Final Exam

Lucy Hackett

May 11, 2021

1 Identifying assumptions for regression

For each of the questions in this section provide a short answer and argument. Note the quality and concision of the argument matters much more than the answer!

1. Evaluate the truth of following statement: ''In the linear regression $y = X\beta + u$ the usual identifying assumption $\mathbb{E}(u|X) = 0$ implies $E(h(X)u) = 0$ for any function h satisfying some regularity conditions related to measurability."

   Answer For functions satisfying regularity conditions, this statement is true. Intuitively, because X provides no information about the expected value of u, no function of X is able to either. To the contrary, we would be able to gain information about u from an informationless random variable by transforming it, which is not possible. Formally,

   $\mathbb{E}(u|X) = 0 \Rightarrow \mathbb{E}(u|h(X)) = 0 \Rightarrow \mathbb{E}(h(X)u) = 0$

2. Consider the same linear regression $y = X\beta + u$, but now suppose an alternative identifying assumption $\mathbb{E}(X|u) = 0$. Construct a simple estimator based on this alternative. Compare the usual and alternative identifying assumptions; are they equivalent? Is one stronger than the other?

   Answer A simple moment estimator given by this condition is derived by:

   $$\mathbb{E}(X|u) = 0 \Rightarrow \mathbb{E}(u'X) = 0 \Rightarrow \mathbb{E}((y - X\beta)'X) = 0$$
   $$\mathbb{E}(y'X) = \beta'\mathbb{E}(X'X)$$

   which can be solved using the sample analog.

   The usual identifying assumption is $\mathbb{E}(u|X) = 0$; here, the condition goes the other direction, $\mathbb{E}(X|u) = 0$. Mathematically one is not stronger than the other in the sense that one does not imply the other. However, from a data generating point of view the alternative identifying assumption is more restrictive. Consider a regression model with an intercept, for example. If the disturbances had a non-zero mean, we could ''demean" the disturbances by subtracting the mean and adding it to the intercept. Then, $\mathbb{E}(u|X)$ is not restrictive (See Greene chapter 2 for a discussion of this).

   However, if we believe that each of the columns of X come from some real data-generating process, then $\mathbb{E}(X|u) = 0 \Rightarrow \mathbb{E}(X) = 0$ (by the law of iterated expectations) is restrictive. The columns of X may each come from a different random process, and "demeaning" X is not straightforward as before.

3. Suppose that $y = f(X) + u$ for some unknown but continuous function f. Suppose we want to use observed data on X to predict outcomes y, and seek a predictor $\hat{y}(X)$ which is ''best" in the sense that the mean squared prediction error $\mathbb{E}(y - \hat{y}(X)|X))^2$ is minimized. What can we say about $\hat{y}$ and its relation to the conditional expectation $E(y|X)$? Its relation to u?

   Answer First note that $\hat{y}(X) = \mathbb{E}(y|X)$ is the minimizer of the objective function:

$$\min_{\hat{y}(X)} \mathbb{E}(y - \hat{y}(X)|X))^2$$

   This implies that:

   $\mathbb{E}(\hat{y}(X)|X) = \mathbb{E}(y|X) = \mathbb{E}(f(X)|X) + \mathbb{E}(u|X) = f(X) + \mathbb{E}(u|X)$

   Thus under the identifying assumption $\mathbb{E}(u|X) = 0$, which in turn implies $\mathbb{E}(u|\hat{y}(X)) = 0$, $\hat{y}(X)$ unbiasedly estimates $f(X)$.

2 Omitted variables

You are asked to serve as a referee for a paper submitted to a top field journal. In the submitted paper the researcher uses a sample of size N to estimate a model

$$y = \alpha + \beta x + u$$

The coefficient β seems to be significantly different from zero, but the researcher is concerned about omitted variable bias, so they also estimate a variety of alternative specifications of the form

$$y = \alpha + \beta x + \gamma w + u$$

where w is one of a number of other variables that the researcher hypothesizes might have some effect on y as a way of testing the first model.

The researcher finds a particular variable w which enters the regression significantly, and so

(i) rejects the first model, concluding that the first estimate of β was in fact affected by omitted variable bias;

(ii) declares the augmented regression to be their ''preferred specification"; and

(iii) proceeds to construct standard t-statistics for β and γ as a way of proceeding with inference.

Peer reviews in economics usually include some "notes for the author." What might your notes say about the paper's approach to omitted variable bias? Comment specifically on each of (i), (ii), and (iii). Try to make your remarks critical yet constructive— what shortcomings do you see, and how might the author address these?

Answer

First I comment on their procedure for including extra control variables w one by one. If the researchers have reasons to believe that a vector of relevant variables W may be omitted, then their hypothesis is that $\mathbb{E}(u|X \cup W) = 0$, not $\mathbb{E}(u|X \cup w) = 0, w \in W$. Therefore I would suggest that they also run a specification where they include the entire vector W. Now, for their procedure:

(i) They reject the model based on the significance of the parameter associated with w. However, the fact that w has significant covariance with y does not imply that X is endogeneous

and consequently that β is biased. For example, if w is correlated with Y but orthogonal to X, then by the Frisch–Waugh–Lovell theorem, we know that the estimate of $\hat{\beta}$ will be unaffected.
I am also concerned that there may be other specifications in which some w' does not enter significantly, but whose presence significantly alters the estimate of $\hat{\beta}$. Even though γ in this case would not be signficant, the fact that $\hat{\beta}$ changes dramatically could be a sign of OVB.

(ii) Based on the above, I believe the author's criteria for choosing their preferred specification is flawed. They should choose their ''preferred specification'' based on the behavior of $\hat{\beta}$, and not on the significant of w (or a vector W). If the estimate of $\hat{\beta}$ is fairly consistent across specifications, then this choice may be based on theory or choosing an estimate that represents a middle ground or conservative estimate of the effect they are trying to measure. If the inclusion of a set of extra controls causes the estimate to change (regardless of the significance of W), this should probably be their preferred specification because it is the estimate that takes into account the possible confounding factors.

If the different $\hat{\beta}$'s vary widely, they may have a serious OVB problem, and their identification model may need to be rethought.

(iii) Once a preferred specification is chosen, the authors may construct t-statistics for any of the coefficients in the regression, though because it seems that the authors are primarily interested in conducting inference on β, these statistics may be irrelevant for γ. A more informative way to report their results may be to report the point estimates and t statistics for $\hat{\beta}$ from various specifications, so that the reader can judge if the value and significance of $\hat{\beta}$ appears stable when including additional controls, and which specification they find most convincing.

3 Breusch-Pagan Extended

Consider a linear regression of the form

$$y = \alpha + \beta x + u$$

with (y, x) both scalar random variables, where it is assumed that

(a.i) $\mathbb{E}(ux) = \mathbb{E}(u) = 0$

(a.ii) $\mathbb{E}(u^2|x) = \sigma^2$

1. The condition (a.i) is essentially untestable, but Breusch and Pagan (1979) argue that one can test (a.ii) via an auxiliary regression $\hat{u}^2 = c + dx + e$, where the $\hat{u}^2$ are the residuals from the first regression, and the test of (a.ii) then becomes a test of $H_0 : d = 0$. Explain both why (a.i) is untestable, and the logic of the test of (a.ii.)

   Answer

   (a.i) is untestable because it is an assumption in the construction of OLS. Computationally, we can see the untestability of this assumption with the following. Suppose we wanted to test (a.i), so we construct a sample analog for this object. Let $X = [\mathbb{1}_N \; x]$.

   $X'e = X'(y - Xb) = X'y - X'Xb = X'y - X'X(X'X)^{-1}X'y = X'y - X'y = 0$

   This object is zero always, so we cannot test the condition (a.i) in any way.

The logic of (a.ii) follows from intuition of conditional homoskedasticity. This condition intuitively states that u does not vary with x in any systematic way. Therefore, we should not be able to predict the value of u^2 with x, which in a regression setup would imply $d = 0$.

2. Use the two conditions (a.i) and (a.ii) to construct a GMM version of the Breusch-Pagan test.

   Answer My answer draws from Shreya Sarkar's formulation of a test for homoskedasticity on piazza @33_f2. (a.ii) can be transformed as follows (again, letting $X = [\mathbb{1}_N \; x]$)

$$\begin{aligned}\mathbb{E}(u^2|X) = \sigma^2 &\Rightarrow \mathbb{E}(u^2 - \sigma^2|X) = 0 \\ &\Rightarrow \mathbb{E}(X'(u^2 - \sigma^2)) = 0 \\ &\Rightarrow \mathbb{E}(X'(y - X'\tilde{\beta})^2 - X'\sigma^2) = 0\end{aligned}$$

   where $\tilde{\beta} = [\alpha \; \beta]'$. Stacking this condition with (a.i) gives a vector of 4 moment conditions:

$$\mathbb{E}g_i(\tilde{\beta}, \sigma^2) = \mathbb{E}\begin{bmatrix} X_i'(y_i - X_i'\tilde{\beta}) \\ X_i'(y_i - X_i'\tilde{\beta})^2 - X_i'\sigma^2 \end{bmatrix} = \mathbf{0}_4$$

   Therefore:

$$g_N(\tilde{\beta}, \sigma^2) = \begin{bmatrix} \frac{1}{n}\sum_i^N X_i'(y_i - X_i'\tilde{\beta}) \\ \frac{1}{n}\sum_i^N X_i'(y_i - X_i'\tilde{\beta})^2 - X_i'\sigma^2 \end{bmatrix}$$

   and the objective function to minimize by GMM with a generic weighting matrix W is given by:

$$J(\tilde{\beta}, \sigma^2) = N g_N(\tilde{\beta}, \sigma^2)' W g_N(\tilde{\beta}, \sigma^2)$$

   So that the GMM test for homoskedasticity would be based on the statistic $J(\tilde{\beta}_{GMM}, \sigma^2_{GMM}) \sim \chi^2_1$

3. What can you say about the performance or relative merits of the Bruesch-Pagan test versus your GMM alternative?

   Answer One downside to the GMM test for homoskedasticity is that the test uses two distinct types of moments: moments related to the exogeneity of x, and moments related to homoskedasticity. Therefore, if the null hypothesis is rejected, we may not know if it was because of homoskedasticity or because of the exogeneity condition. However, the GMM test is more flexible than the Breusch-Pagan test, which assumes normality of the errors and, in this formulation, that any heterskedasticity is linear in x.

   I explore this in the notebook $\text{BP}_G MM, where I compare the J statistics for a data process where X is endogeneous versus$

4. Suppose that in fact that x is distributed uniformly over the interval $[0, 2\pi]$, and $E(u^2|x) = \sigma^2(x) = \sigma^2 + \sin(x)$, thus violating (a.ii). What can you say about the performance of the Breusch-Pagan test in this circumstance? Can you modify your GMM test to provide a superior alternative?

   Answer

   The Breusch-Pagan test will not perform will in this circumstance, and will be prone to accepting the null hypothesis of homoskedasticity when it should be rejected. This is because $\sin(x)$ cycles, so the net effect of x on $\hat{u}^2$ will tend to give an estimate of $b = 0$.

If we suspected that the variance of the disturbances took that form, we could directly incorporate that information into the moment conditions, replacing our homoskedasticity moments with:

$$\mathbb{E}(X_i' u_i^2 - X_i'(\sigma^2 + \sin(X_i))$$

5. In the above, we've considered a test of a specific functional form for the variance of u. Suppose instead that we don't have any prior information regarding the form of $\mathbb{E}(u^2|x) = f(x)$. Discuss how you might go about constructing an extended version of the Breusch-Pagan test which tests for $f(x)$ non-constant.

   Answer

   In a similar vein to the previous proposal, if we suspect that $\mathbb{E}(u^2|x) = f(x)$, we may want to approximate this non-linear relationship with basis functions. For example, on the interval $[0, 2\pi]$, the function $\sigma^2 \sin(x)$ may be reasonably approximated by a cubic basis function. Therefore we could modify the Breusch-Pagan test to be:

$$\hat{u}^2 = c + d_1 x + d_2 x^2 + d_3 x^3 + e$$

   with the joint hypothesis

$$H_0 : \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

6. Show that you can use your ideas about estimating $f(x)$ to construct a more efficient estimator of β if $f(x)$ isn't constant. Relate your estimator to the optimal generalized least squares (GLS) estimator.

   Answer Using the estimated basis function from the previous part, we can use $\hat{f}(x)$ to estimate GLS, which will yield

$$b_{GLS} = (X'\hat{f}(X)^{-1}X)^{-1}X'\hat{f}(X)^{-1}y$$

4 Black Lives Matter

Fryer Jr (2019) uses data on encounters between police and civilians of different races in the US to explore how police use of force is related to a civilian's race. While Fryer finds that Black and Hispanic civilians are much more likely to ''experience some form of force" from the police and while the probability of being shot by the police is much higher for a civilian who is Black or Hispanic, Fryer's most prominent result is that for ''the most extreme use of force— officer-involved shootings— we find no racial differences either in the raw data or when contextual factors are taken into account."

Introducing some notation, let R denote the civilian's race; U some variables observed by the police officer prior to any interaction (e.g., observing ''suspicious" behavior) but not the econometrician; D a binary variable indicating the event ($D = 1$) of an encounter between a given civilian and a police officer; V a set of ''contextual factors" related to the encounter and reported by the officer; and S the event that the civilian is shot by the officer. We can then express Fry's finding regarding shootings as not being able to reject either

$$Pr(S|D = 1, R) = Pr(S|D = 1) \tag{1}$$

or

$$Pr(S|D = 1, V, R) = Pr(S|D = 1, V) \tag{2}$$

1. Durlauf and Heckman (2020) criticize this conclusion of Fryer's, on the grounds that D may be an endogenous variable. You needn't read their paper, but explain in your own words what sorts of endogeneity might undermine Fryer's conclusion that the probability of being shot by the police doesn't depend on race.

   Answer

   If officers racially profile civilians, whether because they are personally prejudiced, unconsciously biased, or are following official guidelines that more heavily police minority neighborhoods, then they may interact more with innocent black civilians. That means that racial profiling is positively correlated with the probability of having an interaction with a police officer. If innocent people are more likely to comply with police officers (for example, less likely to flee from officers) then racial profiling may be negatively correlated with the probability of being shot by an officer. i.e., racial profiling is clearly positively correlated with $P(D|R = black)$ and negatively correlated with S. This omitted variable therefore would bias the estimate of the marginal effect of R on S downwards downwards, which would throw doubt on our ability to estimate (1).

   A related but distinct source of endogeneity is officer bias given that an interaction has already ocurred (the previous was focused on bias that increases $P(D|R)$). If some police officers are prejudiced against black people and believe (consciously or inconsciously) that they are more dangerous than other ethnic groups, then this bias could be correlated with an increased propensity to exaggerate the variables V experienced by the officer during the interaction, as well as an increased propensity to use deadly force given a perceived level of V. This source of endogeneity decreases our confidencein being able to test (2), because V is reported by the officer, so it may not be accurate and may be correlated with their propensity to use force.

2. Spell out conditions on (R, S, U, V, D, Z) (perhaps using a causal diagram) which would suffice to interpret (1) and (2) as evidence that there are no racial differences in the victims of police shootings. In particular, what does one need to assume about $Pr(D = 1|R, U)$?

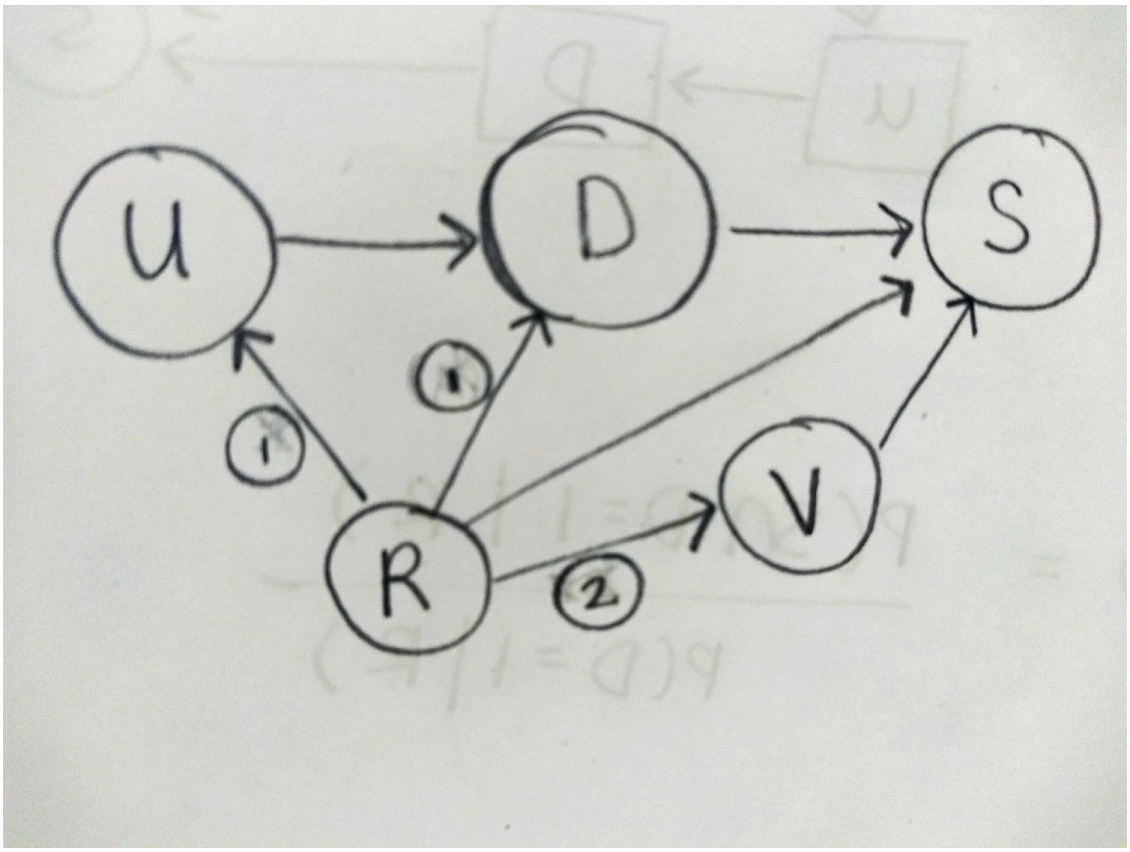
   Answer The below causal diagram draws the relationships between (R, S, U, D, V). The omitted variable of racial profiling described above is what causes the relationship between race and factors previous to the encounter or directly the probability of being stopped marked by (1). The omitted variable of bias that ocurs during the encounter is marked by the path from R to V, marked with (2). All of these signalled arrows must "not exist" (i.e., there is no relationship between race and getting stopped or the officer's perception of the danger level of the encounter) in order for equations (1) and (2) to be interpreted as evidence of no racial differences in the propensity to use force, because those equations attempt to only measure the direct line from R to S.

   Note that equation (1) above attempts to do inference on $Pr(S|D = 1, R, U)$. However,

$$Pr(S|D = 1, R, U) = \frac{Pr(S \cap D = 1|R, U)}{P(D = 1|R, U)}$$

   but $P(D = 1|R, U)$ is not observable just from police incidence reports. Therefore, we need the

condition $Pr(D = 1|R, U) = Pr(D = 1|U)$ (the probability of being stopped is independent of race) so that prejudice does not confound estimation by interfering in the denominator.



3. Consider an alternative ("driving while Black") model in which the police use race as a criterion for stopping or otherwise interacting with a given civilian. Compare the causal structure of this model with your answer to (1). Viewed through the lens of this model, how would one interpret Fry's failure to reject $Pr(S|D = 1, R) = Pr(S|D = 1)$?

   Answer The "driving while Black" model explicitly posits what I call the "racial profiling" confounding factor above. If police use race as a criteran for stopping a civilian, then R directly influences D in the causal diagram, so our "exclusion restriction" is violated. Fry fails to reject:

$$Pr(S|D = 1, R) = Pr(S|D = 1)$$
$$\frac{Pr(S \cap D = 1|R)}{Pr(D = 1|R)} = \frac{Pr(S \cap D = 1)}{Pr(D = 1)}$$

   If we suspect that the driving while Black model has relevance here, then its prediction that $Pr(D = 1|R) > Pr(D = 1)$ implies that it is likely that Fry's estimated probabilities are equal because $Pr(S \cap D = 1|R) > Pr(S \cap D = 1)$.

4. The Justice Department should care about which (Fry's or the "driving while Black") is the better model. How might one go about testing one against the other?

Answer

The main conflict between these models is that "driving while Black" proposes that $P(D = 1|R) > P(D = 1)$, whereas Fry's model requires $P(D = 1|R) = P(D = 1)$ for his identification strategy. Therefore, testing these models against eachother only requires evidence of $P(D = 1)$ vs. $P(D = 1|R)$. For example, Horrace and Rohlin (2016) exploit the relative darkness of nightand hence deteriorated ability of police officers to identify the race of someone they may stop) to measure racial profiling, and find that the "odds of a black driver being stopped (relative to nonblack drivers) increase 15% in daylight compared to darkness" (Horrace and Rohlin, 2016).

References:

William C. Horrace, Shawn M. Rohlin; How Dark Is Dark? Bright Lights, Big City, Racial Profiling. The Review of Economics and Statistics 2016; 98 (2): 226–232. doi: https://doi.org/10.1162/REST_a_00543

5 Weighting of Linear IV Estimators

Consider the Linear IV model

$$y = X\beta + u \qquad \mathbb{E}(Z'u) = 0$$

1. Exploiting the moment condition, under what conditions does the estimator b_{IV} satisfying $Z'y = (Z'X)b_{IV}$ consistently estimate β?

   Answer Hansen (2021) outlines the conditions under which IV is consistent. The moment conditions already provide the assumption that $\mathbb{E}(Z'u) = 0$. Additionally supposing the data is iid, and that each of y, X, Z have finite variance, then we only need:

   1. $\mathbb{E}(Z'Z)$ positive definite, and
   2. $\mathbb{E}(Z'X)$ full column rank (instrument relevance)

   for consistency.

2. Suppose that Z has l columns. Construct a symmetric, $l \times l$ full rank matrix W, and a corresponding estimator b_W satisfying $WZ'y = W(Z'X)b_W$. Under what conditions will this estimator consistently estimate β?

   Answer This estimator will consistently estimate β for any full rank matrix W such that $\mathbb{E}(WZ'X)$ is full column rank, plus the regularity conditions about the data being iid and y, X, Z having finite variance.

3. Describe the GMM criterion function that b_W minimizes.

   Answer Let:

   $$g_N(\beta) = \sum_{i=1}^{N} W_i Z_i' y - W_i Z_i' X_i \beta = WZ'y - WZ'X\beta$$

   then the GMM criterian function is:

$$J(\beta) = N(WZ'y - WZ'X\beta)'(WZ'y - WZ'X\beta)$$

4. Consider Hansen's description of the two-stage least squares estimator (Section 12.12). What is W for this estimator? Under what conditions is this the optimally weighted GMM estimator?

   Answer Minimizing the above, we can see that:

   $$J_\beta = -2X'ZW(Z'y - Z'Xb_{gmm}) = 0 \Rightarrow b_{gmm} = (X'ZWZ'X)^{-1}(X'ZWZ'y)$$

   In two-stage least squares, substituting $\hat{X} = Z(Z'Z)^{-1}Z'X$, we have:

   $$\begin{aligned} b_{2sls} &= (X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \end{aligned}$$

   Comparing these two expressions, we can see that W for this estimator is $(Z'Z)^{-1}$. Two stage least squares therefore is the optimally weighted GMM estimator under conditional homoskedasticity, because the scaling does not matter for the weighting matrix, and under conditional homoskedasticity the efficient weighting matrix is $\hat{\Omega}^{-1} = \mathbb{E}(Z'Z)^{-1}$. When distrubances are not homoskedastic, $W = \hat{\Omega}^{-1}$ would take a different form, and these would no longer be equal.

5. $W = I$ for the b_{IV} estimator described above. Under what conditions is b_{IV} the optimally weighted GMM estimator?

   Answer When IV is just identified, the IV and GMM estimators are identical. To see this, note that in the just-identified case, $X'Z$ is a square matrix. Let Ω^{-1} be the efficient weight matrix for GMM. Then

   $$\begin{aligned} b_{gmm} &= (X'Z\Omega^{-1}Z'X)^{-1}(X'Z\Omega^{-1}Z'y) = (Z'X)^{-1}\Omega(X'Z)^{-1}(X'Z\Omega^{-1}Z'y) \\ &= (Z'X)^{-1}Z'y = b_{iv} \end{aligned}$$

   Because these estimators are exactly the same and GMM is efficiently weighted, b_{IV} is the efficiently weighted GMM estimator in the just-identified case. This is underscored when examining the variances directly. The variance of the efficiently weighted GMM estimator in the square case reduces to:

   $$V_{gmm} = (Z'X)^{-1}\Omega(X'Z)^{-1} = (Z'X)^{-1}Z'ee'Z(X'Z)^{-1} = V_{IV}$$

6. For the estimator described in (2), suppose that W is diagonal, with diagonal elements proportional to $(1, 1/2, 1/4, \ldots, 2^{1-l})$. Under what conditions is the estimator b_W the optimally weighted estimator? Can you think of a practical example where the optimal weighting matrix might have this structure?

   Answer b_W is the optimally weighted estimator if b_W is proportional to $(Z'Z)^{-1}$. If this is the case, then this essentially states that the l instruments have no covariance between themselves, and the variance of each instrument is half of the one preceding it.

   As a practical example, this could be the case if the Z_l are independent measurements of a variable X, each with different units of measurement that result in different variances. FIXME