14.661 Review Exercise - Answers

Konrad Menzel (menzel@mit.edu)

September 13, 2006

A Income and Substitution Effects

The graphs illustrating how the response of demand to a price change can be decomposed into income and substitution effects can be found in any standard intermediate microeconomics textbook.

B Consumer Choice

This exercise is supposed to illustrate the mechanics of the two dual optimization problems - the utility maximization problem (UMP) and the expenditure minimization problem (EMP) - in order to review some of the most important concepts from classical consumer choice theory. Later in the lecture, this will be the basic framework for the discussion of individual labor supply decisions (where one good will be interpreted as "leisure"), and standard results like Roy's identity or Shepard's lemma will often turn out to be useful when deriving household responses to changes in wages or income taxes.

(1) The marginal rate of substitution between the two goods x_1 and x_2 is

$$MRS_{2,1} := \left.\frac{dx_2}{dx_1}\right|_{du=0} = -\frac{\frac{\partial}{\partial x_1}u(x_1,x_2)}{\frac{\partial}{\partial x_2}u(x_1,x_2)} = \frac{\alpha}{1-\alpha}\frac{x_2-\gamma_2}{x_1-\gamma_1}$$

(2) The utility maximization problem (UMP) is

$$\max_{x_1,x_2}\left\{\alpha\log(x_1-\gamma_1)+(1-\alpha)\log(x_2-\gamma_2)\right\} \qquad \text{s.t. } p_1x_1+p_2x_2\le y$$

While it is possible to solve the utility maximization problem by plugging the budget constraint into $u(x_1,x_2)$ in order to reduce the UMP to a one-dimensional unconstrained problem (which we can since the budget constraint will clearly be binding), the mechanics of the problem will be more transparent if we solve it through its Lagrangian

$$\mathcal{L}(x_1,x_2,\lambda)=\alpha\log(x_1-\gamma_1)+(1-\alpha)\log(x_2-\gamma_2)+\lambda\Big(y-p_1x_1-p_2x_2\Big)$$

The corresponding first-order conditions (FOC) are

$$\begin{aligned}
\textbf{(I)}\ \frac{\partial\mathcal{L}}{\partial x_1}=0 &\iff \frac{\alpha}{x_1-\gamma_1}=\lambda p_1\\
\textbf{(II)}\ \frac{\partial\mathcal{L}}{\partial x_2}=0 &\iff \frac{1-\alpha}{x_2-\gamma_2}=\lambda p_2\\
\textbf{(III)}\ \frac{\partial\mathcal{L}}{\partial \lambda}=0 &\iff p_1x_1+p_2x_2=y
\end{aligned}$$

From **(I)** and **(II)**, we get

$$\textbf{(I')}\ \frac{p_1}{p_2} = \frac{\alpha}{1-\alpha}\frac{x_2-\gamma_2}{x_1-\gamma_1} = -MRS_{2,1}$$

Graphically, **(III)** states that the optimal solution lies on the budget line, whereas **(I')** implies that the budget line and the consumer's indifference curve actually are tangents at the optimal bundle.

(3) From the budget constraint **(III)**, we know that the Marshallian demand for good 2, $x_2^M = \frac{y-p_1x_1^M}{p_2}$. Plugging this into the tangency condition **(I')**, we obtain

$$p_1 = \frac{\alpha}{1-\alpha}\frac{y-p_1x_1^M-p_2\gamma_2}{x_1^M-\gamma_1} \iff p_1x_1^M\left(1+\frac{\alpha}{1-\alpha}\right) = p_1\gamma_1 + \frac{\alpha}{1-\alpha}(y-p_2\gamma_2)$$

and, solving for x_1^M

$$x_1^M = (1-\alpha)\gamma_1 + \alpha\frac{y-p_2\gamma_2}{p_1} = \gamma_1 + \alpha\frac{y-p_1\gamma_1-p_2\gamma_2}{p_1}$$

By symmetry,

$$x_2^M = \gamma_2 + (1-\alpha)\frac{y-p_1\gamma_1-p_2\gamma_2}{p_2}$$

This gives us the indirect utility function

$$v(p_1,p_2,y) = u\left(\,x_1^M(p_1,p_2,y), x_2^M(p_1,p_2,y)\right) = \log(y-p_1\gamma_2-p_2\gamma_2) - \alpha\log(p_1) - (1-\alpha)\log(p_2) + \text{const.}$$

Therefore,

$$\begin{aligned}\frac{\partial}{\partial p_1}v(p_1,p_2,y) &= -\frac{\gamma_1}{y-p_1\gamma_2-p_2\gamma_2} - \frac{\alpha}{p_1}\\ \frac{\partial}{\partial y}v(p_1,p_2,y) &= \frac{1}{y-p_1\gamma_2-p_2\gamma_2}\end{aligned}$$

Plugging this into Roy's identity,

$$-\frac{\frac{\partial}{\partial p_1}v(p_1,p_2,y)}{\frac{\partial}{\partial y}v(p_1,p_2,y)} = \gamma_1 + \alpha\frac{y-p_1\gamma_2-p_2\gamma_2}{p_1} = x_1^M(p_1,p_2,y)$$

In a similar fashion, we can check that x_2^M is consistent with Roy's identity.

(4)&(5) The cost function is the solution to the expenditure minimization problem (EMP)

$$\min_{x_1,x_2}\left\{\,p_1x_1+p_2x_2\right\} \qquad \text{s.t. } \alpha\log(x_1-\gamma_1)+(1-\alpha)\log(x_2-\gamma_2) \geq \overline{u}$$

and the corresponding Lagrangian is

$$\mathcal{L}(x_1,x_2,\mu) = p_1x_1+p_2x_2+\mu\left(\,\alpha\log(x_1-\gamma_1)+(1-\alpha)\log(x_2-\gamma_2)-\overline{u}\right)$$

with first-order conditions

$$\begin{aligned}\textbf{(A)}\ \frac{\partial\mathcal{L}}{\partial x_1} = 0 \quad &\iff \quad p_1 = \mu\frac{\alpha}{x_1^H-\gamma_1}\\ \textbf{(B)}\ \frac{\partial\mathcal{L}}{\partial x_2} = 0 \quad &\iff \quad p_2 = \mu\frac{1-\alpha}{x_2^H-\gamma_2}\\ \textbf{(C)}\ \frac{\partial\mathcal{L}}{\partial \mu} = 0 \quad &\iff \quad \overline{u} = \alpha\log(x_1^H-\gamma_1)+(1-\alpha)\log(x_2^H-\gamma_2)\end{aligned}$$

As in the previous part of this problem, the first two FOCs combine to

$$\textbf{(A')}\ \frac{p_1}{p_2} = \frac{\alpha}{1-\alpha}\frac{x_2^H - \gamma_2}{x_1^H - \gamma_1} = -MRS_{2,1} \Longrightarrow x_2^H - \gamma_2 = \frac{1-\alpha}{\alpha}\frac{p_1}{p_2}(x_1^H - \gamma_1)$$

This should not come as a surprise, since each solution to the EMP is also the solution to a UMP. The main difference between the two programs is in the comparative statics with respect to price changes: the UMP holds the budget fixed but allows for changes in overall (indirect) utility, whereas the EMP traces out compensated demand moving along the indifference curve and adjusts the budget (the expenditure function) accordingly.
Substituting **(A')** into the constraint and collecting terms, we get

$$\textbf{(C')}\ \overline{u} = \log(x_1^H - \gamma_1) + (1-\alpha)\log\left(\frac{1-\alpha}{\alpha}\frac{p_1}{p_2}\right)$$

Solving for Hicksian/compensated demand, x_1^H,

$$\log(x_1^H - \gamma_1) = \log\left(\exp(\overline{u})\left[\frac{\alpha}{1-\alpha}\frac{p_2}{p_1}\right]^{1-\alpha}\right) \Longrightarrow x_1^H = \gamma_1 + \exp(\overline{u})\left[\frac{\alpha}{1-\alpha}\frac{p_2}{p_1}\right]^{1-\alpha}$$

In the same fashion,

$$x_2^H = \gamma_2 + \exp(\overline{u})\left[\frac{1-\alpha}{\alpha}\frac{p_1}{p_2}\right]^{\alpha}$$

Therefore, the expenditure function is

$$E(p_1,p_2,\overline{u}) = p_1x_1^H + p_2x_2^H = p_1\left\{\gamma_1 + \exp(\overline{u})\left[\frac{\alpha}{1-\alpha}\frac{p_2}{p_1}\right]^{1-\alpha}\right\} + p_2\left\{\gamma_2 + \exp(\overline{u})\left[\frac{1-\alpha}{\alpha}\frac{p_1}{p_2}\right]^{\alpha}\right\}$$

Differentiating with respect to p_1 using the product rule,

$$\begin{aligned}\frac{\partial}{\partial p_1}E(p_1,p_2,\overline{u}) &= x_1^H + p_1\frac{\partial}{\partial p_1}x_1^H + p_2\frac{\partial}{\partial p_1}x_2^H \\ &= x_1^H + (\alpha-1)p_1\exp(\overline{u})\left[\frac{\alpha p_2}{1-\alpha}\right]^{1-\alpha}p_1^{\alpha-2} + \alpha p_2\exp(\overline{u})\left[\frac{1-\alpha}{\alpha p_2}\right]^{\alpha}p_1^{\alpha-1} \\ &= x_1^H - \alpha^{1-\alpha}(1-\alpha)^{\alpha}\exp(\overline{u})\left(\frac{p_2}{p_1}\right)^{1-\alpha} + \alpha^{1-\alpha}(1-\alpha)^{\alpha}\exp(\overline{u})\left(\frac{p_2}{p_1}\right)^{1-\alpha} = x_1^H\end{aligned}$$

(6) Expenditure on x_1 as a function of income is simply

$$p_1x_1^M = p_1\left(\gamma_1 + \alpha\frac{y - p_1\gamma_2 - p_2\gamma_2}{p_1}\right)$$

Since $p_1\gamma_1 + p_2\gamma_2$ is the expenditure for the subsistence level in each good, we can interpret $y - p_1\gamma_1 - p_2\gamma_2$ as the portion of income that is still freely disposable after basic needs are satisfied. As with standard Cobb-Douglas preferences, this surplus income is spent in constant shares on each consumption good. But note that other than for Cobb-Douglas preferences, income and substitution effects on the total amount of consumption do not necessarily cancel out.

C Returns to Scale, Hypothesis Tests, Random vs. Fixed Effects

(1) A production function $f(L,K)$ is said to have constant returns to scale (CRS) if for any $\lambda \geq 0$, $f(\lambda L, \lambda K) = \lambda f(L,K)$. For the Cobb-Douglas production function, we have

$$Q(\lambda L, \lambda K) = \gamma(\lambda L)^{\alpha}(\lambda K)^{\beta} = \gamma \lambda^{\alpha+\beta} L^{\alpha} K^{\beta}$$

so that Q has CRS whenever $\alpha + \beta = 1$.

(2) Modifying the production function to

$$Q_t = \gamma L_t^{\alpha} K_t^{\beta} \exp(\varepsilon_t)$$

and taking logs, we get the desired specification. Note that even if $\mathbb{E}[\varepsilon_t | L_t, K_t] = 0$, in general $\mathbb{E}[Q_t] > \gamma L_t^{\alpha} K_t^{\beta}$ because $\mathbb{E}[\exp(\varepsilon_t)|\cdot] > \exp\left(\mathbb{E}[\varepsilon_t|\cdot]\right) = 1$.

Assuming $\mathbb{E}[\varepsilon_t | L_t, K_t] = 0$, we can estimate the parameters by regressing $\log(Q_t)$ on $\log(L_t)$, $\log(K_t)$, and a constant (which will give us the intercept $\log(\gamma)$).

(i) The two most important types of tests for OLS are the t-test and the F-test. For a t-test, note that under the assumption of constant returns to scale, we can rewrite the problem as

$$\log(Q_t) = \log(\gamma) + \alpha \log(L_t) + (1-\alpha)\log(K_t) + \varepsilon_t = \log(\gamma_t) + \alpha \log\left(\frac{L_t}{K_t}\right) + \log(K_t) + \varepsilon_t$$

Therefore for an t-test we would regress log output on log-capital, the log of workers per unit of capital, and a constant, and test whether the coefficient on capital, say π, is equal to one,

$$\mathcal{T} := \left.\frac{\hat{\pi} - 1}{\sigma_{\hat{\pi}}}\right|_{H_0:\pi=1} \overset{A}{\sim} t_{T-3}$$

where T is the number of periods over which the firm is observed.
Alternatively, could do an F-test (Wald/LR/LM). The easiest way of doing this would be to estimate the unrestricted model (as on the problem set) and a restricted model

$$[\log(Q_t) - \log(K_t)] = \log(\gamma) + \alpha\left[\log(L_t) - \log(K_t)\right] + \varepsilon_t$$

in order to obtain the respective sums of squared residuals, SSR_U (unrestricted model), and SSR_R (restricted model) in order to compute $\mathcal{F}$ for which under the null hypothesis of CRS,

$$\mathcal{F} := (T-3)\left.\frac{SSR_R - SSR_U}{SSR_U}\right|_{H_0:\alpha+\beta=1} \overset{A}{\sim} F_{(1,T-3)}$$

(ii) Differences in managerial efficiency across firms translate to variation in the parameter γ, so that the new estimating equation becomes

$$\log(Q_{it}) = \log(\gamma_i) + \alpha \log(L_{it}) + \beta \log(K_{it}) + \varepsilon_{it}$$

If we believe that input choices (L_{it}, K_{it}) have nothing to do with managerial efficiency (i.e. $(L_{it}, K_{it}) \perp \log(\gamma_i)$), we could in principle pool all data points and estimate the model via OLS as before. However, there are two reasons for not doing that. Since the new model now is

$$\log(Q_{it}) = \overline{\log(\gamma)} + \alpha \log(L_{it}) + \beta \log(K_{it}) + \underbrace{\left[\log(\gamma_i) - \overline{\log(\gamma)}\right] + \varepsilon_{it}}_{=:\,\eta_{it}}$$

(where $\overline{\log(\gamma)} := \frac{1}{N}\sum_{i=1}^{N}\log(\gamma_i)$), there are two reasons why the Gauss-Markov assumptions (on the new error term η_{it}) will in general not hold:

1. The conditional mean restriction $\mathbb{E}[\eta_{it}|K_{it}, L_{it}] = 0$ does not hold unless $\log(\gamma_i)$ is uncorrelated with (K_{it}, L_{it}).
2. Even if the conditional mean restriction holds, $\text{Cov}(\eta_{it}, \eta_{is}) = \text{Var}(\log(\gamma_i))$ which is greater than zero as long as there is some heterogeneity across firms.

A violation of the first kind will lead to biased and inconsistent estimates, whereas in the second case, OLS will only be inefficient.

So if we suspect that e.g. more efficient management uses more inputs, our estimates of returns to scale using pooled OLS or GLS will be biased (Q: which way?). A standard way to address this problem with this type of data is to run a fixed-effects regression, which consists in subtracting the within-firm means from the data - e.g. $\widehat{\log(Q_{it})}_{FE} := \log(Q_{it}) - \frac{1}{T}\sum_{s=1}^{T}\log(Q_{is})$ - and then run OLS over the transformed data. This procedure ensures that $\log(\gamma_i)$ drops out of the estimating equation and can therefore not affect our estimates.

If endogeneity (i.e. the violation of the zero conditional mean restriction) was not an issue, we could still use the additional information on firm heterogeneity from the panel structure of the data set in order to improve the precision of our estimator and estimate the model using GLS for random-effects panel data models.[1] The fixed-effects estimator is unbiased under these assumption, but not efficient. A reason for this is that while the fixed-effects transformation completely removes the "noise" emanating from the individual intercept $|log(\gamma_i)$, it also subtracts something off the data which contains some noise on its own, and the optimal random-effects GLS solves this trade-off optimally.

D Choice under Uncertainty, Insurance

Given a random payoff X, the certainty equivalent CE_X with respect to a strictly increasing (Bernoulli) utility function $u(\cdot)$ over money is defined by

$$u(CE_X) = \mathbb{E}[u(X)] \Longleftrightarrow CE_X := u^{-1}\left(\mathbb{E}[u(X)]\right)$$

The agent is said to be risk-averse if for any lottery X,

$$u(\mathbb{E}[X]) \geq \mathbb{E}[u(X)]$$

i.e. the agent would always choose to receive the expected value of the payoffs for sure over the lottery itself. Being risk-averse is equivalent with $u(\cdot)$ being concave.[2]

[1] The Random Effects GLS estimator for random-effects models runs a regression after subtracting from the data (in our case $\log(Q_{it})$, $\log(K_{it})$, and $\log(L_{it})$) a fraction of the respective within-firm means (e.g. $\widehat{\log(Q_{it})}_{RE} := \log(Q_{it}) - \lambda\frac{1}{T}\sum_{s=1}^{T}\log(Q_{is})$ etc.), and then runs pooled OLS on the transformed data. This doesn't fully cancel out the firm-specific intercept, but it lowers its contribution to the overall OLS residual $[\varepsilon_{it} - \bar{\varepsilon}_i] + (1-\lambda)\log(\gamma_i)$. The optimal value for λ actually turns out to be

$$\lambda = 1 - \frac{1}{\sqrt{1 + T\frac{\sigma_{c_i}^2}{\sigma_{\varepsilon}^2}}}$$

where in our example, $c_i = \log(\gamma_i)$ (see e.g. Wooldridge, "Econometric Analysis of Cross Sectional and Panel Data", pp.286-288). If there is much variation in managerial efficiency across firms, GLS puts more weights on the within-firm means to "filter out" the common component, if on the other hand idiosyncratic variation over time (i.e. in ε_{it} dominates or if the panel is relatively short, putting too much weight on the within-means will only tune up the noise, and therefore GLS will be closer to pooled OLS

[2] Jensen's inequality states that for any concave function $f(\cdot)$, $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$

If $u(\cdot)$ is increasing,

$$u(\mathbb{E}[X]) \geq \mathbb{E}[u(X)] = u(CE_X)$$

implies that $\mathbb{E}[X] \geq CE_X$. Therefore the agent would be willing to forego $\mathbb{E}[X] - CE_X \geq 0$ on average for not having to bear the risk. In uncertain environments, labor contracts usually include an insurance component.

E Investment Decisions

The most important type of investment decisions you'll see in Labor Economics concerns human capital. In order to be able to evaluate the return to an investment e.g. in education, it is often necessary to evaluate the present value of income streams over a longer period. If interest is paid at fixed intervals (say annually), an interest rate r translates to a discount factor $\delta := \frac{1}{1+r}$ so that the present value of a x per period over t periods is

$$V_0 = \sum_{s=1}^{t} \delta^s x = \frac{1-\delta}{1-\delta} x \sum_{s=1}^{t} \delta^s = \frac{x}{1-\delta} \sum_{s=1}^{t} (\delta^s - \delta^{s+1}) = \frac{x}{1-\delta} \left(\sum_{s=1}^{t} \delta^s - \sum_{s=2}^{t+1} \delta^s \right) = \frac{x(\delta - \delta^{t+1})}{1-\delta}$$

where in the last step, all summands that appear in both sums simply cancel out.
With interest compounded continuously at rate ϱ, the present value is calculated as

$$V_0 = \int_0^t x \exp(-\varrho s) ds = \left[-\frac{x \exp(-\varrho s)}{\varrho} \right]_{s=0}^{t} = \frac{x}{\varrho} \Big(1 - \exp(-\varrho t) \Big)$$

F Market Demand

One possible linear specification for the demand function is $Q = Ay - Bp$ where $A, B \geq 0$. The price elasticity of demand is

$$\eta := \frac{\partial Q}{\partial p} \frac{p}{Q} = -\frac{Bp}{Ay - Bp} = -\frac{Bp}{Q}$$

G Long versus Short Regression

Suppose we have estimated the two equations

$$\begin{aligned} W_i &= \hat{\alpha}_L + S_i \hat{\beta}_L + TS_i \hat{\gamma} + \hat{\varepsilon}_i \\ W_i &= \hat{\alpha}_S + S_i \hat{\beta}_S + \hat{\eta}_i \end{aligned}$$

where $(\hat{\alpha}_L, \hat{\beta}_L, \hat{\gamma})$ and $(\hat{\alpha}_S, \hat{\beta}_S)$ are the *estimated* regression coefficients and $(\hat{\varepsilon}_i, \hat{\eta}_i)$ the corresponding OLS residuals, implying that in the sample

$$\text{Cov}(S_i, \hat{\varepsilon}_i) = \text{Cov}(TS_i, \hat{\varepsilon}_i) = \text{Cov}(S_i, \hat{\eta}_i) = 0$$

by the usual properties of the OLS estimator. The coefficient on schooling from the "short" regression is

$$\hat{\beta}_S = \frac{\text{Cov}(W_i, S_i)}{\text{Var}(S_i)} = \frac{\text{Cov}(S_i \hat{\beta}_L + TS_i \hat{\gamma} + \hat{\varepsilon}_i, S_i)}{\text{Var}(S_i)} = \frac{\text{Var}(S_i) \hat{\beta}_L + \text{Cov}(TS_i, S_i) \hat{\gamma}}{\text{Var}(S_i)} = \hat{\beta}_L + \frac{\text{Cov}(TS_i, S_i) \hat{\gamma}}{\text{Var}(S_i)}$$

Therefore, the estimate from the short and long regressions differ only if the coefficient on test scores in the long regression is different from zero *and* if, in addition, test scores and schooling are correlated.
We often think of the long regression as the regression "we would like to run" and interpret the discrepancy between the two estimates as an "omitted variables bias" in the short regression. However, in this particular example, it is not so clear whether that's exactly the right way to think about the estimation problem (Q: under which circumstances would it be better to run the short regression if we want to give the coefficient on schooling a causal interpretation?).
More generally, suppose we have a long regression

$$Y_i = X_{1i}\beta_L + X_{2i}\gamma_L + \varepsilon_i$$

where X_{1i} and X_{2i} are row vectors. Then estimating the short regression

$$Y_i = X_{1i}\beta_S + \eta_i$$

would give us

$$\beta_S = (\mathbf{X_1'X_1})^{-\mathbf{1}}\mathbf{X_1'Y} = (\mathbf{X_1'X_1})^{-\mathbf{1}}\mathbf{X_1'}(\mathbf{X_1}\beta_\mathbf{L} + \mathbf{X_2}\gamma_\mathbf{L} + \varepsilon) = \beta_\mathbf{L} + (\mathbf{X_1'X_1})^{-\mathbf{1}}\mathbf{X_1'X_2}\gamma_\mathbf{L} + \mathbf{0}$$

H Partitioned Regression

Suppose your regression model is

$$Y_i = X_{1i}\beta + X_{2i}\gamma + \varepsilon_i$$

Instead of obtaining an estimate for β by regressing $\mathbf{Y}$ on the full set of regressors, $[\mathbf{X_1}, \mathbf{X_2}]$, it is also possible to start with two auxiliary regressions

$$\begin{aligned} Y_i &= X_{2i}\theta + \nu_i \\ X_{1i} &= X_{2i}\pi + \eta_i \end{aligned}$$

and construct residuals

$$\begin{aligned} \mathbf{Y}^\perp &:= \mathbf{Y} - \mathbf{X_2}\hat{\theta} = (\mathbf{I} - \mathbf{X_2}(\mathbf{X_2'X_2})^{-\mathbf{1}}\mathbf{X_2'})\mathbf{Y} =: (\mathbf{I} - \mathbf{P_{X_2}})\mathbf{Y} \\ \mathbf{X_1^\perp} &:= \mathbf{X_1} - \mathbf{X_2}\hat{\pi} = (\mathbf{I} - \mathbf{X_2}(\mathbf{X_2'X_2})^{-\mathbf{1}}\mathbf{X_2'})\mathbf{X_1} =: (\mathbf{I} - \mathbf{P_{X_2}})\mathbf{X_1} \end{aligned}$$

where $\mathbf{P_{X_2}} := X_2(X_2'X_2)^{-1}X_2'$ is a matrix that represents a projection on the space spanned by the column vectors in $\mathbf{X_2}$ with the properties $\mathbf{P_{X_2}P_{X_2}} = \mathbf{P_{X_2}}$ and $(\mathbf{I} - \mathbf{P_{X_2}})(\mathbf{I} - \mathbf{P_{X_2}}) = (\mathbf{I} - \mathbf{P_{X_2}})$ (idempotence, verify this using the definition of the projector). Also note that

$$(\mathbf{I} - \mathbf{P_{X_2}})\mathbf{X_2} = \mathbf{X_2} - \mathbf{X_2}(\mathbf{X_2'X_2})^{-\mathbf{1}}\mathbf{X_2'X_2} = \mathbf{X_2} - \mathbf{X_2} = \mathbf{0}_{(\mathbf{N},\mathbf{k_2})}$$

In a second stage we then regress

$$Y_i^\perp = X_{1i}^\perp\beta_2 + \zeta_i$$

Using the results on projectors, we can verify that

$$\begin{aligned} \hat{\beta}_2 &\equiv (\mathbf{X_1^{\perp}}'\mathbf{X_1^\perp})^{-\mathbf{1}}\mathbf{X_1^{\perp}}'\mathbf{Y}^\perp = (\mathbf{X_1'}(\mathbf{I} - \mathbf{P_{X_2}})\mathbf{X_1})^{-\mathbf{1}}\mathbf{X_1}(\mathbf{I} - \mathbf{P_{X_2}})(\mathbf{X_1}\beta + \mathbf{X_2}\gamma + \varepsilon) \\ &= \beta + (\mathbf{X_1'}(\mathbf{I} - \mathbf{P_{X_2}})\mathbf{X_1})^{-\mathbf{1}}\mathbf{X_1}(\mathbf{I} - \mathbf{P_{X_2}})\mathbf{X_2}\gamma + (\mathbf{X_1'}(\mathbf{I} - \mathbf{P_{X_2}})\mathbf{X_1})^{-\mathbf{1}}\mathbf{X_1}(\mathbf{I} - \mathbf{P_{X_2}})\varepsilon \equiv \hat{\beta}_\mathbf{LS} \end{aligned}$$

so that this procedure is numerically identical to OLS on the full set of regressors. This is also true for the estimated standard errors from the second stage.
This procedure is often also referred to as *partialing out* X_2, or partitioned regression. In principle this method is useful for reducing the computational cost of running a regression with many regressors (which is of little practical relevance today), but it also helps understand what multivariate regression does: it attributes to each regressor the joint variation with the dependent variable which cannot be "explained" by (i.e. which is orthogonal to) all other variables on the right-hand side of the regression equation.

I Probit/Discrete Choice

The likelihood of the sample is defined as

$$\begin{aligned}
\ell(\beta,\sigma;y_1,\ldots,y_N,X_1,\ldots,X_N) &:= \mathbb{P}\Big\{\ Y_1=y_1,Y_2=y_2,\ldots\Big| X_1,X_2,\ldots,\beta,\sigma\Big\}\\
&\overset{iid}{=} \prod_{i=1}^N \mathbb{P}\{X_i\beta>\varepsilon_i\}^{y_i}\mathbb{P}\{X_i\beta\le\varepsilon_i\}^{1-y_i}\\
&= \prod_{i=1}^N \Phi\left(\frac{X_i\beta}{\sigma}\right)^{y_i}\left[1-\Phi\left(\frac{X_i\beta}{\sigma}\right)\right]^{1-y_i}
\end{aligned}$$

where $\Phi(z)$ is the standard normal cdf. Taking logs, we get the log-likelihood function

$$\begin{aligned}
L(\beta,\sigma;y_1,\ldots,y_N,X_1,\ldots,X_N) &:= \log\Big(\ \ell(\beta,\sigma;y_1,\ldots,y_N,X_1,\ldots,X_N)\Big)\\
&= \sum_{i=1}^N\left\{y_i\log\left[\Phi\left(\frac{X_i\beta}{\sigma}\right)\right]+(1-y_i)\log\left[1-\Phi\left(\frac{X_i\beta}{\sigma}\right)\right]\right\}
\end{aligned}$$

The maximum-likelihood estimator (MLE) for the parameters is defined as

$$(\hat\beta,\hat\sigma)_{ML} := \arg\max_{\beta\in\mathbb{R}^k,\sigma\in\mathbb{R}_+} L(\beta,\sigma;y_1,\ldots,y_N,X_1,\ldots,X_N)$$

For those not familiar with nonlinear estimation techniques, MLEs have a number of desirable properties as consistency and asymptotic efficiency (if the specification of the model is correct).
At this point you may have noticed that the parameters we want to estimate always appear in ratios $\frac{1}{\sigma}\beta$ which means that if we replace the "true" value of σ, σ_0, with, say, $2\sigma_0$, scaling up β by the same factor would leave the value of the log-likelihood unchanged. Therefore the maximum can't be unique unless we normalize σ to some positive value, typically $\sigma=1$. We say that σ is *not identified* from the data. Such a normalization (often it's made only implicitly) is necessary for any type of discrete choice model (multiple choice, ordered choice, other distributions for the error term etc.).
If you think of $u_i=-X_i\beta+\varepsilon_i$ as the utility differential between working (and being paid) and not working, the fundamental reason for the identification problem is that, as you may recall from an intermediate micro class, a preference relation is only ordinal. I.e. any strictly monotone transformation of a given utility function represents exactly the same preferences, so that the scale of β has no empirical meaning.
The MLE solves the first-order conditions to the maximization problem (recall that we normalized σ to one)

$$\begin{aligned}
0=\frac{\partial}{\partial\beta}L(\beta,1;y_1,\ldots,X_1,\ldots) &= \sum_{i=1}^N\left\{\frac{y_i}{\Phi(X_i\beta)}-\frac{1-y_i}{1-\Phi(X_i\beta)}\right\}\varphi(X_i\beta)X_i\\
&= \sum_{i=1}^N\frac{y_i[1-\Phi(X_i\beta)]-(1-y_i)\Phi(X_i\beta)}{\Phi(X_i\beta)[1-\Phi(X_i\beta)]}\varphi(X_i\beta)X_i\\
&= \sum_{i=1}^N\frac{y_i-\Phi(X_i\beta)}{\Phi(X_i\beta)[1-\Phi(X_i\beta)]}\varphi(X_i\beta)X_i
\end{aligned}$$

Now, instead of estimating the parameters, consider an alternative approach: you could always try to approximate the conditional mean $\mathbb{E}[Y_i|X_i]$ by just fitting a standard normal cdf $\Phi(X_i\gamma)$ via weighted nonlinear least squares, i.e.

$$\hat{\gamma}_{NLS} := \arg\min_{\gamma \in \mathbb{R}^k} \sum_{i=1}^N w_i \left(y_i - \Phi(X_i\gamma) \right)^2$$

where w_i are weights you may want to use for estimation. The first-order conditions for this problem are

$$0 = \sum_{i=1}^N w_i(y_i - \Phi(X_i\gamma))\varphi(X_i\gamma)X_i$$

Now, plugging in

$$w_i := \frac{1}{\Phi(X_i\gamma)[1 - \Phi(X_i\gamma)]}$$

these are exactly the same first-order conditions as for the maximum-likelihood estimator, so that for these particular weights, $\hat{\gamma}_{NLS} = \hat{\beta}_{ML}$. Since we don't know the value of γ beforehand, we could start with a preliminary estimate $\gamma^{(0)}$ and update the weights using the estimate from the last stage as we go along.

The main insight from this is that - regardless of whether the specification of the error distribution is correct - maximum-likelihood estimators for discrete-choice models approximate the conditional mean of y_i with a weighted least-squares fit of the cdf chosen by the researcher.

J Experimental Design

Suppose you have N units, a share s of which is will be assigned to the treatment group, whereas the remaining $(1-s)N$ units serve as controls.

Now you could simply go ahead and calculate the variance of the difference in outcomes Y as a function of s,

$$\text{Var}(\bar{Y}_T - \bar{Y}_C) = \text{Var}(\bar{Y}_T) + \text{Var}(\bar{Y}_C) = \dots$$

However, it is slightly more instructive to set this up as a regression problem: define a treatment dummy

$$D_i = \begin{cases} 1 & \text{if unit } i \text{ is treated} \\ 0 & \text{otherwise} \end{cases}$$

and think of the problem as estimating the equation

$$Y_i = \alpha + D_i\beta + \varepsilon_i$$

The OLS coefficient will be an efficient estimate of the difference in means, and from standard results from regression analysis we know that since $\text{Var}(\varepsilon_i|D_i) := \sigma^2 = \text{const.}$

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{N\text{Var}(D_i)} = \frac{\sigma^2}{N(\mathbb{E}[D_i^2] - \mathbb{E}[D_i]^2)} = \frac{\sigma^2}{Ns(1-s)}$$

Therefore the variance of the estimator for the treatment effect is minimized for the proportion s which maximizes the variance of the treatment indicator in the regression, $s(1-s)$, so that $s^* = \frac{1}{2}$.