The Linear Model Notes

Vasco Villas-Boas

April 6, 2025

1 Conditional Mean and Projection

Consider a model where Y is a dependent variable and $X := [X_1, ..., X_k]'$ are regressors with *joint pdf* $f_{Y,X}(y, x_1, ..., x_k)$. The joint pdf defines the *conditional pdf* $f_{Y|X}(y|x_1, ..., x_k) := \frac{f_{Y,X}(y,x_1,...,x_k)}{f_X(x_1,...,x_k)}$ provided that $f_X(x_1, ..., x_k) > 0$. We also define the *conditional expectation function*:

$$\begin{aligned} m(x) &:= \mathbb{E}[Y|X_1, ..., X_k] \\ &= \int_{-\infty}^{\infty} y f_{Y|X}(y|X_1, ..., X_k)\mathrm{d}y \end{aligned}$$

Lastly we also define the *regression error* $e := y - m(x) \implies y = m(x) + e$.

Assumption 1 (Finite Second Moments)**.** $\mathbb{E}[Y^2] < \infty,\ \mathbb{E}[X_j^2] < \infty.$

Proposition 1 (Mean-Independence and No Correlation)**.** Under Assumption 1, (i) $\mathbb{E}[e|X] = 0$ and (ii) $\mathbb{E}[h(X)e] = 0$ for any $h : \mathbb{R}^K \to \mathbb{R}$ such that $\mathbb{E}[(h(X))^2] < \infty$.

Proof (i).

$$\begin{aligned} \mathbb{E}[e|X] &= \mathbb{E}[Y - m(X)|X] \\ &= \mathbb{E}[Y|X] - m(X) \\ &= 0 \end{aligned}$$

Proof (ii).

$$\begin{aligned} \mathbb{E}[h(X)e] &= \mathbb{E}[\mathbb{E}[h(X)e|X]] \\ &= \mathbb{E}[h(X)\overset{0}{\cancel{\mathbb{E}[e|X]}}],\ \text{by } (i) \\ &= 0, \text{because } h(X) \text{ is square integrable and by Assumption 1 (ie., use Cauchy-Schwarz).} \end{aligned}$$

□

Note that by using proposition 1, for $h(X) = 1 \implies \mathbb{E}[e] = 0$ and for $h(X) = X \implies \mathbb{E}[Xe] = 0$.

1.1 Mean Squared Error

Let $g(X)$ be any predictor of Y with X. We define the mean squared error of the prediction, denoted MSE as

$$MSE := \mathbb{E}[(y - g(X))^2]$$

We also denote $MSE^* = \mathbb{E}[(Y - m(X))^2] = \mathbb{E}[e^2]$.

Theorem 2 (Best Predictor (MSE))**.** For any predictor *g(X),* $\mathbb{E}[(y - m(X))^2] \leq \mathbb{E}[(y - g(X))^2]$.

Proof.

$$\begin{aligned}
\mathbb{E}[(y - g(X))^2] &= \mathbb{E}[(m(X) + e - g(X))^2] \\
&= \mathbb{E}[(m(X) - g(X))^2] + \mathbb{E}[e^2] + 2\overset{0}{\cancel{\mathbb{E}[(m(X) - g(X))e]}}, \text{ by proposition 1} \\
&\geq 0 + MSE^* \\
&= MSE^*
\end{aligned}$$

□

1.2 Conditional Variance, Homoskedasticity, and Heteroskedasticity

We define the *conditional variance* of e as $\sigma^2(X) := \mathrm{Var}(e|X)$. If $\exists \sigma^2 \in \mathbb{R}_+$ such that $\forall X$ with $f_X(x_1, ..., X_k) > 0$, $\sigma^2(X) = \sigma^2$, then we say that e is *homoskedastic*. Otherwise, we say that e is *heteroskedastic*.

2 Linear Regression Model

Assume that

$$\begin{aligned}
m(X) &= \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \\
&= X'\beta
\end{aligned}$$

where $X := [1, X_1, ..., X_k]'$ and $\beta := [\beta_0, ..., \beta_k]'$. The linear regression model is characterized by:

Linear Regression Model

$$Y = X'\beta + e, \ \mathbb{E}[e|X] = 0$$

The linear regression model jointly consists of both the population regression equation and the population conditional mean equation. In fact, one can think that it's the population conditional mean equation that defines the coefficient in the population regression equation. When estimating β from data, we will use a sample analogue of the population conditional mean equation[1], to identify β. If we wish to use a linear model to predict Y from X in a 'centered sense', this is the model that we must assume. That comes from Proposition 1. By 'in a centered sense', I mean that, for any X, our prediction of Y given X is correct on average. In Figure 1, we see two examples where this population conditional assumption is clearly violated.

[1] Or more typically projection equations that are implied by the conditional mean equation as seen in Section 3.1

2.1 Best Linear Predictor/ Linear Projection

Assumption 2 (Non-Singularity of Population Design Matrix Outer-Product)**.** $\mathbb{E}[XX']$ is non-singular.

For any $b \in \mathbb{R}^{K+1}$, define

$$\begin{aligned} S(b) &:= \mathbb{E}[(Y - X'b)^2] \\ &= \mathbb{E}[Y^2] - 2b'E[XY] + b'\mathbb{E}[XX']b \end{aligned}$$

In the *best linear predictor* of Y given X, also known as the *linear projection* of Y on to X, given by $g(X) := X'\beta$, β solves the following problem

$$\beta := \text{argmin}_{b \in \mathbb{R}^{K+1}} S(b) \tag{1}$$

Proposition 3 (Linear Projection Existence and Uniqueness)**.** Under Assumptions 1 and 2, the projection $X'\beta$ exists and is unique. Further, define $\tilde{e} := Y - X'\beta$. Then, $\sigma^2 := \mathbb{E}[\tilde{e}^2] < \infty$, $\mathbb{E}[X\tilde{e}] = 0$, and $\text{Cov}(X, \tilde{e}) = 0$.

Proof.
When computing the FOCs of the problem in Equation (1), under Assumption 1, we can switch the order of the $\mathbb{E}[\cdot]$ and $\frac{\partial S(\cdot)}{\partial b}$ operators. Also, the objective is convex so that the first order conditions characterize β, giving existence and uniqueness. The first order conditions are

$$\begin{aligned} 0 &= \frac{\partial S(\beta)}{\partial b} \qquad (2) \\ &= -2\mathbb{E}[XY] + 2\mathbb{E}[XX']\beta \\ \implies \beta &= \mathbb{E}[XX']^{-1}\mathbb{E}[XY], \text{ under Assumption 2} \qquad (3) \end{aligned}$$

Next, since β is well-defined, that implies that $\tilde{e} = Y - X'\beta$ is also well-defined. Note that $\mathbb{E}[(X'\beta)^2] \leq \mathbb{E}[||X||^2]||\beta||^2$, by Cauchy-Schwarz. Then, note that $\mathbb{E}[||X||^2] = \sum_{j=0}^{K} E[X_j^2] < \infty$ since each term is also finite by Assumption 1. By transitivity, $\mathbb{E}[(X'\beta)^2] < \infty$.

Then, note that $\mathbb{E}[\tilde{e}^2] = \mathbb{E}[(Y - X'\beta)^2] \leq \mathbb{E}[Y^2] + \mathbb{E}[(X'\beta)^2] + 2\mathbb{E}[|Y(X'\beta)|] < \infty$ where the first term is finite by Assumption 1, the second by the previous paragraph, and the third by using another Cauchy-Schwarz argument. As a result, by transitivity, $\mathbb{E}[\tilde{e}^2] < \infty$.

Then, $\mathbb{E}[X\tilde{e}]$ is well-defined and finite by Cauchy-Schwarz. So,

$$\begin{aligned} \mathbb{E}[X\tilde{e}] &= \mathbb{E}[X(Y - X'\beta)] \\ &= \mathbb{E}[XY] - \mathbb{E}[XX']\beta \\ &= \mathbb{E}[XY] - \cancel{\mathbb{E}[XX']\mathbb{E}[XX']^{-1}}\mathbb{E}[XY] \\ &= \mathbb{E}[XY] - \mathbb{E}[XY] \\ &= 0 \end{aligned}$$

Finally, $\mathbb{E}[\tilde{e}] = 0$ since we assume that X contains a constant[2]. Then, $\text{Cov}(X, \tilde{e}) = \overset{0}{\cancel{\mathbb{E}[X\tilde{e}]}} - \mathbb{E}[X]\overset{0}{\cancel{\mathbb{E}[\tilde{e}]}} = 0$. □

As a remark, we call $X'\beta$ the projection of Y on X as the first order condition in Equation (2) can be rewritten as $\mathbb{E}[X(Y - X'\beta)] = 0 \iff \mathbb{E}[X\tilde{e}] = 0$. $X'\beta$ is the 'closest' linear function in the span of X to Y. That is characterized by the projection error $Y - X'\beta = \tilde{e}$ being orthogonal to each element of X.

[2] We see this in the FOC with respect to the constant term of β.

2.2 Linear Projection of Nonlinear $g(X)$

Suppose that we're looking for the linear projection of $g(X)$ onto X for some nonlinear function $g : \mathbb{R}^{K+1} \to \mathbb{R}$ that satisfies $\mathbb{E}[(g(X))^2] < \infty$. For any $b \in \mathbb{R}^{K+1}$, we can define,

$$\begin{aligned} d(b) &:= \mathbb{E}[(g(X) - X'b)^2] \\ &= \mathbb{E}[(g(X))^2] - 2b'\mathbb{E}[Xg(X)] + b'\mathbb{E}[XX']b \end{aligned}$$

We define the linear projection of $g(X)$ onto X by $X'\tilde{\beta}$ where $\tilde{\beta}$ solves

$$\tilde{\beta} := \text{argmin}_{b \in \mathbb{R}^{K+1}} d(b) \tag{4}$$

Again, by the normality conditions, we can switch the order of the operators $\mathbb{E}[\cdot]$ and $\frac{\partial d(\cdot)}{\partial b}$ when computing the first order conditions. Further, again, the objective is convex so that the first order conditions characterize the optimum. The first order conditions are

$$\begin{aligned} & 0 = -2\mathbb{E}[Xg(X)] + 2\mathbb{E}[XX']\tilde{\beta} \\ \implies & \tilde{\beta} = \mathbb{E}[XX']^{-1}\mathbb{E}[Xg(X)], \text{ taking Assumption 2 to hold} \end{aligned}$$

As a special case, if $g(X) = m(X)$, then

$$\begin{aligned} \tilde{\beta} &= \mathbb{E}[XX']^{-1}\mathbb{E}[Xm(X)] \\ &= \mathbb{E}[XX']^{-1}\mathbb{E}[X\mathbb{E}[Y|X]] \\ &= \mathbb{E}[XX']^{-1}\mathbb{E}[XY], \text{ where I used the law of iterated expectations} \\ &= \beta, \text{ as defined in Equation (3)} \end{aligned}$$

2.2.1 Example

Suppose that $g(X) = X^2$. One can show that for $X \sim U[-1, 1]$, the linear projection of $g(X)$ on to $[1, X]'$ is given by $Y = \frac{1}{3}$. Meanwhile, for $X \sim U[0, 1]$, the linear projection of $g(X)$ on to $[1, X]'$ is given by $Y = X - \frac{1}{6}$. These functions are plotted in Figure 1.

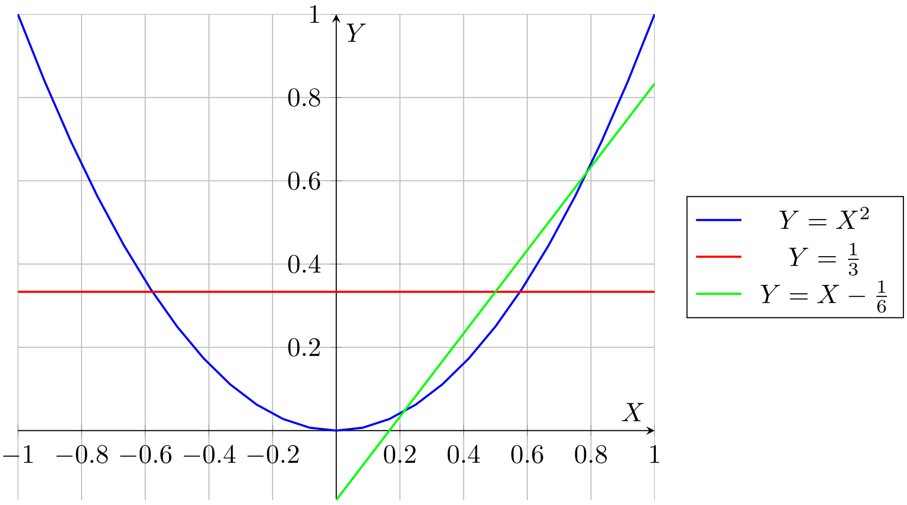

Figure 1. $g(X) = X^2$ Linear Projections

For didactic purposes, consider the projection of X^2 on to $[1, X]'$ when $X \sim U[-1, 1]$. We can define the projection error $\tilde{e} := X^2 - \frac{1}{3}$. By the first order conditions of the problem in Equation (4), we have that $\mathbb{E}[X\tilde{e}] = 0$ and $\mathbb{E}[\tilde{e}] = 0$. However, from the plot in Figure 1, it's clear that $\mathbb{E}[\tilde{e}|X] = X^2 - \frac{1}{3} \neq 0$ almost surely. Thus, while $\frac{1}{3} + (0)X$ is the linear projection of X^2 on to $[1, X]'$ when $X \sim U[-1, 1]$, that does not mean that it's a good predictor of X^2. If we use the linear projection $\frac{1}{3} + (0)X$ to predict X^2 using $[1, X]'$, we will systematically underestimate X^2 for some values of X and overestimate for others.

3 Linear Projection Model

The linear projection model is characterized by

Linear Projection Model

$$Y = X'\beta + e, \ \mathbb{E}[Xe] = 0$$

Again, the linear projection model consists of both the population regression equation and the population projection equation. One can think that it's the population projection equation that defines the coefficient in the population regression equation. Recall that a simple corollary of Proposition 1 is the linear regression model implies the linear projection model under Assumption 1 (ie., $\mathbb{E}[e|X] = 0 \implies \mathbb{E}[Xe] = 0$).

3.1 Estimation of the Projection Coefficient

We assume that we observe data $D := \{(x_n, y_n)\}_{n=1}^N$ from the linear projection model

$$y_i = x_i'\beta + \tilde{e}_i, \ \mathbb{E}[x_i\tilde{e}_i] = 0$$

where x_i implicitly contains a constant term. As for dimensions, we have that $x_i \in \mathbb{R}^{K+1}$, $y_i \in \mathbb{R}$, and $\beta \in \mathbb{R}^{K+1}$ for some $K \in \mathbb{N}$. I can stack the data D in the form

$$Y := \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{pmatrix}, \ X := \begin{pmatrix} x_1' \\ x_2' \\ \cdots \\ x_N' \end{pmatrix}, \ \tilde{e} := \begin{pmatrix} \tilde{e}_1 \\ \tilde{e}_2 \\ \cdots \\ \tilde{e}_N \end{pmatrix}, \ \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_K \end{pmatrix}$$

where $x_n = [1, x_{n1}, ..., x_{nk}]'$. As for dimensions, $Y, \tilde{e} \in \mathbb{R}^N, X \in \mathbb{R}^{N \times (K+1)}$, and $\beta \in \mathbb{R}^{K+1}$. I can then relate the introduced variables

$$Y = X\beta + \tilde{e}$$

Recall from Equation (1), that the linear projection coefficient $\beta = \text{argmin}_{b \in \mathbb{R}^{K+1}} \mathbb{E}[(y_i - x_i' b)^2]$. We can define the unnormalized empirical analogue

$$SSR_N(b) := \sum_{n=1}^{N} (y_n - x_n' b)^2 \tag{5}$$

and estimate β using this[3]

$$\begin{aligned} \hat{\beta}_N^{LS} &:= \text{argmin}_{b \in \mathbb{R}^{K+1}} SSR_N(b) & (6) \\ &= \text{argmin}_{b \in \mathbb{R}^{K+1}} \frac{1}{N} SSR_N(b) \\ &= \text{argmin}_{b \in \mathbb{R}^{K+1}} \frac{1}{N} \sum_{n=1}^{N} (y_n - x_n' b)^2 \\ &= \text{argmin}_{b \in \mathbb{R}^{K+1}} \frac{1}{N} (Y - Xb)'(Y - Xb) \\ &= \text{argmin}_{b \in \mathbb{R}^{K+1}} \frac{1}{N} [Y'Y - 2Y'Xb + b'X'Xb] & (7) \end{aligned}$$

The objective is convex so that the first order conditions characterize the optimum of the problem in Equation (7). The first order conditions of the problem are

$$\begin{aligned} 0 &= \left. \frac{\partial SSR_N(b)}{\partial b} \right|_{b = \hat{\beta}_N^{LS}} \\ &= -\frac{2}{N} X'Y + \frac{2}{N} X'X \hat{\beta}_N^{LS} & (8) \\ \implies \hat{\beta}_N^{LS} &= (X'X)^{-1} X'Y, \text{ when } \det(X'X) \neq 0 & (9) \end{aligned}$$

Note that this derivation is precisely analagous to the population derivation that terminates at Equation (3). As a second remark, for $\det(X'X) \neq 0$, we need X to have full column rank of $K + 1$. If it's the case that $N < K + 1$, then we don't have enough data to fit the model and we say the estimation suffers from *micronumerosity*. If it's the case that $N \geq K + 1$ but $\text{rank}(X) \neq K + 1$ (ie., X doesn't have full rank), then we say that the estimation suffers from *multicollinearity*.

We can also write estimator in Equation (9) in index form, introducing N by multiplying and dividing by the same number, as

[3] Note that estimating β by minimizing the empirical analogue of $\mathbb{E}[(y_i - x_i' b)^2]$ (ie., $\frac{1}{N} \sum_{n=1}^{N} (y_n - x_n' b)^2$) is equivalent to minimizing the unnormalized version (ie., $SSR_N(b)$).

$$\hat{\beta}_N^{LS} = (\frac{1}{N}\sum_{n=1}^{N} x_n x_n')^{-1}(\frac{1}{N}\sum_{n=1}^{N} x_n y_n)$$

Here, it's quite clear that β in Equation (3) is estimated using it's sample analogue. Next, we define the least squares fit $\hat{y}_n := x'_n\hat{\beta}$ and the least squares residual as $\hat{e}_n := y_n - \hat{y}_n$. In vector form, $\hat{Y} := [\hat{y}_1, ..., \hat{Y}_N]'$ and $\hat{e} := [\hat{e}_1, ..., \hat{e}_N]'$ so that $Y = \hat{Y} + \hat{e}$. Note that the empirical first order condition in Equation (8) is equivalent to

$$\begin{aligned} 0 &= \frac{1}{N}X'(Y - X\hat{\beta}_N^{LS}) \\ &= \frac{1}{N}X'\hat{e} \\ &= \frac{1}{N}\sum_{n=1}^{N} x_n\hat{e}_n \end{aligned}$$

which makes $X\beta$ the empirical analogue of the linear projection of Y on to X. $X\beta$ can be viewed as the linear projection of Y on to the column space of X. See Figure 2 for a visual representation of the various vectors.

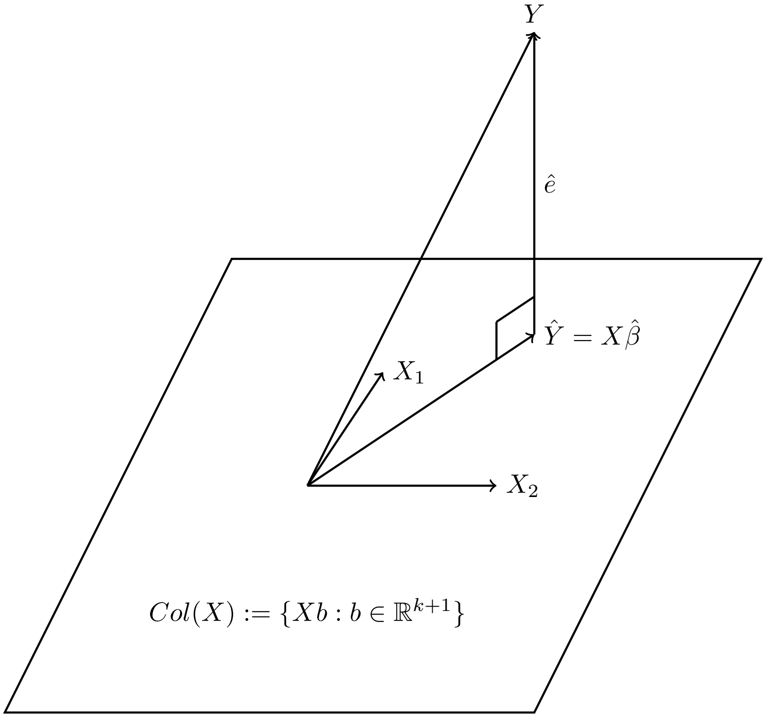

Figure 2. Projection Vectors Plot

3.2 Goodness of Fit (R^2)

In sample, for $Z_1, Z_2 \in \mathbb{R}^N$, define $\hat{\text{Cov}}_N(Z_1, Z_2) := \frac{1}{N}\sum_{n=1}^{N}(Z_{1n} - \bar{Z}_1)(Z_{2n} - \bar{Z}_2)$ where $\bar{Z}_j = \frac{1}{N}\sum_{n=1}^{N} Z_{jn}$ for $j \in \{1, 2\}$. Also, for $Z_1 \in \mathbb{R}^N$, define $\hat{\text{Var}}_N(Z_1) := \hat{\text{Cov}}_N(Z_1, Z_1)$. As a result,

$$\begin{aligned}
\hat{\text{Var}}_N(Y) &= \hat{\text{Var}}_N(\hat{Y} + \hat{e}) \\
&= \hat{\text{Var}}_N(\hat{y}) + 2\overset{0}{\cancel{\hat{\text{Cov}}_N(\hat{Y}, \hat{e})}} + \hat{\text{Var}}_N(\hat{e}), \text{ since } X\beta \text{ is the projection of } Y \text{ on to columns of } X \\
&= \hat{\text{Var}}_N(\hat{y}) + \hat{\text{Var}}_N(\hat{e})
\end{aligned}$$

We define the quantity R^2, which is a measure of goodness of fit as

$$\begin{aligned}
R^2 :=& \frac{\hat{\text{Var}}_N(\hat{Y})}{\hat{\text{Var}}_N(Y)} \\
=& \frac{\hat{\text{Var}}_N(Y) - \hat{\text{Var}}_N(\hat{e})}{\hat{\text{Var}}_N(Y)} \\
=& 1 - \frac{\hat{\text{Var}}_N(\hat{e})}{\hat{\text{Var}}_N(Y)}
\end{aligned} \tag{10}$$

There are a few different interpretations for this expression one of which I list here. The intuition for Equation (10) is that we're capturing the fraction of the empirical variance in the dependent variable (denominator) that we can explain by using the independent variables X (numerator).

R^2 as a measure of 'goodness of fit' does have its limitations. It's possible to artificially inflate the variance of Y and produce a higher R^2 if we can keep our residuals of similar magnitudes. R^2 also does not tell us whether our predictions are biased at any particular covariates X.

3.3 The Projection and Complementary Projection Matrices

Define the *projection matrix* $P_X := X(X'X)^{-1}X'$ and the *complementary projection matrix* $M_X := I_N - P_X$. As for dimensions, one sees that $P_X, M_X \in \mathbb{R}^{N \times N}$. One also sees that $\hat{Y} = P_X Y$ and $\hat{e} = M_X Y$. Now, I wish to list some useful properties of these matrices.

$$\begin{aligned}
\text{(Symmetry } P_X\text{) } P_X' &= (X(X'X)^{-1}X')' \\
&= X(X'X)^{-1}X' \\
&= P_X \\
\text{(Symmetry } M_X\text{) } M_X' &= (I_N - X(X'X)^{-1}X')' \\
&= I_N - X(X'X)^{-1}X' \\
&= M_X \\
\text{(Idempotence } P_X\text{) } P_X P_X &= X(X'X)^{-1}\cancel{X'X(X'X)^{-1}}X' \\
&= X(X'X)^{-1}X' \\
&= P_X \\
\text{(Idempotence } M_X\text{)} M_X M_X &= (I_N - P_X)(I_N - P_X) \\
&= I_N - 2P_X + P_X P_X \\
&= I_N - 2P_X + P_X \\
&= I_N - P_X \\
&= M_X \\
\text{(Trivial Projection) } P_X X &= X\cancel{(X'X)^{-1}X'X} \\
&= X
\end{aligned}$$

$$\begin{aligned}
\text{(Orthogonality) } M_X P_X &= (I_N - P_X)P_X \\
&= P_X - P_X P_X \\
&= 0 \\
\text{(Trace } P_X\text{) } \operatorname{tr}(P_X) &= \operatorname{tr}(X(X'X)^{-1}X') \\
&= \operatorname{tr}(\cancel{X'X(X'X)^{-1}}) \\
&= \operatorname{tr}(I_{K+1}) \\
&= K+1 \\
\text{(Trace } M_X\text{) } \operatorname{tr}(M_X) &= \operatorname{tr}(I_N - P_X) \\
&= N - (K+1) \\
\text{(Positive Semidefiniteness } P_X\text{) } a'P_X a &= a'P_X' P_X a \\
&= (P_X a)'(P_X a) \\
&\geq 0
\end{aligned}$$

For the trace properties, I use the fact that the trace operator produces identical values under circular rearrangements of the inputed matrices and that the trace operator is linear. One last property of the projection matrix P_X is that all of its eigenvalues are either 0 or 1. To see that, suppose that (λ, v) is an eigenpair of P_X. Then, $P_X v = \lambda v$. But, P_X is idempotent and λv is also an eigenpair of P_X, so that $P_X P_X v = \lambda^2 v \implies P_X v = \lambda^2 v \implies \lambda = \lambda^2 \implies \lambda \in \{0,1\}$.

4 Partitioned Regression and Frisch-Waugh-Lovell

Consider the model

$$\begin{aligned}
y_i &= x_i'\beta + e_i \\
&= x_{1i}'\beta_1 + x_{2i}'\beta_2 + e_i,\ \mathbb{E}[x_i e_i] = 0
\end{aligned}$$

where $x_i = [x_{1i}', x_{2i}']'$ and $\beta = [\beta_1', \beta_2']'$ Consider the following two procedures to compute an estimate $\hat{\beta}_1$ for β_1 using design matrix using data (X, Y) where $X = [X_1, X_2] \in \mathbb{R}^{N \times (K+1)}$ and $Y \in \mathbb{R}^{K+1}$.

Procedure 4 (Residual Regression)**.** Regress $(I_N - P_{X_2})Y$ on $(I - P_{X_2})X_1$ to identify $\hat{\beta}_1^{RR}$

Procedure 5 (Long Regression)**.** Regress Y on X and select out $\hat{\beta}_1^{LR}$

Theorem 6 (Frisch-Waugh-Lovell)**.** Procedure 4 and Procedure 5 produce the same estimate of $\hat{\beta}_1$.

Proof.
Consider first the output from the Long Regression $\hat{\beta}^{LR} = [\hat{\beta}_1^{LR\prime}, \hat{\beta}_2^{LR\prime}]' = (X'X)^{-1}X'Y$.
Define the in-sample residuals to be $\hat{e} := Y - X\hat{\beta}^{LR}$, I can rearrange to write that in sample

$$\begin{aligned}
Y &= X_1\hat{\beta}_1^{LR} + X_2\hat{\beta}_2^{LR} + \hat{e} \\
\implies (I_N - P_{X_2})Y &= (I_N - P_{X_2})X_1\hat{\beta}_1^{LR} + \overset{0}{\cancel{(I_N - P_{X_2})X_2}}\hat{\beta}_2^{LR} + (I_N - P_{X_2})\hat{e} \\
&= (I_N - P_{X_2})X_1\hat{\beta}_1^{LR} + I_n\hat{e} - \overset{0}{\cancel{P_{x_2}\hat{e}}} \\
&= (I_N - P_{X_2})X_1\hat{\beta}_1^{LR} + \hat{e} \\
\implies X_1'(I_N - P_{X_2})Y &= X_1'(I_N - P_{X_2})X_1\hat{\beta}_1^{LR} + \overset{0}{\cancel{X_1'\hat{e}}} \\
&= X_1'(I_N - P_{X_2})X_1\hat{\beta}_1^{LR} \\
\implies \hat{\beta}_1^{LR} &= (X_1'(I_N - P_{X_2})X_1)^{-1}(X_1'(I_N - P_{X_2})Y)
\end{aligned}$$

Next, consider the output from the Residual Regression. Define $\tilde{Y} := (I_N - P_{X_2})Y$ *and* $\tilde{X}_1 := (I_N - P_{X_2})X_1$. *Then,*

$$\begin{aligned}\hat{\beta}_1^{RR} &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} \\ &= (X_1'(I_N - P_{X_2})'(I_N - P_{X_2})X_1)^{-1}X_1'(I_N - P_{X_2})'(I_N - P_{X_2})Y \\ &= (X_1'(I_N - P_{X_2})X_1)^{-1}X_1'(I_N - P_{X_2})Y \\ &= \hat{\beta}_1^{LR}\end{aligned}$$

□

5 Least Squares as Maximum Likelihood Estimation

Assumption 3 (IID Gaussian Errors). $e_i|x_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Assume the linear regression model and Assumption 3, which jointly imply that

$$y_i|x_i \overset{iid}{\sim} \mathcal{N}(x_i'\beta, \sigma^2)$$

The log-likelihood of a single sample, where $\phi(\cdot)$ is the pdf of $\mathcal{N}(0,1)$ is

$$\begin{aligned}l(y_i|x_i;\beta,\sigma^2) :=& \log\left(\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right) \\ =& \log(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(y_i - x_i'\beta)^2}{2\sigma^2}\right)) \\ =& -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - x_i'\beta)^2}{2\sigma^2}\end{aligned}$$

Thus, observing $D := \{(y_n, x_n)\}_{n=1}^N$ iid samples from this population, I get that the sample log-likelihood is

$$\begin{aligned}\mathcal{L}_N(\beta,\sigma^2) :=& \sum_{n=1}^N l(y_n|x_n;\beta,\sigma^2) \\ =& -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^N(y_n - x_n'\beta)^2 \\ =& -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}SSR_N(\beta)\end{aligned}$$

where $SSR_N(\cdot)$ is defined in Equation (5). The *Maximum Likelihood Estimator* of (β, σ^2) is given by

$$(\hat{\beta}_N^{MLE}, \hat{\sigma}^2_{MLE,N}) := \text{argmax}_{(b,s^2)\in\mathbb{R}^{K+1}\times\mathbb{R}_+}\mathcal{L}_N(b, s^2) \tag{11}$$

The objective of the problem in Equation (11) is strictly concave so that the optimum is characterized by the first order conditions. From this maximization problem, we see that $\hat{\beta}_N^{MLE}$ is precisely equal to

$$\begin{aligned}\hat{\beta}_N^{MLE} &= \text{argmin}_{b\in\mathbb{R}^{K+1}} SSR_N(b)\\ &= \hat{\beta}_N^{LS}\end{aligned}$$

where the last step comes as a result of Equation (6). The first order condition with respect to σ^2 is

$$\begin{aligned} 0 &= \frac{-N}{2(2\pi)\hat{\sigma}^2_{MLE,N}}(2\pi) + \frac{1}{2(\hat{\sigma}^2_{MLE,N})^2}SSR_N(b)\bigg|_{b=\hat{\beta}_N^{MLE},s^2=\hat{\sigma}^2_{MLE,N}}\\ \implies \hat{\sigma}^2_{MLE,N} &= \frac{1}{N}SSR_N(\hat{\beta}_N^{MLE})\\ &= \frac{1}{N}\sum_{n=1}^{N}(y_n - x_n'\hat{\beta}_N^{MLE})^2\end{aligned}$$

As we will see later, this estimator for $\hat{\sigma}^2_{MLE,N}$ is in fact downward biased. An unbiased estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{N-K-1}\sum_{n=1}^{N}(y_n - x_n'\hat{\beta}_N^{MLE})^2 \tag{12}$$

6 Finite Sample Properties of Least Squares

Assumption 4 (Homoskedastic Errors)**.** $\text{Var}(e_i|x_i) = \sigma^2$ a.s.

We take as given Assumptions 1, 2, 4, the linear regression model, and assume that we observe $\{(x_n, y_n)\}_{n=1}^N$ iid samples from the population that has data-generating process $\mathcal{P}$ for which I will abuse notation and let it jointly refer to unconditional or conditional relevant distributions.

6.1 Conditional Unbiasedness of Least Squares Estimator

To show that $\hat{\beta}_N^{LS}$ is a conditionally unbiased estimator[4] of β (ie., $\mathbb{E}_{\mathcal{P}}[\hat{\beta}_N^{LS}|X] = \beta$), we first note that

$$\begin{aligned}\mathbb{E}_{\mathcal{P}}[(e_1,...,e_N)|X] &= (\mathbb{E}_{\mathcal{P}}[e_1|x_1,...,x_N],...,\mathbb{E}_{\mathcal{P}}[e_N|x_1,...,x_N])\\ &= (\mathbb{E}_{\mathcal{P}}[e_1|x_1],...,\mathbb{E}_{\mathcal{P}}[e_N|x_N]),\ \text{by independence of samples}\\ &= \mathbf{0}\end{aligned}$$

As a result,

$$\begin{aligned}\mathbb{E}_{\mathcal{P}}[\hat{\beta}_N^{LS}|X] &= \mathbb{E}_{\mathcal{P}}[(X'X)^{-1}X'Y|X]\\ &= \mathbb{E}_{\mathcal{P}}[(X'X)^{-1}X'(X\beta+e)|X]\\ &= \beta + \mathbb{E}_{\mathcal{P}}[(X'X)^{-1}X'e|X]\\ &= \beta + (X'X)^{-1}X'\cancelto{0}{\mathbb{E}_{\mathcal{P}}[e|X]},\ \text{by above}\\ &= \beta\end{aligned}$$

□

[4] We typically care about the *conditional* unbiasedness of an estimator rather than the unconditional unbiasedness. The former obviously implies the latter.

6.2 Conditional Variance of Least Squares Estimator

Assumption 5 (Heteroskedastic Errors)**.** $\mathrm{Var}(e_i|x_i) = \sigma_i^2$

We now temporarily relax Assumption 4 and that the data are independently drawn from the population. We define $D := \mathrm{Cov}(e|X) = \mathbb{E}_{\mathcal{P}}[ee'|x_1, ..., x_N]$.

As special cases, under independent data from the population and Assumption 4, $D = \sigma^2 I_N$. Also under independent data from the population and Assumption 5, $D = diag([\sigma_1^2, ..., \sigma_N^2]')$. In either special case, the off diagonal entries are 0 because under the assumption of independent samples from the population, for $i \neq j \in \{1, ..., N\}$,

$$\begin{aligned}\mathbb{E}_{\mathcal{P}}[e_i e_j|x_1, ..., x_n] &= \overset{0}{\cancel{\mathbb{E}_{\mathcal{P}}[e_i|x_i]}}\overset{0}{\cancel{\mathbb{E}_{\mathcal{P}}[e_j|x_j]}} \\ &= 0\end{aligned}$$

Finally[5],

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_N^{LS}|x_1, ..., x_n) &= \mathrm{Var}((X'X)^{-1}X'Y|x_1, ..., x_n) \\ &= \mathrm{Var}((X'X)^{-1}X'(X\beta + e)|x_1, ..., x_n) \\ &= \mathrm{Var}(\beta + (X'X)^{-1}X'e|x_1, ..., x_n) \\ &= (X'X)^{-1}X' \mathrm{Var}(e|x_1, ..., x_n)X(X'X)^{-1} \\ &= (X'X)^{-1}X'DX(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}, \text{ iff Assumption 4 holds}\end{aligned}$$

7 Least Squares Efficiency and Gauss-Markov

Definition 7 (Linear Estimator)**.** Consider a set of data $\{(x_i, y_i)\}_{i=1}^N$. An estimator $\tilde{\beta}_N$ is a linear estimator of y given x if it can be written as $\tilde{\beta}_N := A'y$ for some $A \in \mathbb{R}^{N \times K}$. In other words, it is a linear estimator if it is a linear function of the data. Note the difference between a *linear model* which is linear in parameters and what we have here which is a *linear estimator*.

Proposition 8. Consider a set of data $\{(x_i, y_i)\}_{i=1}^N$ that arises from the linear regression model with $\mathrm{Cov}(e|X) = D$. We make the additional assumption that $\mathbb{E}[e|X] = 0$ where $X = [x_1, ..., x_N]'$.. A linear estimator $\tilde{\beta}_N := A'y$ is *conditionally unbiased* as in Section 6.1 if and only if $A'X = I_K$ where $X := [x_1, x_2, ..., x_N]'$.

Proof.

$$\begin{aligned}\mathbb{E}[\tilde{\beta}_N|X] &= \mathbb{E}[A'y|X] \\ &= \mathbb{E}[A'(X\beta + e)|X] \\ &= A'X\beta + A'\overset{0}{\cancel{\mathbb{E}[e|X]}} \\ &= A'X\beta\end{aligned}$$

Now for conditional unbiasedness, it must be the case that $\mathbb{E}[\tilde{\beta}_N|X] = \beta$ for any $\beta \in \mathbb{R}^K$. As a result, it must be that $A'X = I_K$. □

Definition 9 (Efficiency)**.** An estimator $\hat{\beta}_N^1$ is *efficient* relative to an estimator β_N^2 if $[\mathrm{Var}(\hat{\beta}_N^2) - \mathrm{Var}(\hat{\beta}_N^1)]$ is positive semi-definite.

[5] In the fourth step of the derivation, I use the fact that for a random vector $Z \in \mathbb{R}^M$ and a non-stochastic matrix $A \in \mathbb{R}^{K \times M}$, $\mathrm{Var}(AZ) = A\,\mathrm{Var}(Z)A'$.

Claim 10 (Cholesky Decomposition)**.** Suppose that $D \in \mathbb{R}^{N \times N}$ is a positive definite matrix. Then, there exists a lower triangular matrix $C \in \mathbb{R}^{N \times N}$ such that $CC' = D$.

Theorem 11 (Gauss-Markov Theorem)**.** Consider a set of data $\{(x_i, y_i)\}_{i=1}^N$ with $X = [x_1, ..., x_N]'$. (i) Under Assumption 4, the linear regression model, and the additional assumption that $\mathbb{E}[e|X] = 0$, the best (most efficient) linear (conditionally) unbiased estimator is the Least Squares Estimator $\hat{\beta}_N^{LS}$. (ii) In the linear regression model with $\text{Cov}(e|X) = D \in \mathbb{R}^{N \times N}$, the best (most efficient) linear (conditionally) unbiased estimator is $\hat{\beta}_N^{GLS} := (X'D^{-1}X)^{-1}X'D^{-1}y$.

It suffices to prove (ii) as (i) is just a special case of (ii) with $D = \sigma^2 I_N$.

Proof (ii).
Consider $\hat{\beta}_N^{GLS} := (X'D^{-1}X)^{-1}X'D^{-1}y$. We have that

$$\begin{aligned}
\text{Var}(\hat{\beta}_N^{GLS}|X) &= \text{Var}(X'D^{-1}X)^{-1}X'D^{-1}y(X'D^{-1}X)^{-1}X'D^{-1}y)'|X) \\
&= \text{Var}(X'D^{-1}X)^{-1}X'D^{-1}yy'D^{-1}X(X'D^{-1}X)^{-1}|X) \\
&= (X'D^{-1}X)^{-1}X'D^{-1}\,\text{Var}((X\beta + e)(X\beta + e)'|X)D^{-1}X(X'D^{-1}X)^{-1} \\
&= (X'D^{-1}X)^{-1}X'D^{-1}\,\text{Var}(ee'|X)D^{-1}X(X'D^{-1}X)^{-1} \\
&= (X'D^{-1}X)^{-1}X'D^{-1}DD^{-1}X(X'D^{-1}X)^{-1} \\
&= (X'D^{-1}X)^{-1}
\end{aligned}$$

Now consider any other linear estimator (conditionally) unbiased estimator $\tilde{\beta}_N := A'y$. We have that

$$\begin{aligned}
\text{Var}(\tilde{\beta}_N|X) &= \text{Var}(A'y|X) \\
&= A'\,\text{Var}(y|X)A \\
&= A\,\text{Var}(X\beta + e|X)A' \\
&= ADA'
\end{aligned}$$

Now onto the proof of interest

$$\begin{aligned}
\text{Var}(\tilde{\beta}_N) - \text{Var}(\hat{\beta}_N^{GLS}) &= A'DA - (X'D^{-1}X)^{-1} \\
&= A'DA - A'X(X'D^{-1}X)^{-1}X'A, \text{ using the fact that } A'X \text{ is conditionally unbiased} \\
&= A'C\left[I_N - C^{-1}X(X'(CC')^{-1}X)^{-1}X'(C')^{-1}\right]C'A \\
&= A'C\left[I_N - C^{-1}X((C^{-1}X)'C^{-1}X)^{-1}(C^{-1}X)'\right]C'A \\
&= A'C\left[I_N - H(H'H)^{-1}H'\right]C'A, \text{ defining } H := C^{-1}X \\
&= A'CM_HC'A, \text{ noting } M_H = I_N - H(H'H)^{-1}H' \\
&= A'CM_H'M_HC'A, \text{ by complementary projection matrix idempotence and symmetry} \\
&= B'B, \text{ defining } B := M_HC'A \\
&\text{which is PSD}
\end{aligned}$$

□

8 Variance Estimation

Let's continue from the line of thought that terminated in Equation (12) where we were estimating the variance of a homoskedastic. We can write $\hat{\sigma}^2_{MLE,N} = \frac{1}{N}y'(I_N - P_X)y$ where $P_X = X'(X'X)^{-1}X$. In turn, we can check whether $\hat{\sigma}^2_{MLE,N}$ is (conditionally) unbiased:

$$
\begin{aligned}
\mathbb{E}[\hat{\sigma}^2_{MLE,N}|X] &= \frac{1}{N}\operatorname{tr}(\mathbb{E}[y'(I_N - P_X)y|X]) \\
&= \frac{1}{N}\mathbb{E}[\operatorname{tr}(y'(I_N - P_X)y)|X] \\
&= \frac{1}{N}\mathbb{E}[\operatorname{tr}(e'(I_N - P_X)e)|X] \\
&= \frac{1}{N}\mathbb{E}[\operatorname{tr}((I_N - P_X)ee')|X], \text{ since } \operatorname{tr}(\cdot) \text{ is invariant under circular permutations} \\
&= \frac{1}{N}\operatorname{tr}((I_N - P_X)\mathbb{E}[ee'|X]) \\
&= \frac{1}{N}\operatorname{tr}((I_N - P_X)I_N\sigma^2) \\
&= \frac{1}{N}\operatorname{tr}(I_N - P_X)\sigma^2 \\
&= \frac{N-K-1}{N}\sigma^2
\end{aligned}
$$

where we assume that X has K nontrivial features and 1 constant feature. As a result, we see that $\hat{\sigma}^2_{MLE,N}$ is downwards biased. To have a (conditionally) unbiased estimator of σ^2, we need to report $\frac{N}{N-K-1}\hat{\sigma}^2_{MLE,N}$.

9 Asymptotics of Least Squares

Definition 12 (Consistency of Estimator)**.** An estimator $\hat{\beta}_N$ is consistent for β if $\lim_{N\to\infty}\Pr(||\hat{\beta}_N - \beta|| > \epsilon) = 0\,\forall\epsilon > 0$. In other words, we have that $\hat{\beta}_N \xrightarrow{\mathbb{P}} \beta$.

Proposition 13 (Consistency of Least Squares Estimator)**.** Assume that we have N iid observations $\{(x_i, y_i)\}_{i=1}^N$ from the linear projection model. Suppose that Assumptions 1 and 2 so that we have finite second moments and a non-singular population design matrix outer-product. Let $Q := \mathbb{E}[X_iX_i']$. Then, $\hat{\beta}_N^{LS} \to \beta$.

Proof.
Using Cauchy-Schwarz, we first see that $\mathbb{E}[|X_{ij}X_{il}|] \leq \sqrt{\mathbb{E}[X_{ij}^2]\mathbb{E}[X_{il}^2]} < \infty$ since both squared moments are finite by Assumption 1. As a result, by Khimchin's LLN, we have that $\frac{1}{N}\sum_{i=1}^N x_ix_i' \xrightarrow{\mathbb{P}} \mathbb{E}[X_iX_i'] = Q$, which we assume to be non-singular by Assumption 2.

We can argue that $\mathbb{E}[|X_{ij}e_i|] < \infty$ using a similar argument so that $\frac{1}{N}\sum_{i=1}^N X_ie_i \xrightarrow{\mathbb{P}} \mathbb{E}[X_ie_i]$ by Khimchin's LLN. By the assumption of the linear projection model, we have that $\mathbb{E}[X_ie_i] = 0$. Then, we can now proceed to show the consistency of the least square estimator,

$$
\begin{aligned}
\hat{\beta}_N^{LS} &= \left(\frac{1}{N}\sum_{i=1}^N x_ix_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N x_iy_i\right) \\
&= \beta + \overset{\xrightarrow{\mathbb{P}}Q^{-1}}{\cancel{\left(\frac{1}{N}\sum_{i=1}^N x_ix_i'\right)^{-1}}}\overset{\xrightarrow{\mathbb{P}}0}{\cancel{\left(\frac{1}{N}\sum_{i=1}^N x_ie_i\right)}}, \text{ by the continuous mapping theorem} \\
&\xrightarrow{\mathbb{P}} \beta, \text{ by the continuous mapping theorem}
\end{aligned}
$$

□

Assumption 6 (Finite Fourth Moments)**.** $\mathbb{E}[Y^4] < \infty,\ \mathbb{E}[X_j^4] < \infty$.

Proposition 14 (Asymptotic Distribution of Least Squares Estimator)**.** Assume that we have N iid observations $\{(x_i, y_i)\}_{i=1}^N$ from the linear projection model. Suppose that Assumptions 1 and 2 so that we have finite second moments and a non-singular population design matrix outer-product. Let $Q := \mathbb{E}[X_i X_i']$. Additionally suppose Assumption 6 holds so that we assume that we have finite fourth moments. Also let $\Omega := \mathbb{E}[X_i X_i' e_i^2]$. Then, we have that $\sqrt{N}(\hat{\beta}_N^{LS} - \beta) \xrightarrow{d} \mathcal{N}(0, Q^{-1}\Omega Q^{-1})$. Further, under linear regression model and Assumption 4, we have that $\sqrt{N}(\hat{\beta}_N^{LS} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q^{-1})$.

Proof.
...

REFERENCES