ECO375 Exam Review

Matt Tudball

University of Toronto Mississauga

November 30, 2017

ECO375 Exam Review

Welcome back!

Today's coverage:

- Review of statistical terminology
- Least squares estimation
- MLR assumptions and properties of the least squares estimator
- Inference
- Heteroscedasticity
- Instrumental variables
- In-class exercise

Do not exclusively study from these slides. I do not cover everything that may appear on the exam.

The Statistical Problem

- **Data Generating Process:** The data generating process is the true underlying mechanism that is generating the data we observe in our sample. We often assume that it takes a particular functional form. For example,
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$
  Given $x_{i1}, x_{i2}, ..., x_{ik}$ this is the (unobservable) process which is generating the data y_i.
- **Parameters:** The above example is a **parametric model**, meaning that it can be characterised by a set of parameter values, $\{\beta_0, \beta_1, \beta_2, ..., \beta_k\}$.
- The **statistical problem** is to use data coming out of the data generating process in order to make inferences about the true underlying parameters; in our example $\{\beta_0, \beta_1, \beta_2, ..., \beta_k\}$.

Least Squares Estimation

- **Estimator:** An estimator is a statistic (i.e. a function of the data) that is used to infer the value of an unknown parameter in a statistical model.
- **Residual:** A residual is the difference between an observed value y_i and the fitted (or predicted) value $\hat{y}_i$ implied by an estimator, often denoted $\hat{u}_i = y_i - \hat{y}_i$.
- Given the parametric model assumed in the previous slide, $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$ and some data drawn from it, $\{y_i, x_{i1}, x_{i2}, ..., x_{ik}\}_{i=1}^{n}$, a natural way of calculating our estimators for the true parameters $\{\beta_0, \beta_1, \beta_2, ..., \beta_k\}$ is by a process called least squares.
- The intuition behind least squares is that we want to choose estimators $\{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k\}$ to minimise the sum of square residuals, which can be thought of as the 'error' in our prediction.

Least Squares Estimation

- We can write the sum of square residuals in the following form, which we call our **objective function**,

$$\min_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)^2$$

- Naturally to find $\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k\}$ which minimise this objective function, we need to take the first derivative with respect to each of the estimators, set each derivative equal to 0 and then solve the resulting system of $k+1$ equations in $k+1$ unknowns.

$$0 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

$$0 = \sum_{i=1}^{n} x_{i1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

$$0 = \sum_{i=1}^{n} x_{i2} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

$$\vdots$$

$$0 = \sum_{i=1}^{n} x_{ik} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik} \right)$$

Least Squares Estimation

- We can simplify the system of equations in the previous slide considerably by recognising that,
  $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - ... - \hat{\beta}_k x_{ik} = y_i - \hat{y}_i = \hat{u}_i$.
- Let's rewrite it,
  $0 = \sum_{i=1}^{n} \hat{u}_i$
  $0 = \sum_{i=1}^{n} x_{i1} \hat{u}_i$
  $0 = \sum_{i=1}^{n} x_{i2} \hat{u}_i$
  $\vdots$
  $0 = \sum_{i=1}^{n} x_{ik} \hat{u}_i$
- This implies that the sample mean of $\hat{u}_i$ is always equal to 0 and that the sample covariance between $\hat{u}_i$ and each of the explanatory variables $x_{i1}, x_{i2}, ..., x_{ik}$ is also 0.
- This does **not** imply anything about the properties of the true population error term u_i. $\hat{u}_i$ only has these properties due to the way in which the least squares objective function is solved.

Simple Regression

- It is very difficult to explicitly solve this system of $k+1$ unknowns. If $k=1$, however, then we are running a **simple regression** and it is relatively straight-forward to solve for $\{\hat{\beta}_0, \hat{\beta}_1\}$.
- Let's look at the FOCs,
  $0 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} \right)$
  $0 = \sum_{i=1}^{n} x_{i1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} \right)$
- Solving these 2 equations in 2 unknowns, we obtain the following closed-form expressions,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1) y_i}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1$$

- The least squares prediction for observation i is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$. The least squares residual for observation i is $\hat{u}_i = y_i - \hat{y}_i$.

Partialling Out

- What if we are interested in an expression for a particular estimator $\hat{\beta}_1$ when $k > 1$?
- We might have some intuition that if we could "transform" $\hat{\beta}_1$ into a simple regression, then we could find a closed-form expression as in the previous slide. This is the idea behind partialling out.
- Steps:
  1. Regress x_1 on x_2, x_3, ..., x_k and calculate the residual $\hat{r}_1$. We can think of this residual as capturing the unique variation in x_1 that is not explained by x_2, x_3, ..., x_k.
     $x_{i1} = \hat{\alpha}_0 + \hat{\alpha}_2 x_{i2} + ... + \hat{\alpha}_k x_{ik} + \hat{r}_{i1}$
     $\hat{r}_1 = x_{i1} - \hat{x}_{i1}$
  2. Regress y on $\hat{r}_1$.
     $y_i = \hat{\lambda}_0 + \hat{\lambda}_1 \hat{r}_{i1} + \hat{\epsilon}_i$
  3. The resulting slope estimate $\hat{\lambda}_1$ is always equal to $\hat{\beta}_1$.

$$\hat{\beta}_1 = \hat{\lambda}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{r}_{i1} - 0)}{\sum_{i=1}^{n}(\hat{r}_{i1} - 0)^2} = \frac{\sum_{i=1}^{n} y_i \hat{r}_{i1}}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$$

MLR Assumptions

- Notice that, so far, we haven't made any assumptions about the underlying population from which $\{y_i, x_{i1}, x_{i2}, ..., x_{ik}\}_{i=1}^{n}$ is being drawn. We would calculate these same OLS estimators even if the true data generating process were not linear!
- However, the OLS estimators may not have "good properties" in the absence of certain assumptions over the population.
- What properties do we want our estimators to have?
- **Unbiasedness:** $\mathbb{E}(\hat{\beta}|x_1, x_2, ..., x_k) = \beta$
- **Consistency:** $\hat{\beta}$ converges to β as n goes to infinity.
- **Efficiency:** $\hat{\beta}$ has minimum variance (i.e. most precision) within the class of linear estimators.
- **Known Sampling Distribution:** $\hat{\beta}$ has a known finite sample distribution (this will allow us to do inference in finite sample).
- The next slide will cover the assumptions needed for our OLS estimators to achieve these properties.

MLR Assumptions

MLR.1 The model is linear in parameters:
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$

MLR.2 We have a random sample $\{(x_{i1}, x_{i2}, ..., x_{ik}, y_i)\}_{i=1}^{n}$ of size n following the model in MLR.1.

MLR.3 None of the independent variables is constant and there is no perfect multicollinearity (i.e. no exact linear relationship among the independent variables).

MLR.4 The error u_i has an expected value of zero conditional on all x_i
$\mathbb{E}(u_i | x_{i1}, x_{i2}, ..., x_{ik}) = 0$ for $i = 1, ..., n$.

MLR.5 The error u_i homoscedastic, i.e. it has the same variance for all x_i
$\text{Var}(u_i | x_{i1}, x_{i2}, ..., x_{ik}) = \sigma^2$

MLR.6 The error u_i is independent of the explanatory variables $x_{i1}, x_{i2}, ..., x_{ik}$ and is normally distributed with mean 0 and variance σ^2
$u_i \sim \mathcal{N}(0, \sigma^2)$

Unbiasedness

- Why is the OLS estimator unbiased under these assumptions?
- Let's consider the OLS estimator in the simple regression model. First note,

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1) y_i}{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2} \\
&= \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1)(\beta_0+\beta_1 x_{i1}+u_i)}{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2} \\
&= \beta_0 \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1)}{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2} + \beta_1 \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1) x_{i1}}{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2} + \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1) u_i}{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2} \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1) u_i}{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2}
\end{aligned}$$

- Then,

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}_1 | x_1\right) &= \mathbb{E}\left(\beta_1 + \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1) u_i}{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2} | x_1\right) \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_{i1}-\bar{x}_1) \mathbb{E}(u_i | x_1)}{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2} \\
&= \beta_1
\end{aligned}$$

Omitted Variable Bias

- What if we omit a relevant variable? Will $\hat{\beta}$ still be unbiased?
- Let's assume that the actual population relationship between y and a set of x's is,
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$
  satisfying the assumptions MLR.1 to MLR.4
- Denote the estimator for β_1 and β_k by running this correctly specified regression as $\hat{\beta}_1$ and $\hat{\beta}_k$ respectively.
- But suppose that we accidentally specify the the model,
  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{k-1} x_{ik-1} + u_i$
  whereby we omit a relevant variable x_k.
- Denote the estimator for β_1 obtained in this misspecified model by $\tilde{\beta}_1$.

- By the "partialling out" procedure we know that $\tilde{\beta}_1$ can be written as,

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} \tilde{r}_{i1} y_i}{\sum_{i=1}^{n} \tilde{r}_{i1}^2}$$

  where $\tilde{r}_{i1}$ are the sample residuals from the regression, $x_{i1} = \alpha_0 + \alpha_2 x_{i2} + \alpha_3 x_{i3} + ... + \alpha_{k-1} x_{ik-1} + r_i$

- It can be shown that,

$$\begin{aligned} \tilde{\beta}_1 &= \hat{\beta}_1 + \hat{\beta}_k \frac{\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik}}{\sum_{i=1}^{n} \tilde{r}_{i1}^2} \\ &= \hat{\beta}_1 + \hat{\beta}_k \tilde{\delta}_1 \end{aligned}$$

  where $\tilde{\delta}_1$ is the "partialling out" coefficient for x_1 from the regression of x_k on x_1, x_2, ..., x_{k-1}.

- The bias in $\tilde{\beta}_1$ is
  $\text{Bias}(\tilde{\beta}_1) = \text{E}(\tilde{\beta}_1 - \beta_1 | x_1, x_2, ..., x_k) =$
  $\text{E}(\hat{\beta}_1 + \hat{\beta}_k \tilde{\delta}_1 - \beta_1 | x_1, x_2, ..., x_k) = \beta_1 + \beta_k \tilde{\delta}_1 - \beta_1 = \beta_k \tilde{\delta}_1$.

Known Sampling Distribution

- Why does the OLS estimator have a known sampling distribution under these assumptions?
- Let's again consider the simple regression model.
- Since we know that $u_i \sim \mathcal{N}(0, \sigma^2)$ by MLR.5, it follows from the properties of the normal distribution that,

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \text{Var}\left(\frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)u_i}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}\right)\right)$$

where $\text{Var}(\frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)u_i}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}$ from MLR.5.

- The **standard error** $se(\hat{\beta}_1)$ is the square root of the variance.
- Then the test statistic,

$$\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim \mathcal{N}(0, 1)$$

Known Sampling Distribution

- However, this expression is a function of the true variance of u denoted by σ^2, which we cannot observe.
- We must therefore use an estimator for σ^2 (not derived here) which is usually taken to be

$$\widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{u}_i^2$$

where, for the simple regression case, $k = 2$.

- We denote the estimator for the standard error of $\hat{\beta}_j$ using $\widehat{\sigma}^2$ as $\widehat{se}(\hat{\beta}_j)$.
- Note that again by MLR.6,

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{se}(\hat{\beta}_j)} \sim t_{n-k-1}$$

which is the t-distribution on $n-k-1$ degrees of freedom.

Relaxing MLR.6

- Recall:
  (Lindeberg–Lévy) Central Limit Theorem: Consider n random variables $X_1, X_2, ..., X_n$ drawn independently from the same distribution with expected value $\mathbb{E}(X_i) = \mu$ and variance $\text{Var}(X_i) = \sigma^2 < \infty$. Then the Central Limit Theorem says that as n converges to infinity $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to $\mathcal{N}(0, \sigma^2)$.
- By the Central Limit Theorem,

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \xrightarrow{a} \mathcal{N}(0, 1)$$

  where $\text{se}(\hat{\beta}_j)$ is the usual OLS standard error.
- This tells us that our test statistic from the previous slides is "well approximated" by a normal distribution as long as the sample size is large.

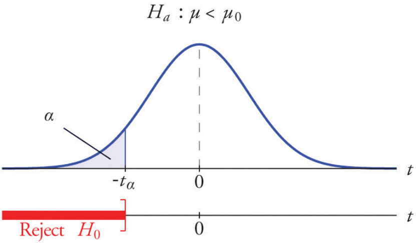
Hypothesis Testing: t-test

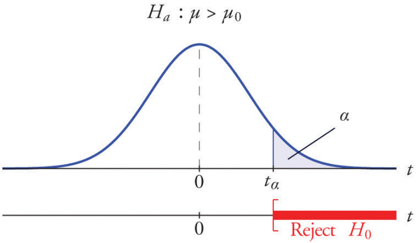
- It is very useful to know that the **t-statistic** $t = \frac{\hat{\beta}_j - \beta_j}{\widehat{se}(\hat{\beta}_j)} \sim t_{n-k-1}$ since the t-distribution, given a **null hypothesis** value of β_j, does not depend on any of the unobserved true parameter values.
- This means that we can start to ask questions about the likelihood of observing a given estimate $\hat{\beta}_j$ given a null hypothesis value of β_j. This is the main intuition behind hypothesis testing.
- The simple hypothesis test is related to the value of a single parameter value,
- **Two-Sided Hypothesis Test:** $\mathrm{H}_0 : \beta_j = a$ and $\mathrm{H}_1 : \beta_j \neq 0$ where a is some fixed constant, H_0 is the null hypothesis and H_1 is the **alternative hypothesis**.
- **One-Sided Hypothesis Test:** $\mathrm{H}_0 : \beta_j > a$ and $\mathrm{H}_1 : \beta_j < a$ or $\mathrm{H}_0 < a$ and $\mathrm{H}_1 : \beta_j > a$.

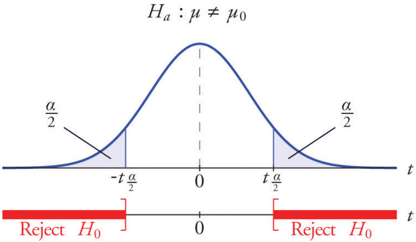
Hypothesis Testing: t-test

- How do we actually test these null and alternative hypotheses?
- We need to specify one more parameter, called the **significance level** and usually denoted by α. α represents the probability under the null of rejecting the null hypothesis when it is actually true.
- From the significance level α we can obtain a **critical value** T which is a value on the t-distribution such that, if we observed a t-statistic more extreme than this, we would reject the null hypothesis.
- For a two-sided test, we obtain a critical value $T_{\alpha/2}$ such that if $|t| > |T_{\alpha/2}|$ then we would reject the null hypothesis, where $|\cdot|$ denotes absolute value.
- For a one-sided test $\mathrm{H}_0 : \beta_j > a$ ($\mathrm{H}_0 : \beta_j < a$), we obtain a critical value T_α such that if $t > T_\alpha$ ($t < T_\alpha$) we would reject the null hypothesis.

Hypothesis Testing: t-test







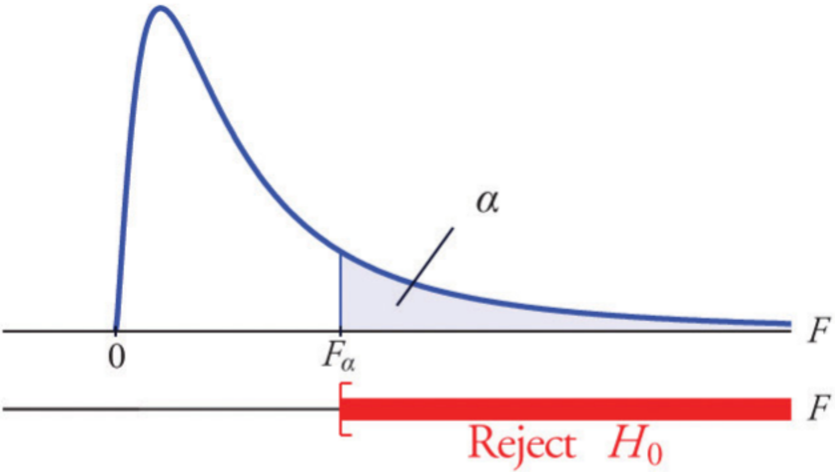
Hypothesis Testing: F-test

- What if we want to test hypotheses over multiple estimates? For example, $\text{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$.
- For hypotheses of this form we will use an **F-test**. The **F-statistic** takes the form,

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(R^2_{ur} - R^2_r)/q}{(1-R^2_{ur})/(n-k-1)} \sim F_{q,n-k-1}$$

  - SSR_r denotes the sum of squared residuals of the **restricted model** in which we assume the null hypothesis is true. R^2_r is the R^2 from this restricted model.
  - SSR_{ur} denotes the sum of squared residuals of the **unrestricted model** in which we make no assumptions over the coefficients in the null hypothesis. R^2_{ur} is the R^2 from this unrestricted model.
  - q is the number of restrictions being tested. In the example above, it is 3.
  - $n-k-1$ is the degrees of freedom of the unrestricted model.

Hypothesis Testing: F-test



Heteroscedasticity

- All of our hypothesis tests so far have assumed **homoscedasticity** of the error term u. This says that it has the same variance for all x_i $\text{Var}(u_i|x_{i1}, x_{i2}, ..., x_{ik}) = \sigma^2$
- We define **heteroscedasticity** as allowing the variance of u_i to vary across x_i such that $\text{Var}(u_i|x_{i1}, x_{i2}, ..., x_{ik}) = \sigma_i^2$, where the i subscript is what allows the variance to differ across observations.
- While heteroscedasticity does not affect the consistency or unbiasedness of the OLS estimator, it does make the OLS estimator inefficient and means that our t-statistics and F-statistics may not follow their respective distributions.

Heteroscedasticity

- What can we do?
- One solution is to use **White (robust) standard errors**. For the simple regression model $y_i = \beta_0 + \beta_1 x_i + u$ under heteroscedasticity, the variance of the estimator is $\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2}$
- A natural way to find an estimator for the variance is to find an estimator for σ_i^2. What is the simplest estimator? $\hat{u}_i^2$ of course! This looks like $\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \hat{u}_i^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2}$

Heteroscedasticity: GLS and FGLS

- Suppose u_i is the true error term with variance σ_i^2. We can think that there exists a transformation p_i such that $u_i^* = p_i u_i$ has constant variance σ^2. This is known as **Generalised Least Squares (GLS)**.
- A popular choice of transformation is $p_i = 1/\sqrt{h_i}$ where $\sigma_i^2 = \mathbb{E}(u_i^2) = h_i \sigma^2$.
- Then if we transform the data in our OLS regression by $p_i = 1/\sqrt{h_i}$ we have that $\mathbb{E}(u_i^{*2}) = \mathbb{E}(u_i^2/h_i) = \mathbb{E}(u_i^2)/h_i = \sigma^2 h_i/h_i = \sigma^2$.
- The problem with this approach is that we need to know the functional form of h_i.
- A more tractable approach is to model heteroscedasticity as a function of $\boldsymbol{x}_i$. This is called **Feasible Generalised Least Squares (FGLS)**. Here are the steps,
  1. Estimate the original model by OLS and recover $\hat{u}_i$.
  2. Estimate the model $ln(\hat{u}_i) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + ... + \alpha_k x_{ik} + e$
  3. Set $\hat{h}_i = \exp(\hat{\alpha}_0 + \hat{\alpha}_1 x_{i1} + ... + \hat{\alpha}_{ik})$
  4. Transform the model using weights $1/\sqrt{\hat{h}_i}$ and estimate by OLS.

Instrumental Variables

- Instrumental variables are used to deal with **endogeneity** in multiple regression models, which is when the error term u_i is correlated with the explanatory variables $x_{i1}, x_{i2}, ..., x_{ik}$. This violates MLR.4.
- An **instrumental variable** z_i is a variable which satisfies the following properties:

$$\begin{array}{ll} (1) & \text{Cov}(z_i, u_i) = 0 \quad \text{exogeneity condition} \\ (2) & \text{Cov}(z_i, x_{i1}) \neq 0 \quad \text{relevance condition} \end{array}$$

- In the simple regression model, we can use the instrument to obtain a consistent estimator of β_1 of the following form:

$$\hat{\beta}_1^{IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_{i1} - \bar{x}_1)} = \frac{\widehat{\text{Cov}}(z,y)}{\widehat{\text{Cov}}(z,x)}$$

- This is called the **IV estimator**.

Two-Stage Least Squares (2SLS): Procedure

- When we have more instruments $z_1, z_2, ..., z_l$ than endogenous variables x_1, we use the Two-Stage Least Squares (2SLS) estimator.
- The reason for this name is that the estimator can be obtained by running two OLS regressions.
    1. Run OLS regression on the first stage regression

    $$x_{i1} = \pi_0 + \pi_1 z_{i1} + \pi_2 z_{i2} + \epsilon_i$$

    and calculate the predicted values $\hat{x}_{i1}$.
    2. Run OLS in the **second stage regression** which replaces x_{i1} with $\hat{x}_{i1}$

    $$y_i = \beta_0 + \beta_1 \hat{x}_{i1} + u_i$$

    obtaining $\hat{\beta}_1^{2SLS}$.
- It can be shown that $\hat{\beta}_1^{2SLS} = \hat{\beta}_1^{IV}$.
- Between step 1 and step 2 we can also compute an F-test for $H_0 : \pi_1 = \pi_2 = 0$ to test the relevance condition.

In-Class Exercise

In this in-class exercise we are going to analyse a dataset using all of the methods covered in this course. Load the dataset CARD.dta from my website (matthewtudball.com). This dataset contains information on log wages (*lwage*), education (*educ*), IQ (*IQ*) and other variables.

1. Regress *lwage* on *educ* and *exper* and report the coefficient on *educ*.
2. Is the coefficient on *educ* statistically significant at the 5% level? Are *educ* and *exper* both jointly significant at the 5% level?
3. Suppose that the previous regression is misspecified and *IQ* is a relevant omitted variable. Calculate the coefficient on *educ* from the regression of *lwage* on *educ*, *exper* and *IQ* using partialling out (Hint: Look at slide 8).
4. Also obtain the coefficient on *IQ* by partialling out. What sort of bias is there in *educ* (i.e. upward/downward)? What do the partialling out coefficient on *IQ* tell you about the relationship between *educ* and *IQ* (i.e. positive/negative)?

In-Class Exercise

5. Suppose we think that there is heteroscedasticity of the form $\sigma_i^2 = \sigma^2 exper_i^2$. Implement GLS using an appropriate transformation.
6. Suppose we now think that *educ* is endogenous and *nearc*4 is a valid instrument. Calculate the IV estimate using Stata (without including *exper* or *IQ* as explanatory variables).
7. Recall that we can calculate the IV estimator by dividing the sample covariance between z and y by the sample covariance between z and x (see slide 25). Calculate the IV estimate manually in this way. Check that your manual estimate is the same as Stata's estimate from Q6.