Empirical Finance: Study Questions Week 2
Imperial College London

Hormoz Ramian

Question 1: Consider estimating the elasticity of a consumption commodity demand denoted by $q_i^d = \alpha_0 + \alpha_1 p_i + u_i$ where $p_i = \ln P_i$, P_i is the actual price, and $q_i^d = \ln Q_i^d$ with Q_i^d denoting the actual quantity demanded. The term u_i represents other factors besides price that affect demand, such as income and consumer taste. The supply equation is in the same form, and is given by $q_i^s = \beta_0 + \beta_1 p_i + v_i$ where the term v_i represents the factors that affect supply, such as weather conditions production factor prices, and union status. Assuming first, that the two error terms are uncorrelated, and second that the equilibrium condition $q_i^d = q_i^s = q_i$ holds, then:

(1.1) Derive the the reduce form system in terms of p_i and q_i and provide the terms inside the 2×2 symmetric variance-covariance matrix $\text{cov}(p_i, q_i)$.

(1.2) Show that the ordinary least squares estimator resulting from the regression of q_i on p_i is biased for both structural parameters α_1 and β_1.

(1.3) Now suppose that variable z_i (e.g. weather condition) is available which is uncorrelated with u_i such that $\text{cov}(z_i, u_i) = 0 \ \forall_i$ but $\text{cov}(z_i, p_i) \neq 0$. Show that using z_i, an instrumental variable approach is able to estimate α_1.

Question 2 Suppose $\mathbb{E}[u_i|x_i] \neq 0$ and that $y_i = \beta_0 + \beta_1 x_i + u_i$. Consider an instrument z which is correlated with the explanatory variable but not the error term together with the following tabulated values where a, b and c are some unknowns binary values.

y	x	z
1	1	0
1	b	0
0	1	1
a	1	0
0	1	c

(2.1) Tabulate the joint distribution of X and Y from the the table and compute $\mathbb{E}[X|y = 0]$.

(2.2) Suppose $x_i = \delta_0 + \delta_1 z_i + v_i$ then compute the instrumental variable estimator $\widehat{\beta}_{1,IV}$ obtained from regressing y on a constant and $\widehat{x}_i$, when $a = 0$, $b = 0$ and $c = 0$.

(2.3) Repeat (2.2) first when $a = 0$, $b = 0$ and $c = 1$, then with $a = 1$, $b = 0$ and $c = 0$.

(2.4) Repeat (2.2) first when $a = 1$, $b = 1$ and $c = 1$.

Question 3 Consider the following regression model, $i = 1, \ldots, n$:

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i \tag{1}$$

(3.1) Derive the OLS estimates when the outcome variable is a constant, $y_i = 1$, $\forall_i$.

(3.2) Suppose $x_{j,i}$ is a risky return on an investment ($j = 1, 2$) and an investor wishes to split \$1.00 between the two assets, assuming a preference proportional to higher expected return and inversely proportional to variance: use the results in previous part and propose an econometric specification to decide how much the investor allocates to each asset[1].

[1] The results in part (3.1) are essentially the answer to this part, except that you may need to apply a normalization

(3.3) Now suppose x_j's are linearly independent with $\overline{x}_1 = \overline{x}_2$, $\sum_i x_{j,i}^2 = n\delta^{j-1}\overline{x}_1$. What are the OLS estimates when $\delta \in (0,1)$. Show the results numerically as a function of δ.

Question 4 Suppose $\mathbb{E}[u_i|S\&P_i, leverage_i] = 0$, $\forall_i$ and $i = 1, \dots, N$ where N is the sample size, then the share price of an asset can be expressed in terms of leverage (on the liabilities side of the balance sheet of the company that issues equity) and whether the asset is included in the S&P500 universe according to the following linear regression:

$$\begin{aligned} shareprice_i &= \alpha + \beta S\&P_i + \theta leverage_i + u_i \\ S\&P_i &= \begin{cases} 1 & \text{if Listed} \\ 0 & \text{if Non-listed} \end{cases} \end{aligned}$$

(4.1) Quantify the $S\&P$-listed price effect. Interpret the result.

(4.2) Quantify the $S\&P$-listed leverage effect. Interpret the result.

(4.3) Now consider running a similar regression to examine bond prices with various ratings in the following way and explain why construction of $Rating_i$ with multiple values can be a useful explanatory variable.

$$\begin{aligned} bondprice_i &= \alpha + \beta Rating_i + \theta leverage_i + u_i \\ Rating_i &= \begin{cases} 6 & \text{if} \quad \text{AAA} \\ 5 & \text{if} \quad \text{AA} \\ 4 & \text{if} \quad \text{A} \\ 3 & \text{if} \quad \text{BBB} \\ 2 & \text{if} \quad \text{BB} \\ 1 & \text{if} \quad \text{B} \\ 0 & \text{if} \quad \text{lower} \end{cases} \end{aligned}$$

(4.4) How would the results change if we re-write the regression in the following way, particularly, using each rating separately as an additional regressor (e.g AAA$_i$ is equal to one when a bond has a triple-A rating and zero otherwise):

$$bondprice_i = \alpha + \beta_{\text{AAA}}\text{AAA}_i + \beta_{\text{AA}}\text{AA}_i + \ldots + \beta_{\text{B}}\text{B}_i + \beta_2 leverage_i + u_i \tag{2}$$

(4.5) Suppose we collapse bond ratings into broader bands. Discuss how this specification could provide a better way to examine the data.

$$Rating_i = \begin{cases} 2 & \text{if} \quad \text{A-band} \\ 1 & \text{if} \quad \text{B-band} \\ 0 & \text{if} \quad \text{lower} \end{cases} \tag{3}$$

$$bondprice_i = \alpha + \beta_{\text{A}}\text{A-band}_i + \beta_{\text{B}}\text{B-band}_i + \theta leverage_i + u_i \tag{4}$$

---

such that OLS estimates in previous part sum to 1. For further reference (optional reading) see "The Sampling Error in Estimates of Mean-Variance Efficient Portfolio Weights", Britten-Jones (1999), The Journal of Finance, https://onlinelibrary.wiley.com/doi/full/10.1111/0022-1082.00120

Answers

Question 1

(1.1) A regressor is endogenous if it is not predetermined (i.e., not orthogonal to the error term), that is, if it does not satisfy the orthogonality condition. When the equation includes the intercept, the orthogonality condition is violated and hence the regressor is endogenous, if and only if the regressor is correlated with the error term. In the present example, the regressor p_i is necessarily endogenous in both equations. To see why, treat the system of two simultaneous equations and solve for (p_i, q_i) as,

$$p_i = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1} \tag{5}$$

$$q_i = \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_i - \beta_1 u_i}{\alpha_1 - \beta_1} \tag{6}$$

such that the price is a function of the two error terms. We can calculate the covariance of the regressor p_i with the demand shifter u_i and the supply shifter v_i:

$$\mathrm{cov}(p_i, u_i) = \mathrm{cov}\left[\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1}, u_i\right] = -\frac{1}{\alpha_1 - \beta_1}\mathrm{var}(u_i) \tag{7}$$

$$\mathrm{cov}(p_i, v_i) = \mathrm{cov}\left[\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1}, v_i\right] = \frac{1}{\alpha_1 - \beta_1}\mathrm{var}(v_i) \tag{8}$$

which are not zero (unless $\mathrm{var}(u_i) = 0$ and $\mathrm{var}(v_i) = 0$). Therefore, price is correlated positively with the demand shifter and negatively with the supply shifter, if the demand curve is downward-sloping ($\alpha_1 < 0$) and the supply curve upwardsloping ($\beta_1 > 0$). In this example, endogeneity is a result of market equilibrium.

(1.2) When quantity is regressed on a constant and price, it estimates neither the demand curve or the supply curve, because price is endogenous in both the demand and supply equations. The OLS estimator is consistent for the least squares projection coefficients but in this setting projection of q_i on a constant and p_i, the coefficient of p_i is:

$$\frac{\mathrm{cov}(p_i, q_i)}{\mathrm{var}(p_i)} = \alpha_1 + \frac{\mathrm{cov}(p_i, u_i)}{\mathrm{var}(p_i)} = \beta_1 + \frac{\mathrm{cov}(p_i, v_i)}{\mathrm{var}(p_i)} = \frac{\alpha_1\,\mathrm{var}(v_i) + \beta_1\,\mathrm{var}(u_i)}{\mathrm{var}(v_i) + \mathrm{var}(u_i)} \in (\alpha_1, \beta_1)$$

therefore the OLS estimator for the slope measures a weighted average of α_1 and β_1. Such a bias is called the simultaneous equations bias. The OLS cannot consistently estimate the true values because both curves are shifted by other factors besides price, and we cannot tell from data whether the change in price and quantity is due to a demand shift or a supply shift. If, however, $u_i = 0$ (that is, the demand curve stays still), then the equilibrium prices and quantities will trace out the demand curve and the OLS is consistent to obtain α_1.

(1.3) From the results in previous parts, p_i has one part correlated with u_i which is $-\frac{u_i}{\alpha_1-\beta_1}$ and one part uncorrelated with u_i which is $\frac{v_i}{\alpha_1-\beta_1}$. If we can isolate the second part, then we can focus on those variations in p_i that are uncorrelated with u_i and disregard the variations in p_i that biases the OLS estimator. Take a supply shifter z_i (e.g., weather), which can be considered to be uncorrelated with the demand shifter u_i then:

$$\mathrm{cov}(z_i, q_i) = \alpha_1.\,\mathrm{cov}(z_i, p_i)$$

which yields $\widehat{\alpha} = \widehat{\mathrm{cov}}(z_i, q_i)/\widehat{\mathrm{cov}}(z_i, p_i)$ which is the IV estimator. Another method to estimate

α_1 as suggested above is to run regression:

$$q_i = \alpha_0 + \alpha_1 \widehat{p}_i + \widetilde{u}_i$$

where $\widehat{p}_i$ is the predicted value from the following regression:

$$p_i = \gamma_0 + \gamma_1 z_i + \eta_i$$

and that $\widetilde{u}_i = \alpha_1(p_i - \widehat{p}_i) + u_i$, and that $\operatorname{cov}(\widehat{p}_i, \widehat{u}_i) = 0$ thus the estimation is consistent (two-stage least squares).

Question 2

(2.1) The joint tabulation is:

		X	
		0	1
Y	0	0	0.6
	1	0.2	0.2

then $\mathbb{P}[X = 0|y = 0] = \frac{0}{0+0.6}$ and $\mathbb{P}[X = 1|y = 0] = \frac{0.6}{0+0.6}$

$$\mathbb{E}[X|y = 0] = 0 \times \frac{0}{0 + 0.6} + 1 \times \frac{0.6}{0 + 0.6}$$

(2.2) Given $(a, b, c) = (0, 0, 0)$ then $\mathbb{E}[X] = \frac{4}{5}$, $\mathbb{E}[Y] = \frac{2}{5}$, $\mathbb{E}[Z] = \frac{1}{5}$ and $\mathbb{E}[YZ] = 0$, $\mathbb{E}[XZ] = \frac{1}{5}$ then:

$$\begin{aligned} \operatorname{cov}(Y, Z) &= \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z] = -0.08 \\ \operatorname{cov}(X, Z) &= \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z] = 0.04 \end{aligned}$$

thus $\widehat{\beta}_{IV} = \operatorname{cov}(Y, Z)/\operatorname{cov}(X, Z) = -2$.

(2.3) First, given $(a, b, c) = (0, 0, 1)$ then $\mathbb{E}[X] = \frac{4}{5}$, $\mathbb{E}[Y] = \frac{3}{5}$, $\mathbb{E}[Z] = \frac{1}{5}$ and $\mathbb{E}[YZ] = 0$ and $\mathbb{E}[XZ] = \frac{1}{5}$ then $\widehat{\beta}_{IV} = \operatorname{cov}(Y, Z)/\operatorname{cov}(X, Z) = -2$. Second, given $(a, b, c) = (1, 0, 0)$ then $\mathbb{E}[X] = \frac{4}{5}$, $\mathbb{E}[Y] = \frac{3}{5}$, $\mathbb{E}[Z] = \frac{1}{5}$ and $\mathbb{E}[YZ] = 0$ and $\mathbb{E}[XZ] = \frac{1}{5}$ then:

$$\begin{aligned} \operatorname{cov}(Y, Z) &= \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z] = -0.12 \\ \operatorname{cov}(X, Z) &= \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z] = 0.04 \end{aligned}$$

thus $\widehat{\beta}_{IV} = \operatorname{cov}(Y, Z)/\operatorname{cov}(X, Z) = -3$

(2.4) Given $(a, b, c) = (1, 1, 1)$ then $\mathbb{E}[X] = 1$, $\mathbb{E}[Y] = \frac{3}{5}$, $\mathbb{E}[Z] = \frac{2}{5}$, $\mathbb{E}[YZ] = 0$ and $\mathbb{E}[XZ] = \frac{2}{5}$,

$$\begin{aligned} \operatorname{cov}(Y, Z) &= \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z] = -0.6 \\ \operatorname{cov}(X, Z) &= \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z] = 0 \end{aligned}$$

thus $\widehat{\beta}_{IV} = \operatorname{cov}(Y, Z)/\operatorname{cov}(X, Z)$ does not exist. This is because X is no longer a random variable and cannot correlate with the instrument.

Question 3

(3.1) Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2]$ and $\boldsymbol{y} = \mathbf{1}_{n\times 1}$

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{1}_{n\times 1} \\ &= \begin{bmatrix} \sum_i x_{1,i}^2 & 0 \\ 0 & \sum_i x_{2,i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i x_{1,i} \\ \sum_i x_{2,i} \end{bmatrix} \end{aligned}$$

(3.2) , (3.3)

$$\left(\widehat{\beta}_1, \widehat{\beta}_2\right) = \left(\frac{\sum_i x_{1,i}}{\sum_i x_{1,i}^2}, \frac{\sum_i x_{2,i}}{\sum_i x_{2,i}^2}\right) \tag{9}$$

re-write the sum of both regression parameters as $\sum_k \widehat{\beta}_i = \mathbf{1}'_{k\times 1}\widehat{\boldsymbol{\beta}}$ and using the information $\overline{x}_1 = \overline{x}_2$, $\sum_i x_{j,i}^2 = n\delta^{j-1}\overline{x}_1$ for $\delta \in (0,1)$ then:

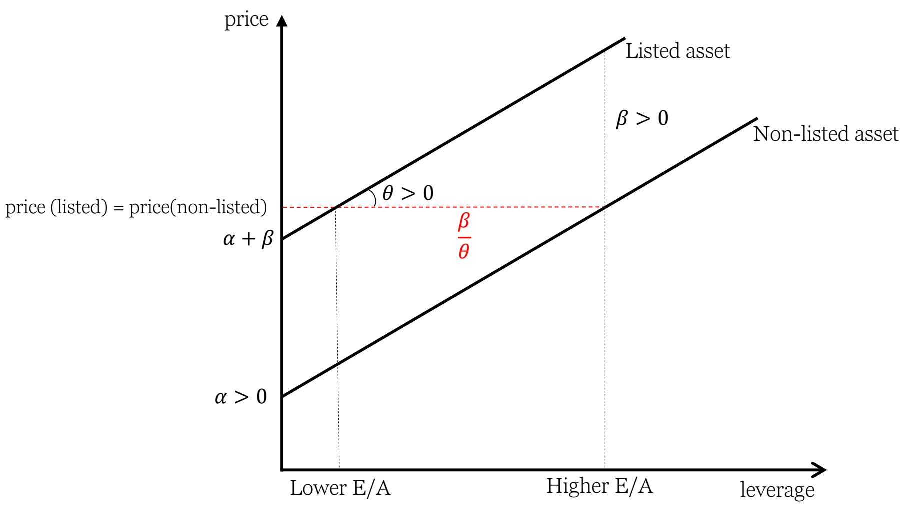
$$\boldsymbol{\omega} := \frac{\widehat{\boldsymbol{\beta}}}{\mathbf{1}'_{k\times 1}\widehat{\boldsymbol{\beta}}} = \left(\frac{\frac{\sum_i x_{1,i}}{\sum_i x_{1,i}^2}}{\frac{\sum_i x_{1,i}}{\sum_i x_{1,i}^2} + \frac{\sum_i x_{2,i}}{\sum_i x_{2,i}^2}}, \frac{\frac{\sum_i x_{2,i}}{\sum_i x_{2,i}^2}}{\frac{\sum_i x_{1,i}}{\sum_i x_{1,i}^2} + \frac{\sum_i x_{2,i}}{\sum_i x_{2,i}^2}}\right) = \left(\frac{1}{1+\frac{1}{\delta}}, \frac{\frac{1}{\delta}}{1+\frac{1}{\delta}}\right) \tag{10}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2)'$ is a 2×1 (estimated) vector of allocation weights on assets 1 and 2 such that $\omega_2 = 1 - \omega_1$ but ω_i is not necessarily between zero and one, however, when $\delta = 1$ then $\omega_1 = \omega_2 = \frac{1}{2}$ and $\lim_{\delta\to 0}\omega_1 = 0$ and $\lim_{\delta\to 0}\omega_2 = 1$, also both allocation weights are monotonic functions of δ thus they remain bounded within $(\frac{1}{2}, 1)$ interval for $\delta \in (0,1)$. The parameter δ together with $\overline{x}_1 = \overline{x}_2$, $\sum_i x_{j,i}^2 = n\delta^{j-1}\overline{x}_1$ specifies the relative importance of each asset based on increased second moment or approximately variance (assuming small mean returns). The second asset becomes increasingly less attractive to a risk-averse decision maker as δ increases from zero to one.

Question 4

(4.1) Given this specification, $\widehat{\beta}$ is the average difference between share prices of a listed asset versus a non-listed asset (with identical other characteristic).

(4.2) Define *leverage* = *equity*/*totalassets* and suppose $\theta > 0$. This implies that more equity on a company's balance sheet has an increasing effect on its share price. Consider two assets A and B which are priced exactly the same and have identical other characteristics except that asset A is listed but asset B is non-listed, then asset B needs to be supported by more equity to compensate for being non-listed. More formally, higher equity requirement or leverage cost can be expressed with β/θ.



(4.3) This is a sensible way to construct an explanatory variable with multiple discrete values because the values are inherently sorted (an ordered set). This indicates that higher numerical values are expected to have a monotonic effect (increasing or decreasing) on the left hand side variable. Note that this is always the case when a variable can take only two values e.g. zero and one, as this can be re-expressed as an ordered set. However, we should be cautious when constructing such variables when there are more than two possibilities. For instance, when constructing a variable to identify if the headquarter of a company belongs to a certain country or continent, then it is not plausible to assign values $1, 2, \ldots, 5$ to Europe, Americas, Asia, Africa and Oceania, because setting these values do not represent an inherent ordering. When interpreting the estimated coefficient for β, we note that when $Rating_i = 0$ then $Rating_i$ is absent from the specification. That is to say the effect of having a rating strictly lower than B is captured by $\widehat{\alpha}$ or the base group. The estimated value $\widehat{\beta}$ then quantifies how much the left hand side variable e.g. increases (decreases), on average and given a confidence level, when rating of an asset increases (decreases), from strictly lower than B to B. Note that the same logic applies when an asset rating increases from B to BB and so on. An important implication of estimating $\widehat{\beta}$ from variable *Rating* is to note that this effect is effectively the same for all increments, for instance, *bondprice* changes the same amount when *rating* changes from B to BB or changes from AA to AAA. In fact $\widehat{\beta}$ shows the averages of all such incremental effects.

(4.4) This specification considers a zero or one dummy variable per each rating which results in estimation of six separate parameters each of which quantifying the individual effect of each rating on prices. Re-expressing the specification in previous part into the one in this part has the benefit of providing more information on whether a rating improvement from e.g. B to BB is statistically different from e.g. AA to AAA.

(4.5) This specification allows us to identify only between broad bands which is based on the assumption or knowledge that within each band, ratings have approximately the same effect on the prices[2]. The advantage of using such specification is to propose a parsimonious way to examine the data as fewer number of parameters should now be estimated.

[2]Think about tax brackets as an example where individuals within the same tax bracket would get the estimated parameters. Such model assumes that two individuals with different incomes, one who earns one cent above the bracket floor, and one who earns one cent less than the same bracket ceiling, approximately share the same typical lifestyle, hence grouping them into the same bracket is sensible without missing out important information.