foundations for ordinary least squares (OLS) regression in predictor functions, introduce the least squares estimator and its properties, and introduce generalized least squares as the best linear unbiased estimator.

2.1. OLS Example: Acemoglu et al. (2001).

Throughout this paper, we use the paper by Acemoglu, Johnson, and Robinson, *The Colonial Origins of Comparative Development: An Empirical Investigation* as an example of linear regression in action. Acemoglu et al. try to determine whether stronger, democratic institutions have a positive effect on national wealth in former colonies. The *outcome variable* is log GDP, whereas one of their *regressors* will be average protection against expropriation risk, a score assigned to different nations reflecting property right strength which serves as a proxy for institutional robustness. Acemoglu et al. also include other regressors to control for factors other than institutional strength. By controlling for other regressors, Acemoglu et al. ensure *ceteris paribus*, or all else equal, allowing them to determine causality. Hence, they control for other variables, such as location. Figure 2.1 on the next page displays the results from Acemoglu et al. regression:

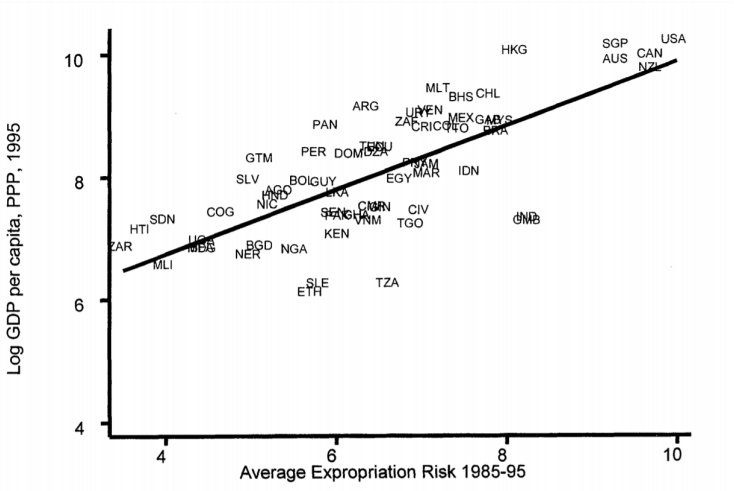
FIGURE 1. OLS regression comparing GDP to institutional strength

OLS Regression			
Regressor	Sample (1)	Sample (2)	Sample (3)
Average protection against expropriation risk, 1985-1995	0.52 (0.06)	0.47 (0.06)	0.41 (0.06)
Latitude	-	1.60 (0.70)	0.92 (0.63)
Asia dummy	-	-	-0.60 (0.23)
Africa dummy	-	-	-0.90 (0.17)
Other continent dummy (exclude America)	-	-	-0.04 (0.32)

Note that a dummy variable has a binary value. For example, the Asia dummy random variable equals 1 when a nation is in Asia, 0 when it is not.

The sample for each regression represents a selection of former colonies. Therefore, our third regression suggests that a 1 point increase in average protection against expropriation is associated with a 0.41 percent increase in GDP. To provide visual intuition, we observe the graph of this OLS regression, directly from [4]:

FIGURE 2. OLS regression of average expropriation and log GDP, Acemoglu et al. (2001)



With this real world application of OLS established, we can dive into the theoretical details of regression. However, also consider some potential flaws with the regression described above. Can we truly infer causality from this correlational

approach? We will investigate this question when we next cover this example in section 3.2, where we will introduce instrumental variables.

2.2. **Conditional Expectation Function and Best Linear Predictor.** This section introduces probabilistic models called *structural models* as the main theoretical foundation for least squares regression in predictor functions. These predictor functions seek to find a value for the *outcome variable* Y given *regressors* X. We begin with the best predictor of Y given X, the conditional expectation function:

Definition 2.1. The conditional expectation function, $m(X)$, represents the conditional expectation of a random variable Y, given a value of the random variable or vector X. We write:

$$m(x) = \mathbb{E}[Y|X = x]$$

We also introduce prediction error, which we will need in order to define optimal prediction functions. We can think of prediction error as the portion of Y unexplained by our regressors.

Definition 2.2. The prediction error, for a given predictor function $g(X)$ for Y, is a random variable representing the difference between the outcome variable and the respective predictor function, denoted:

$$e = Y - g(X)$$

If our goal is to minimize the mean squared prediction error, a function of our prediction error, we can prove that the conditional expectation function is the best predictor of Y.

Proposition 2.3. *The conditional expectation function, $m(X)$, is the best predictor of Y, in that it minimizes mean squared prediction error.*

Proof. We want to minimize the mean squared prediction error for a given function, $g(X)$[1]:

$$\mathbb{E}\left[(Y - g(X))^2\right]$$

We can plug in $m(X) + e_m$ for Y, where e_m represents the prediction error for the conditional expectation function:

$$\begin{aligned}
\mathbb{E}\left[(Y - g(X))^2\right] &= \mathbb{E}\left[(m(X) + e_m - g(X)^2\right] \\
&= \mathbb{E}\left[((m(X) - g(X)) + e_m)^2\right] \\
&= \mathbb{E}\left[(m(X) - g(X))^2 + 2e_m(m(X) - g(X)) + e_m^2\right] \\
&= \mathbb{E}\left[(m(X) - g(X))^2\right] + 2\mathbb{E}\left[e_m(m(X) - g(X))\right] + \mathbb{E}\left[e_m^2\right] \\
&\geq 2\mathbb{E}\left[e_m(m(X) - g(X))\right] + \mathbb{E}\left[e_m^2\right]
\end{aligned}$$

To complete this proof, we use the following lemma:

Lemma 2.4. *For any function $h(x)$ where $\mathbb{E}\left[e_m h(X)\right] < \infty$, then $\mathbb{E}\left[e_m h(X)\right] = 0$.*

[1]Credit to Bruce Hansen's *Econometrics* for the outline of this proof

Proof.

$$\begin{aligned}\mathbb{E}\left[e_m h(X)\right] &= \mathbb{E}\left[(Y - m(X))h(X)\right] \\ &= \mathbb{E}\left[Yh(X) - m(X)h(X)\right] \\ &= \mathbb{E}\left[Yh(X)\right] - \mathbb{E}\left[m(X)h(X)\right] \\ &= \mathbb{E}\left[Yh(X)\right] - \mathbb{E}\left[\mathbb{E}\left[Y|X\right]h(X)\right] \\ &= \mathbb{E}\left[Yh(X)\right] - \mathbb{E}\left[Yh(X)\right] \\ &= 0\end{aligned}$$

The fifth equality holds by *Theorem 2.7* from [2]. □

Therefore, we find:

$$\mathbb{E}\left[(Y - g(X))^2\right] \geq \mathbb{E}\left[e_m^2\right]$$

We also find the mean squared prediction error for $m(X)$:

$$\mathbb{E}\left[(Y - m(X))^2\right] = \mathbb{E}\left[e_m^2\right]$$

Because the mean squared prediction error for $m(X)$ equals the lower bound for the mean squared prediction errors of any $g(X)$, $m(X)$ is the best predictor for Y. □

While the conditional expectation function is the best predictor for Y, we cannot find an expression for such a function. Therefore, while the conditional expectation function provides the theoretical background for regression, it is not useful in applications. We find a useful definition with the best linear predictor:

Definition 2.5. The best linear predictor of Y given X, denoted $\mathscr{P}(Y|X)$, is the function in the form $X'\beta$ with the lowest mean squared prediction error.[2] We call β the linear prediction coefficient.

Consider X and β to be $k \times 1$ column vectors, where k is the total number of regressors. We find an expansion for $X'\beta$:

$$X'\beta = \sum_{j=1}^{k} X_j \beta_j$$

where each regressor X_j is a random variable. When we eventually consider observations, we can think of X_j as a vector. We provide geometric intuition in the section 2.4.

Now we introduce some important assumptions necessary for us to derive an expression for the linear prediction coefficient.

Assumptions 2.1.

(1) $\mathbb{E}[Y^2] < \infty$

(2) $\mathbb{E}\left[\|X\|^2\right] < \infty$

(3) $\mathbb{E}[XX']$ *is positive definite.*

[2]Note that X' denotes the transpose of X

(4) $\mathbb{E}[Xe] = 0$

The first assumption implies the finite variance and expectation of X. Note that the second assumption, where $\|X\|$ denotes the Euclidean norm of X, applies assumption 2.1.1 to the random variables X_j in X.

$$\begin{aligned}\mathbb{E}\|X\|^2 &= \mathbb{E}\left[\sum_{j=1}^{k} X_j^2\right] \\ &= \sum_{j=1}^{k} \mathbb{E}\left[X_j^2\right] < \infty\end{aligned}$$

Therefore, we find:

$$\mathbb{E}\left[X_j^2\right] < \infty \text{ for } j \in [k]$$

We find that, as a consequence of the finiteness of squared expectation, the variance of random variables is finite.

Proposition 2.6. *If a random variable X has an expectation such that:*

$$\mathbb{E}\left[X^2\right] < \infty$$

Then the variance of the random variable is finite.

Proof. First, we want to show that $\mathbb{E}[X] < \infty$. We consider the variance of random variable X

$$\begin{aligned}var(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X\mathbb{E}[X]] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2\end{aligned}$$

Since variance is a squared expectation, it is non-negative. Therefore:

$$0 \leq (\mathbb{E}[X])^2 \leq \mathbb{E}\left[X^2\right] < \infty$$

Therefore, not only is the expectation of X finite, but the variance as well. □

Note that assumption 2.1.3 ensures that $\mathbb{E}[XX']$ is invertible. A positive definite matrix has only positive eigenvalues, and therefore only non-zero eigenvalues, meaning that it is invertible. Finally, assumption 2.1.4 ensures that the regressors are uncorrelated with the error terms, or that the regressors only affect the outcome directly, and not through any unobservable factors represented by the error term.

From here, we derive an expression for β:

Proposition 2.7. *In the best linear predictor model, the linear prediction coefficient $\beta^* = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$ minimizes the mean squared prediction error.*

Proof. Applying a common strategy in econometric proofs[3], we find a lower bound for the mean squared prediction error by considering an arbitrary β. Note that we define the prediction error $e^* = Y - X'\beta^*$.

$$\begin{aligned}
\mathbb{E}\left[(Y - X'\beta)^2\right] &= \mathbb{E}\left[(X'\beta^* + e^* - X'\beta)^2\right] \\
&= \mathbb{E}\left[((X'\beta^* - X'\beta) + e^*)^2\right] \\
&= \mathbb{E}\left[(X'\beta^* - X'\beta)^2 + 2e^*(X'\beta^* - X'\beta) + (e^*)^2\right] \\
&= \mathbb{E}\left[(X'\beta^* - X'\beta)^2\right] + 2\mathbb{E}\left[e^*(X'\beta^* - X'\beta)\right] + \mathbb{E}\left[(e^*)^2\right] \\
&= \mathbb{E}\left[(X'(\beta^* - \beta))^2\right] + 2\mathbb{E}\left[e^*X'(\beta^* - \beta)\right] + \mathbb{E}\left[(e^*)^2\right] \\
&= \mathbb{E}\left[(X'(\beta^* - \beta))(X'(\beta^* - \beta))\right] + 2\mathbb{E}\left[e^*X'(\beta^* - \beta)\right] + \mathbb{E}\left[(e^*)^2\right] \\
&= \mathbb{E}\left[((\beta^* - \beta)'X)(X'(\beta^* - \beta))\right] + 2\mathbb{E}\left[e^*(\beta^* - \beta)'X\right] + \mathbb{E}\left[(e^*)^2\right] \\
&= (\beta^* - \beta)'\mathbb{E}\left[XX'\right](\beta^* - \beta) + 2(\beta^* - \beta)'\mathbb{E}\left[e^*X\right] + \mathbb{E}\left[(e^*)^2\right]
\end{aligned}$$

Here, we note the following:

$$(\beta^* - \beta)'\mathbb{E}\left[XX'\right](\beta^* - \beta) > 0$$

$$\mathbb{E}\left[e^*X\right] = \mathbb{E}\left[Xe^*\right] = 0$$

The former follows from assumption 2.1.3, which is that $\mathbb{E}\left[XX'\right]$ is positive definite. The former assumes $\beta \neq \beta^*$. Obviously, when $\beta = \beta^*$ we have equality. The latter follows by substituting e^* with $Y - X'\beta^*$, substituting β^* with $\mathbb{E}\left[XX'\right]^{-1}\mathbb{E}\left[XY\right]$ and applying the linearity of expectations. As a corollary, if we include a constant intercept regressor, the norm in regression, we also take $\mathbb{E}[e^*] = 0$. This result also satisfies assumption 2.1.4. With these results, we find a lower bound for the mean squared prediction error for all linear models:

$$\mathbb{E}\left[(Y - X'\beta)^2\right] > \mathbb{E}\left[(e^*)^2\right] \text{ for all } \beta \neq \beta^*$$

Finally, we find the mean squared error for β^*, by substituting $Y - X'\beta^*$ with e:

$$\mathbb{E}\left[(Y - X'\beta^*)^2\right] = \mathbb{E}\left[(e^*)^2\right]$$

Therefore, we find that β^* is the coefficient for the best linear predictor. □

Note that while $\beta^* = \beta$ under assumption 2.1, we should not think of β^* as part of the structural form of the linear model.

$$\begin{aligned}
\beta^* &= \mathbb{E}\left[XX'\right]^{-1}\mathbb{E}\left[XY\right] \\
&= \mathbb{E}\left[XX'\right]^{-1}\mathbb{E}\left[XX'\beta + Xe\right] \\
&= \beta + \mathbb{E}\left[XX'\right]^{-1}\mathbb{E}\left[Xe\right] = \beta
\end{aligned} \tag{2.8}$$

Indeed, we need $\mathbb{E}\left[Xe\right] = 0$ in the structural model for β^* to equal β.

[3]For a calculus-based proof of this result, I recommend section 2.18 of [2]

With these properties of the best linear predictor established, we introduce the heteroskedasticity and homoskedasticity assumptions.

2.3. **Assumptions on Error Covariance.** Consider the conditional variance of our prediction error e for the best linear predictor, where we take the variance of e given X:

Definition 2.9. Assuming the finiteness of $\mathbb{E}\left[e^2\right]$, the conditional variance of e given X is denoted:

$$\sigma^2(x) = var\left[e \mid X = x\right]$$

We can simplify our above expression for conditional variance as follows:

$$\begin{aligned}
\sigma^2(x) &= var\left[e \mid X = x\right] \\
&= \mathbb{E}\left[\left(e - \mathbb{E}\left[e \mid X = x\right]\right)^2 \mid X = x\right] \\
&= \mathbb{E}\left[e^2 - 2e\mathbb{E}\left[e \mid X = x\right] + \mathbb{E}\left[e \mid X = x\right]^2 \mid X = x\right] \\
&= \mathbb{E}\left[e^2 \mid X = x\right] - \mathbb{E}\left[2e\mathbb{E}\left[e \mid X = x\right] \mid X = x\right] + \mathbb{E}\left[\mathbb{E}\left[e \mid X = x\right]^2 \mid X = x\right] \\
&= \mathbb{E}\left[e^2 \mid X = x\right] - \left(\mathbb{E}\left[e \mid X = x\right]\right)^2 \\
&= \mathbb{E}\left[e^2 \mid X = x\right]
\end{aligned}$$

We find that $\sigma^2(x)$ is the conditional expectation of squared prediction error given a value of our regressors $X = x$. Hence, we introduce two assumptions on conditional variance:

Definition 2.10. We define a model as homoskedastic when the conditional expectation of the squared prediction error is not dependent on X. That is:

$$\sigma^2(x) = \sigma^2 \text{ for all } x$$

Where σ^2 is a constant.

Definition 2.11. We define a model as heteroskedastic when the conditional expectation of the squared prediction error depends on X. That is, a model is heteroskedastic if and only if it is not homoskedastic.

We can think of heteroskedastic models as models which are variably uncertain depending on our value of X, or rather, as models whose regressors are variably informative depending on their values. An example of a heteroskedastic linear model could be a line fitted to a financial time series spanning periods of high and low volatility, covering both financial crises and uneventful trading days.

2.4. **Least Squares Estimator.** We derive an estimator for the linear predictor coefficient β established in the previous section using geometric intuition. In the previous section, we took

$$\beta = \mathbb{E}\left[XX'\right]\mathbb{E}\left[XY\right].$$

However, we do not know the underlying distributions of X and Y. Therefore, we base an estimate of β on a random sample of n observations.

Definition 2.12. We define the sample as the set

$$\{(Y_i, X_i) \mid i \in [n]\}$$

where i represents an individual sample. X_i is a single observation of k regressors, Y_i is a single observation of the outcome variable Y, and n is the sample size.

Assumption 2.13. The pairs (Y_i, X_i) are independent and identically distributed between values of i.

This assumption may seem technical, but states that observations are taken from a common population. We consider individual observations from a population to be random variables with the same distribution as the population.

From here, we can begin to specify the least squares estimator $\hat{\beta}$. First, however, we define the linear prediction model in updated, matrix form. Note that while we use largely the same notation as that used in the best linear predictor, such as Y and X, we now denote using these variables matrices of observations, as opposed to random variables.

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} \quad \hat{e} = \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \vdots \\ \hat{e}_n \end{pmatrix} \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

We note that Y is an $n \times 1$ column vector containing observations of Y. Additionally, each X_i' in X represents a transposed column vector containing the i^{th} observation of each regressor. Hence, X is an $n \times k$ matrix. $\hat{e}$ represents an $n \times 1$ column vector containing all residuals $Y_i - X_i'\hat{\beta}$. Finally, $\hat{\beta}$ is the least squares estimator, which we define below.

Definition 2.14. Consider the linear prediction model

$$Y = X'\beta + e$$

where β is the linear prediction coefficient. We can estimate β with $\hat{\beta}$, which gives us an estimator for Y_i:

$$\hat{Y}_i = X_i'\hat{\beta}$$

The least squares estimator $\hat{\beta}$ is the $k \times 1$ column vector that minimizes the sum of squared errors, given below:

$$SSE(\beta) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - X_i'\beta)^2$$

We also calculate that:

$$X\hat{\beta} + \hat{e} = \begin{pmatrix} X_1'\hat{\beta} + \hat{e}_1 \\ X_2'\hat{\beta} + \hat{e}_2 \\ \vdots \\ X_n'\hat{\beta} + \hat{e}_n \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

The structural form for this regression, where we include n random variables representing observations, is as follows:

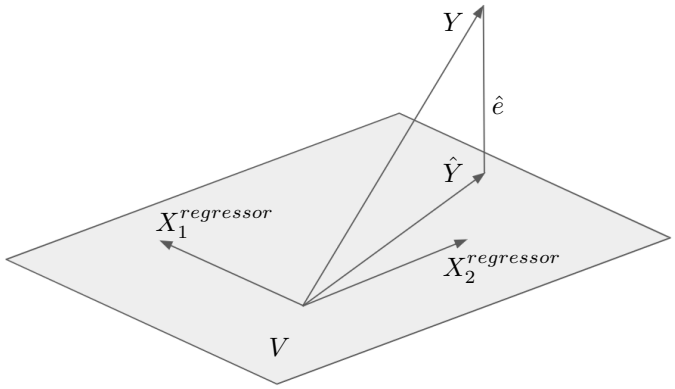
$$Y = X\beta + e \tag{2.15}$$

Now, consider the matrix X in the form of its columns, which represent all n observations for a given regressor.

$$X = \begin{pmatrix} \tilde{X}_1 & \tilde{X}_2 & \dots & \tilde{X}_k \end{pmatrix}$$

Therefore, we can think of $X\hat{\beta}$ as a linear combination of the regressor vectors, $\tilde{X}_j$ in $\mathbb{R}^n$. We denote by $\hat{Y}$ the projection of Y onto the space defined by $span(\tilde{X}_j \mid j \in [k])$, which we call V. We find that the orthogonal projection of Y to V minimizes the value $Y - \hat{Y}$. The diagram in figure 3 provides intuition with two regressors and three observations.

FIGURE 3. Orthogonal projection of Y to the space V



The above figure presents the two regressor vectors in $\mathbb{R}^3$, which represent our three observations for each regressor. We also consider the space generated by our regressors, being the plane V. We find that $\hat{Y} = X\hat{\beta}$ is in the space V. Y is not necessarily a linear combination of our regressors. Hence, Y is not necessarily contained in the space V. We minimize the sum of squared errors, and therefore minimize the norm of $\hat{e}$, squared:

$$SSE(\beta) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{e_i}^2 = \|\hat{e}\|^2$$

In order to minimize $\|\hat{e}\|^2$, we choose $\hat{Y}$ as the orthogonal projection of Y to V. Therefore, $\hat{e}$ is orthogonal to all $\tilde{X}_j$ for all j in $[k]$, meaning:

$$\langle \tilde{X}_j, \hat{e} \rangle = \tilde{X}_j{}' \hat{e} = 0 \tag{2.16}$$

Therefore, we find:

$$0 = X'\hat{e}$$

$$= X'(Y - \hat{Y})$$
$$= X'(Y - X\hat{\beta})$$
$$= X'Y - X'X\hat{\beta}$$

We are now close to our final expression for the least squares estimator.

$$X'X\hat{\beta} = X'Y$$

We find $X'Y$ to be a $k \times 1$ column vector, and $X'X$ to be a $k \times k$ square matrix. We can interpret the above expression as a system of k equations with k unknowns. We find a unique solution for our unknown $\hat{\beta}$ if and only if $X'X$ is invertible. $X'X$ is invertible if and only if X has k linearly independent columns, and $n > k$. In other words, we cannot have any redundant regressors. In this case, we find an expression for the least squares estimator:

$$\hat{\beta} = (X'X)^{-1}X'Y = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i Y_i \tag{2.17}$$

Finally, we define the projection matrix:

Definition 2.18. The projection matrix P is the linear transformation that represents the orthogonal projection of a given vector to the span of regressors, denoted V.

We find that the below expression satisfies the specified properties of P:

$$P = X(X'X)^{-1}X' \tag{2.19}$$

This matrix has the key property that when applied to a vector v, P orthogonally projects v to the span of X. We find a few important properties of projection matrices.

Proposition 2.20.

(1) *P is idempotent, meaning that $PP = P$.*

(2) $PY = X\hat{\beta} = \hat{Y}$

(3) $PX = X$

(4) *P is symmetric, meaning that $P' = P$*

The proofs of these propositions use straightforward algebra and can be solved as an exercise, but the geometric intuition they provide is crucial, particularly for two-stage least squares. Idempotence suggests that once a vector is orthogonally projected onto the space V, all future projections to V keep the result of the first projection. The second fulfills the property of the projection matrix. The third proposition is similar to idempotence. Since the columns of X represent the regressor vectors $X_j^{regressor}$, projecting such vectors onto their span will not change the vectors.

Most importantly, projection matrices serve as an algebraic way to represent the orthogonal projection in least squares regression. These are particularly important

when introducing multiple layers of least squares regression, as they allow for compact and purposeful representation of algebraic expressions.

2.5. **Generalized Least Squares.** Although it is beyond the scope of this paper to prove, we use the Generalized Gauss-Markov theorem to find a lower bound for the variance of all unbiased estimators. We can think of minimizing variance to find the best linear unbiased estimator (BLUE) as finding the unbiased estimator with the lowest sensitivity to changes in our observations for our regressors X.

Theorem 2.21 (Generalized Gauss Markov Theorem)**.** *In a linear model with independent and identically distributed observations, any unbiased estimator $\hat{\beta}^*$ for β will have a lower bound for variance*

$$var\left(\hat{\beta}^* \mid X\right) \geq \left(X'\Sigma^{-1}X\right)^{-1}$$

where Σ is the covariance matrix for our error vector e.

The OLS estimator's variance does not equal this lower bound. Only when we assume homoskedasticity[4] for our unbiased estimators does the variance of the OLS estimator attain a lower bound, as in the standard Gauss-Markov theorem. We therefore need to find the best linear unbiased estimator [5] (BLUE) regardless of homoskedasticity. This motivates generalized least squares: we premultiply the structural form of the now-heteroskedastic linear model by a matrix, creating a homoskedastic model. We apply ordinary least squares to the premultiplied model, deriving the generalized least squares estimator. We will find this matrix in the paragraphs below.

First, we consider the conditional covariance matrix for our prediction errors e:

$$var\left(e \mid X\right) = \mathbb{E}\left[(e - \mathbb{E}[e \mid X])(e - \mathbb{E}[e \mid X])' \mid X\right]$$

Since e is an $n \times 1$ column vector, the covariance matrix of e, which we will call Σ, is an $n \times n$ matrix, where $\Sigma_{ij} = cov(e_i, e_j|X)$. Now, consider our covariance matrix written as the product such that $\Sigma = \sigma^2\Omega$, where we pick $\sigma^2 = \mathbb{E}\left[e_1^2|X\right]$.

$$\Omega = \frac{var(e|X)}{\sigma^2} \tag{2.22}$$

Recall that $\mathbb{E}\left[e_i|X\right] = 0$ for all i in $[n]$. Hence, $cov(e_i, e_j|X) = \mathbb{E}\left[e_i e_j|X\right]$. Since our observations are independent, $\mathbb{E}[e_i e_j|X] = 0$ for $i \neq j$. Therefore, both Ω and Σ are diagonal matrices. In particular, $\Sigma_{ii} = \mathbb{E}\left[e_i^2|X\right]$. By independence, $\Sigma_{ii} = \mathbb{E}\left[e_i^2|X_i\right]$. In the case of homoskedasticity, where $\sigma^2(x) = \sigma^2$, we can write $\Sigma = \sigma^2 I_n$ and hence $\Omega = I_n$, or the identity matrix, since $\mathbb{E}[e_i^2|X] = \sigma^2$.

With Ω being the covariance matrix of error observations divided by a constant (2.22), we can consider $\Omega^{-1/2}$, such that $\Omega^{-1/2}\ \Omega^{-1/2} = \Omega^{-1}$. We know that we can take the inverse and square root of Ω since Ω is a diagonal covariance matrix with non-negative values. Note that because Ω is a diagonal matrix, $\Omega = \Omega'$. We consider a premultiplied version of regression:

$$\Omega^{-1/2}Y = \Omega^{-1/2}X\beta + \Omega^{-1/2}e$$

[4]Recall definition 2.10

[5]An unbiased estimator is an estimator with an expectation equal to the estimated parameter

Most importantly, this premultiplied version of regression is homoskedastic, meaning that OLS is the BLUE for all such premultiplied models.

$$\begin{aligned} var\left(\Omega^{-1/2}e \mid X\right) &= \mathbb{E}\left[\Omega^{-1/2}e\left(\Omega^{-1/2}e\right)' \mid X\right] \\ &= \Omega^{-1/2}\mathbb{E}\left[ee' \mid X\right]\Omega'^{-1/2} \\ &= \sigma^2\Omega^{-1/2}\Omega\Omega'^{-1/2} \\ &= \sigma^2\Omega^{-1/2}\Omega^{1/2}\Omega'^{1/2}\Omega'^{-1/2} \\ &= \sigma^2 I_n \end{aligned}$$

Our GLS estimator, therefore, is the OLS estimator for this premultiplied regression:

$$\begin{aligned} \hat{\beta}_{GLS} &= \hat{\beta}_{OLS} \\ &= \left((\Omega^{-1/2}X)'\Omega^{-1/2}X\right)^{-1}(\Omega^{-1/2}X)'\Omega^{-1/2}Y \\ &= \left(X'\Omega'^{-1/2}\Omega^{-1/2}X\right)^{-1}X'\Omega'^{-1/2}\Omega^{-1/2}Y \\ &= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}Y \end{aligned} \tag{2.23}$$

However, we make the practical consideration that we cannot compute Ω. As a result, GLS as we have derived here is a tool that we can use to derive more complex estimators, particularly in instrumental variables. We can also note that under homoskedasticity $\Omega = I_n$, so in this case the GLS estimator simplifies to the OLS estimator.

3. Regression with Endogeneity

One of the assumptions of the best linear predictor least squares regression is the uncorrelatedness of each regressor with the prediction error in the structural model. However, this assumption does not always hold, as we will show through examples. Instead, there exists the problem of *endogeneity*[6], where regressors may be correlated with prediction error. We introduce *instrumental variables* as a strategy to apply our regression toolkit to problems with endogenous regressors. This section will first cover technical definitions, then provide an example, and finally derive estimators.

3.1. Endogeneity.

Before introducing endogeneity, we recall assumption 2.1.4, stating that $\mathbb{E}[Xe] = 0$ in the linear model. The corrollary that $\mathbb{E}[X_j e] = 0$ implies that all regressors are uncorrelated with the prediction error. This means that all regressors contribute to the outcome variable Y directly, and not through some unobservable represented in the prediction error. However, we cannot always assume that $\mathbb{E}[Xe] = 0$. Instead, we define *exogeneity* and *endogeneity*:

Definition 3.1. A regressor for a structural model $Y = g(X) + e$ is exogenous if the following expectation holds:

$$\mathbb{E}[Xe] = 0$$

where X denotes the regressor and e denotes the prediction error.

[6]See definition 3.2

Definition 3.2. A regressor is endogenous if it is not exogenous, or:

$$\mathbb{E}[Xe] \neq 0$$

We recall the model for linear prediction with n observations, written:

$$Y = X\beta + e$$

We find that the least squares estimator, which converges to the structural parameter β as the sample size n approaches infinity under exogeneity, does not under endogeneity.

Proposition 3.3. *The least squares estimator $\hat{\beta}$ asymptotically approaches the structural parameter β as sample size n approaches infinity only under exogeneity.*

Proof. First, we show that the least squares estimator asymptotically approaches the previously-derived prediction coefficient $\beta^* = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$.

Lemma 3.4. *The least squares estimator $\hat{\beta}$ asymptotically approaches β^* as sample size n approaches infinity.*

Proof.

$$\begin{aligned}\hat{\beta} &= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i' Y_i = n\left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i\end{aligned}$$

By the weak law of large numbers, we find: $\hat{\beta} \underset{p}{\to} \mathbb{E}[XX']\mathbb{E}[XY] = \beta^*$ □

If we refer to (2.8), we can find that under endogeneity, $\beta^* \neq \beta$. □

It is important to note that β^* is not necessarily β, but rather an expression that we derive for β that minimizes the mean squared prediction error. Indeed, β^* is not a parameter, but a coefficient. Because we can no longer apply the least squares estimator under exogeneity, we instead introduce instrumental variables. Our goal is to generate new regressors by projecting our endogenous variables onto a set of exogenous, instrumental variables. We introduce the following requirements for the vector of instrumental variables.

Definition 3.5. We denote our $l \times 1$ vector of instrumental variables with Z. Our instruments Z must satisfy the following assumptions.

Assumptions 3.1.

(1) Z is exogenous

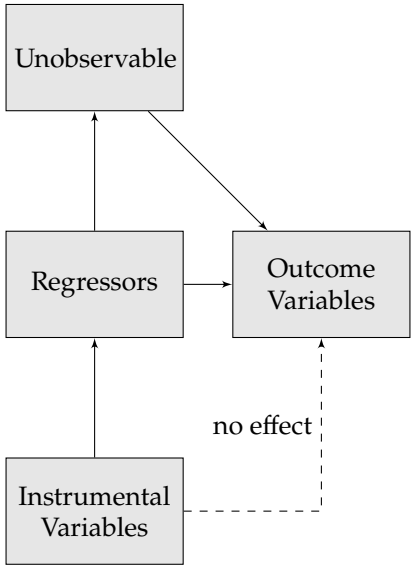
(2) $\mathbb{E}[ZZ']$ is positive definite

(3) $rank(\mathbb{E}[ZX']) = k$

Our first requirement ensures the exogeneity of the instrumental variables, which will assist us in deriving an estimator for β. Our second requirement ensures the invertibility of $\mathbb{E}[ZZ']$, and hence the linear independence of its columns. In other words, we exclude redundant instrumental variables. We call our third requirement the *relevance condition*. We note that since $\mathbb{E}[ZX']$ is an $l \times k$ matrix, the relevance condition requires $l \geq k$. In other words, we require at least as many instruments as regressors.

Assumption 3.1.3 implies that Z and X are correlated. Since $rank\left(\mathbb{E}[Z]\mathbb{E}[X']\right) = 1$ as a vector product, we find that $cov(Z, X') \neq 0$. We cannot have $k = 1$ since we always include a scalar intercept regressor. Assumption 3.1.1 implies the uncorrelatedness of our instruments and our error terms, meaning that our instruments affect our regressors alone. These two results are the principle requirements for instrumental variables. In one sentence, Z can only affect Y through X. The diagram below provides a visualization of this causal chain.

FIGURE 4. The causal chain between instruments, regressors, and the outcome variable.



3.2. **Solving Endogeneity: Acemoglu et al. 2001.** Recall the example from section 2.1. While OLS regression might make sense to understand the causal link between institutional strength and wealth in former colonies, consider the opposite. Institutions could be significant investments that only wealthy nations can afford to build and maintain. As a result, while OLS reveals the correlation between institutional strength and wealth, we cannot assert causation. This is where instrumental variables come into play. We require a potential instrumental variable to affect institutional strength, and also to affect wealth only through its effect on institutional strength. Acemoglu et al. propose using settler mortality from the colonial era as an instrument for modern institutional strength.

FIGURE 5. 2SLS regression finding causal link between institutional strength and GDP

2SLS Regression			
Regressor	Sample (1)	Sample (2)	Sample (3)
Average protection against expropriation risk, 1985-1995	0.92 (0.15)	1.10 (0.22)	1.16 (0.34)
Latitude	-	-	-0.75 (1.70)
GB colony dummy	-	-0.78 (0.35)	-0.80 (0.39)
FR colony dummy	-	-0.12 (0.35)	-0.06 (0.42)
Corresponding OLS Regression			
Average protection against expropriation risk, 1985-1995	0.53 (0.06)	0.53 (0.19)	0.47 (0.07)

Let us investigate the intuition behind this strategy. Acemoglu et al. argue that high settler mortality rates led to brutal and extractive colonies such as the Congo Free State. Hence, high settler mortality rates yield poor institutions. Meanwhile, low settler mortality rates yield higher settlement rates, imported early institutions, stronger modern institutions, and therefore modern wealth. Similarly, settler mortality rates from over a century ago are highly unlikely to impact national wealth today except through the causal chain established above. Therefore we find that the instrument of mortality rates is correlated with the endogenous regressor of institutional strength, and only affects the outcome of GDP through the endogenous regressor. As a result, settler mortality rates are an effective instrument for modern institutional strength. We can think of instrumental variables in the following way: if we can find the portion of our regressors attributable to our instruments, then we can find the portion of our regressors that affect our outcome directly, and not through unobservable criteria. As a result, we can infer causality between institutional strength and national wealth. One of the tools Acemoglu et al. use to do so is two-stage least squares (2SLS) regression. In 2SLS, we estimate our regressors in terms of our instruments and estimate our outcome in terms of these estimated regressors. Hence, we have two stages of regression. The outcome of Acemoglu et al. regression is in figure 5.

The outcome of Acemoglu et al. 2SLS regression suggests (with statistical significance) that institutional strength is causally tied to national wealth. Without 2SLS, researchers cannot argue causality, and hence their results are less convincing. Having established the technical definition and applications of endogeneity and instrumental variables, we can derive some of the estimators economists use to solve the endogeneity problem.

3.3. **Instrumental Variable Estimation.** We derive an estimator for β in the model

$$Y = X\beta + e$$

when we have an equal number of instruments and regressors.[7] We consider n observations, k regressors and l instruments. For now, we will assume that $l = k$. We therefore label Y as our $n \times 1$ column vector of outcome variable observations, X as our $n \times k$ matrix of regressors, Z as our $n \times l$ matrix of instruments, e as our $n \times 1$ column vector of errors, and β as a $k \times 1$ vector of weights for regressors. Before estimating β, we note that e is not equivalent to our residuals $\hat{e}$, since we have yet to take an estimate for Y in the form $X\hat{\beta}$. We can estimate β by premultiplying the structural model by a weighted version of our instruments. We can choose a weighting matrix R such that R is $l \times k$ with rank k. This rank condition ensures that we have no redundant weights. Because $l = k$, R is a square matrix. Hence, $R'Z'$ is a $k \times n$ matrix. We derive the so called *IV estimator* for β:

$$R'Z'Y = R'Z'X\beta + R'Z'e$$

Using the exogeneity of Z, we can approximate $R'Z'e$ by zero, such that we can derive our IV estimator for β:

$$R'Z'Y = R'Z'X\hat{\beta}_{IV}$$

And therefore:

$$\hat{\beta}_{IV} = (R'Z'X)^{-1} R'Z'Y$$

By the relevance condition, we can assume $Z'X$ has full rank. Since R has full rank, we take the inverse of R to find that our choice of R is arbitrary:

$$\begin{aligned}(R'Z'X)^{-1}R'Z'Y &= (Z'X)^{-1}(R')^{-1}R'Z'Y \\ &= (Z'X)^{-1}I_kZ'Y \\ &= (Z'X)^{-1}Z'Y\end{aligned}$$

Hence, we can simplify our IV estimator such that:

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y \tag{3.6}$$

The invertibility of $Z'X$ implies that both Z and X are full rank. In other words, we do not have redundant regressors or instruments.

The IV estimator makes the assumptions that the number of instruments matches the number of regressors. In order to generalize the IV estimator to include cases where the number of instruments l exceeds the number of regressors k, we introduce the two-stage least squares estimator.

3.4. **Two-Stage Least Squares.** Our derivation for the IV estimator relied on the assumption that we have as many instruments as regressors ($j = k$). The two-stage least squares (2SLS) estimator generalizes the IV estimator to cases where $j \geq k$. Just as in OLS where we can use more regressors to ensure apples-to-apples comparisons, economists stand to benefit from the added flexibility of a generalized IV estimator. More high quality instruments can provide a higher resolution view of our endogenous regressors. Therefore, we consider our structural model, but premultiplied by Z':

$$Z'Y = Z'X\beta + Z'e$$

[7]Note that we no longer take X transposed, since we are now considering observations for X.

β represents our unknowns. We note, therefore, that we have k unknowns but at least k equations, since $j \geq k$. We refer to the model where $j > k$ over-identified, and the model where $j = k$ just identified. In the case where $j > k$ therefore, we cannot immediately find a unique solution to estimate β. Here, we cannot approximate $Z'e$ by zero since there is no unique solution to our system of equations. Instead, our solution is the β which minimizes $Z'e$. Since the generalized least squares estimator is unbiased, it shares the error minimization properties of the best linear predictor. Like in generalized least squares, we take the covariance matrix of e to equal $\sigma^2\Omega$. Hence, for the $j \times 1$ column vector $Z'e$, we find the covariance matrix:

$$\begin{aligned} var(Z'e \mid Z'X) &= \mathbb{E}\left[(Z'e - \mathbb{E}[Z'e \mid Z'X])(Z'e - \mathbb{E}[Z'e \mid Z'X])' \mid Z'X\right] \\ &= \mathbb{E}\left[Z'e\,(Z'e)' \mid Z'X\right] \\ &= \mathbb{E}[Z'ee'Z \mid Z'X] \\ &= Z'\mathbb{E}[ee' \mid Z'X]\,Z \\ &= \sigma^2 Z'\Omega Z \end{aligned}$$

Considering the regression with regressors $Z'X$ and outcome variable $Z'Y$, we plug in our new covariance matrix into the GLS estimator (2.23):

$$\hat{\beta}_{GLS} = \left((Z'X)'(Z'\Omega Z)^{-1}Z'X\right)^{-1}(Z'X)'(Z'\Omega Z)^{-1}Z'Y$$

Simplifying, we find:

$$\hat{\beta}_{GLS} = \left(X'Z(Z'\Omega Z)^{-1}Z'X\right)^{-1}X'Z(Z'\Omega Z)^{-1}Z'Y$$

Here, assuming homoskedasticity, we take $\Omega = I_n$ to find the 2SLS estimator:

$$\hat{\beta}_{2SLS} = \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}X'Z(Z'Z)^{-1}Z'Y \tag{3.7}$$

Recall that the projection matrix P_z is eual to $Z(Z'Z)^{-1}Z'$. Hence, we can rewrite the 2SLS estimator as follows:

$$\begin{aligned} \hat{\beta}_{2SLS} &= (X'P_zX)^{-1}X'P_zY \\ &= (X'P_z'P_zX)^{-1}X'P_z'Y \\ &= ((P_zX)'P_zX)^{-1}(P_zX)'Y \end{aligned} \tag{3.8}$$

This simplified form of the 2SLS estimator is crucial to our understanding of the 2SLS estimator's function. 2SLS is the least squares regression of Y on the least squares estimation of X given instruments Z. In clearer terms, we first estimate our regressors in terms of our instruments. These estimates are exogenous. We can therefore take the least squares regression of Y on these estimates of X. This process of layering least squares regression yields the name two-stage least squares.

These derivations of the IV and 2SLS estimators were inspired by [6] and [7].

Acknowledgements

I would like to conclude by thanking my advisor, Colin Aitken, for his support throughout the research process, particularly in explaining many of the more

complex concepts in Econometrics. I would also like to express my gratitude to Professor Peter May, who organized this REU, and who very much made this paper on applied mathematics possible.

References

[1] Bruce Hansen. *Probability and Statistics for Economists*. Princeton University Press. 2021. https://www.ssc.wisc.edu/~bhansen/probability/Probability.pdf

[2] Bruce Hansen. *Econometrics*. Princeton University Press. 2021. https://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf

[3] Rebecca Willett. Least Squares and Geometry. In *STAT 27700: Mathematical Foundations of Machine Learning: Autumn 2020* [Lecture Recording]. https://uchicago.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=29842cc5-4e6b-4d2f-aa67-aae200f22401

[4] Daron Acemoglu, Simon Johnson, James Robinson. *The Colonial Origins of Comparative Development: An Empirical Investigation*. The American Economic Review. 2001. https://economics.mit.edu/files/4123

[5] Joshua Angrist, Jörn-Steffen Pischke. *Mastering Metrics: The Path from Cause to Effect*. Princeton University Press. 2015.

[6] Daniel McFadden. Instrumental Variables. In *ECON 240b: Econometrics: Spring 2010*. [Lecture Notes]. https://eml.berkeley.edu/~mcfadden/e240b_f01/ch4.pdf

[7] Menelaos Karanasos. Instrumental Variables and Two-Stage Least Squares. [Lecture Notes]. http://www.mkaranasos.com/EconomNot_GLS_IV.pdf?fbclid=IwAR3tunY4aRIpQNRWv0YAmt5kupFgdf7A679u0CHGo-_T4ih8_9SJkPD98Wc

[8] Sergei Treil. *Linear Algebra Done Wrong*. 2014. https://www.math.brown.edu/streil/papers/LADW/LADW-2014-09.pdf