# IIA-3 Econometrics: Supervision 5

Emre Usenmez

Lent Term 2025

**Topics Covered**

**Faculty Qs:**

**Supplementary Qs:** Endogeneity, measurement errors; simultaneous equations;

**Related Reading:**

Dougherty, *Introduction to Econometrics*, $5^{th}$ ed, OUP

Chapter 6: Specification of Regression Variables

Chapter 8: Stochastic Regressors and Measurement Errors

Chapter 9: Simultaneous Equations Estimation

Wooldridge J M (2021) *Introductory Econometrics: A Modern Approach*, $7^{th}$ ed,

Chapter 9: More on Specification and Data Issues

Chapter 15: Instrumental Variables in Estimation and Two Stage Least Squares

Chapter 16: Simultaneous Equations Models

Gujarati, D N and Porter, D (2009) *Basic Econometrics*, $7^{th}$ International ed, McGraw-Hill

Chapter 13: Econometric Modeling: Model Specification and Diagnostic Testing

Chapter 20: Simultaneous-Equation Methods

Gujarati, D (2022) *Essentials of Econometrics*, $5^{th}$ ed, Sage

Chapter 7: Model Selection: Criteria and Tests

Very grateful to Dr Oleg Kitov and Dr Clive Lawson for the very informative stylized answers to previous iterations of the supervision questions.

# FACULTY QUESTIONS

## QUESTION 1

# SUPPLEMENTARY QUESTIONS

## QUESTION 1

**(a) Explain what is meant when it is said that the explanatory variables and the disturbance term in a regression equation are not independent. What can be said about the properties of the OLS estimates in this case?**

**Answer:** If the disturbance term and the explanatory variables are not independent then they are correlated. Those explanatory variables that are correlated with the error term are called *endogenous variables*.

Since unibasedness depends on $Cov(\varepsilon_i, X_i) = 0$, this dependency between the error term and the explanatory variables would yield biased estimates.

---

**(b) Suppose that $Y_i = \alpha + \beta X_i + \lambda W_i + \varepsilon_i$ where there also exists a relationship between $X_i$ and $W_i$ of the form $W_i = \rho + \phi X_i + v_i$. Show that if $Y_i$ is estimated using only the $X_i$ variable then the estimate of $\beta$ obtained is biased. Under what circumstances would this estimate of $\beta$ be biased downwards?**

**Answer:** Let's start by substituting in the latter expression into the former:

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \lambda W_i + \varepsilon_i \\ &= \alpha + \beta X_i + \lambda(\rho + \phi X_i + v_i) + \varepsilon_i \\ &= (\alpha + \lambda\rho) + (\beta + \lambda\phi)X_i + (\lambda v_i + \varepsilon_i) \\ &= \gamma_0 + \gamma_1 X_i + u_i \end{aligned}$$

Now notice that both $W_i$ and $u_i$ depend on $v_i$. This means the assumption of exogeneity, i.e. independence between the explanatory variable and the disturbance term, would be violated when $Y$ is regressed on $X$. As a result, $\hat{\gamma}_1$ would be *inconsistent* and *biased*.

To see this, start by looking at the expression for the regression coefficient $\gamma_1$

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \gamma_1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Since $X$ and $u$ are not distributed independently of each other, we can't summarize the distribution of the error term, or obtain an expresssion for its expected value. The most we can do is to determine how the error term would behave if the sample were very large.

However, neither the numerator nor the denominator tends to a particular limit as $n$ increases. To get around this, we can divide both the numerator and the denominator by $n$. Then the probability limit of $\hat{\gamma}_1$ as $n$ tends to infinity becomes

$$plim(\hat{\gamma}_1) = \gamma_1 + \frac{plim\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})\right)}{plim\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)}$$

$$= \gamma_1 + \frac{Cov(X, u)}{Var(X)}$$

$$= \gamma_1 + \frac{Cov\left((\frac{W - \rho - v}{\phi}), (\lambda v + \varepsilon)\right)}{Var\left(\frac{W - \rho - v}{\phi}\right)}$$

$$= \gamma_1 + \frac{Cov\left(\frac{W - \rho}{\phi}, \lambda v\right) + Cov\left(\frac{W - \rho}{\phi}, \varepsilon\right) + Cov\left(\frac{-v}{\phi}, \lambda v\right) + Cov\left(\frac{-v}{\phi}, \varepsilon\right)}{Var\left(\frac{W - \rho - v}{\phi}\right)}$$

If we then assume that the error term in the original model, $\varepsilon$, is distributed independently of $W$, and the error term in the second model, $v$, is distributed independently of $W$ and $\varepsilon$, then the first, second and fourth terms of the numerator are zero. Then

$$plim(\hat{\gamma}_1) = \gamma_1 + \frac{0 + 0 + Cov\left(\frac{-v}{\phi}, \lambda v\right) + 0}{Var\left(\frac{W - \rho - v}{\phi}\right)}$$

$$= \gamma_1 + \frac{-\frac{\lambda}{\phi}Var(v)}{Var\left(\frac{W - \rho}{\phi}\right) + Var\left(\frac{-v}{\phi}\right) + 2Cov\left(\frac{W - \rho}{\phi}, \frac{-v}{\phi}\right)}$$

$$= \gamma_1 + \frac{-\frac{\lambda}{\phi}Var(v)}{\frac{1}{\phi^2}Var(W) + \frac{1}{\phi^2}Var(v) + 0}$$

$$= \gamma_1 - \lambda\phi\frac{Var(v)}{Var(W) + Var(v)}$$

Thus $\hat{\gamma}_1$ is subject to bias whereby the bias is downwards if $\lambda\phi$ is positive.

**(c) Explain why measurement errors and simultaneous equations might also involve correlation of this kind (give simple algebraic examples of each).**

**Measurement Error:** Relatively frequently in economics, the variables we use have not been measured precisely. These may be due to inaccuracies in the surveys or a data available corresponds to a slightly different concept than the variable in our model. Milton Friedman's critique of the consumption function is an example of the latter.[1]

Consider the following model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where $Y_i$ is the permanent consumption expenditure,[2] $X_i$ current income, and $u_i$ stochastic disturbance term.

Since $Y_i$ is not measurable because it is subjectively determined by individual's recent experience and future expectations, we can instead use an observable consumption expenditure variable $Y_i^*$ such that

$$Y_i^* = Y_i + \varepsilon_i$$

where $\varepsilon_i$ are the errors of measurement in $Y_i$. Therefore, we instead estimate the following:

$$\begin{aligned} Y_i* &= (\beta_0 + \beta_1 X_i + u_i) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_i + (u_i + \varepsilon_i) \\ &= \beta_0 + \beta_1 X_i + v_i \end{aligned}$$

where $v_i = u_i + \varepsilon_i$ is a composite error term that contains both the population error term and the measurement error term.

If the classical linear regression assumptions, specifically $\mathbb{E}(u_i) = \mathbb{E}(v_i) = 0$ and $Cov(X_i, u_i)$, as well as $Cov(X_i, \varepsilon_i)$ hold true, then $\hat{\beta}_1$ will be an <u>unbiased</u> estimator of the true $\beta_1$ but the variances, and therefore the standard errors, of $\beta_1$ estimated from this equation will be different because

$$Var(\hat{\beta}_1) = \frac{Var(v)}{\sum(X_i - \bar{X})^2} = \frac{Var(u_i) + Var(\varepsilon_i)}{\sum(X_i - \bar{X})^2} \quad > \quad \frac{Var(u_i)}{\sum(X_i - \bar{X})^2}$$

Therefore, if there is measurement error in the explanatory variable, we will still obtain unbiased estimates of the parameters and their variances, but the estimated variances will be bigger than in the case where there are no such measurement errors.

The situation is different if there is a measurement error in the dependent variable instead. Consider again the model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

but this time $Y - i$ is the *current* consumption expenditure, $X_i$ is *permanent* income, and $u_i$ is the stochastic disturbance term.

Since this time $X_i$ is not measurable, we can instead use an observable income variable $X_i^*$ such that

$$X_i^* = X_i + \varepsilon_i$$

where $\varepsilon_i$ are the errors of measurement in $X_i$. Therefore, we instead estimate the following:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ &= \beta_0 + \beta_1 X_i^* - \beta_1 \varepsilon_i + u_i \\ &= \beta_0 + \beta_1 X_i^* + v_i \end{aligned}$$

[1] Firedman, M (1957) *A Theory of the Consumption Function*, Princeton University Press

[2] permanent consumption expenditure is a term used by Milton Friedman to refer to the level of consumption justified by the level of permanent income. Permanent income can be thought of as a medium term income in that it is the amount that the individual can can more or less depend on for the foreseeable future.

where $v_i = u_i - \beta_1 \varepsilon_i$ is a composite error term that contains both the population error term and the measurement error term.

In this case, even if we assume that the assumptions $\mathbb{E}(u_i) = \mathbb{E}(v_i) = 0$ and $Cov(v_i, u_i)$ hold, we cannot assume that the composite error term $v_i$ is independent of $X_i^*$ because

$$\begin{aligned} Cov(X_i^*, v_i) &= \mathbb{E}[v_i - \mathbb{E}(v_i)][X_i^* - \mathbb{E}(X_i^*)] \\ &= \mathbb{E}(u_i - \beta_1 \varepsilon_i - 0)(X_i + \varepsilon_i - X_i) \\ &= \mathbb{E}(u_i - \beta_1 \varepsilon_i)(\varepsilon_i) \\ &= \mathbb{E}(-\beta_1 \varepsilon_i^2) \\ &= -\beta_1 Var(\varepsilon_i) \end{aligned}$$

Thus $X_i^*$ and $v_i$ are correlated which violates the exogeneity assumption. If this assumption is violated, as shown in part(b) above, the OLS estimators are biased and inconsistent, meaning that they remain biased even if the sample size increases indefinitely.

Notice that this correlation between $X_i^*$ and $v_i$ will cause problems because it means $X_i$ and $\varepsilon_i$ are correlated since $v_i = u_i - \beta_1 \varepsilon_i$. To determine the amount of inconsistency in the OLS we again take the probability limit of $\hat{\beta}_1$:

$$\begin{aligned} plim(\hat{\beta}_1) &= \beta_1 + \frac{Cov(X_1^*, v_i)}{Var(X_1^*)} \\ &= \beta_1 + \frac{-\beta_1 Var(\varepsilon_i)}{Var(X_1) + Var(\varepsilon_i)} \\ &= \beta_1 \left( 1 - \frac{\sigma_\varepsilon^2}{\sigma_{X_1}^2 + \sigma_\varepsilon^2} \right) \\ &= \beta_1 \left( \frac{\sigma_{X_1}^2 + \sigma_\varepsilon^2 - \sigma_\varepsilon^2}{\sigma_{X_1}^2 + \sigma_\varepsilon^2} \right) \\ &= \beta_1 \left( \frac{\sigma_{X_1}^2}{\sigma_{X_1}^2 + \sigma_\varepsilon^2} \right) \end{aligned}$$

Notice that the term multiplying $\beta_1$ is the ratio of $Var(X_1)$ to $Var(X_1^*)$. It is always less than 1, which means $plim(\hat{\beta}_1)$ is always closer to 0 than $\beta_1$. This is called the attenuation bias in OLS: on average, the estimated OLS effect will be attenuated. In particular, if $\beta_1 > 0$, then $\hat{\beta}_1$ will tend to underestimate $\beta_1$.

**Simultaneous Equations:** Another important form of explanatory variables endogeneity is *simultaneity*, which occurs when an explanatory variable and the dependent variable is jointly determined. The main way for estimating simultaneous equations is the same as those for the omitted variables problem and measurement error problem - instrumental variables (IV).

Consider the following model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where $Y_i$ is annual prices growth rate, and $X_i$ is the wages growth rate. Suppose workers want increase in their wages as the prices rise to protect their real wages, but their ability to do so depends on the unemployment rate $J$ in a following manner

$$X_i = \alpha_0 + \alpha_1 Y_i + \alpha_2 J_i + v_i$$

where $u_i$ and $v_i$ are stochastic disturbance terms. Accordingly, $Y_i$ and $X_i$ are both endogeneous variables since their values are determined by the interaction of the relationships in the model, and $J_i$ is an exogeneous variable since its values are determined externally. These equations are called *structural equations*, and if we write the endogeneous variables in terms of exogeneous ones and the disturbance terms, then they are called *reduced form equations*.

To derive the reduced form equation for $Y_i$ and $X_i$ we start with the structural equations, just as we did for measurement errors:

$$\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + u_i \\
&= \beta_0 + \beta_1(\alpha_0 + \alpha_1 Y_i + \alpha_2 J_i + v_i) + u_i \\
(1-\beta_1\alpha_1)Y_i &= (\beta_0 + \beta_1\alpha_0) + \beta_1\alpha_2 J_i + (\beta_1 v_i + u_i) \\
Y_i &= \frac{\beta_0 + \beta_1\alpha_0 + \beta_1\alpha_2 J_i + \beta_1 v_i + u_i}{1-\beta_1\alpha_1}
\end{aligned}$$

and for $X_i$ the reduced form equation is

$$\begin{aligned}
X_i &= \alpha_0 + \alpha_1 Y_i + \alpha_2 J_i + v_i \\
&= \alpha_0 + \alpha_1(\beta_0 + \beta_1 X_i + u_i) + \alpha_2 J_i + v_i \\
(1-\alpha_1\beta_1)X_i &= (\alpha_0 + \alpha_1\beta_0) + \alpha_2 J_i + (\alpha_1 u_i + v_i) \\
X_i &= \frac{\alpha_0 + \alpha_1\beta_0 + \alpha_2 J_i + \alpha_1 u_i + v_i}{1-\alpha_1\beta_1}
\end{aligned}$$

It can be observed that $Y_i$ indirectly depends on the exogeneous variable $J_i$ and the disturbance term $v_i$ through $X_i$. Similarly, $X_i$ depends on $u_i$ indirectly, and $J_i$ and $v_i$ directly. These dependencies mean the OLS would yield inconsistent and biased estimates. To see this lets look at the expression for $\beta_1$:

$$\begin{aligned}
\hat{\beta_1}^{OLS} &= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{(X_i-\bar{X})^2} \\
&= \frac{\sum_{i=1}^{n}(X_i-\bar{X})\big[(\beta_0+\beta_1 X_i + u_i) - (\beta_0 + \beta_1\bar{X}_i + \bar{u}_i)\big]}{(X_i-\bar{X})^2} \\
&= \frac{\sum_{i=1}^{n}\Big((X_i-\bar{X})\beta_1(X_i-\bar{X}) + (X_i-\bar{X})(u_i-\bar{u}_i)\Big)}{(X_i-\bar{X})^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(X_i-\bar{X})(u_i-\bar{u}_i)}{(X_i-\bar{X})^2}
\end{aligned}$$

Since the error term is a nonlinear function of $u_i$, directly, and $v_i$, indirectly, we cannot obtain an analytical expression for its expected value. This is why we look at its probability limit, where we use the rule that the probability limit of a ratio is equal to the ration of probability limit of the numerator to the probability limit of the denominator. In the current form the expression for $\hat{\beta_1}^{OLS}$ does not have a probability limit. For this, we need to divide both the numerator and the denominator by $n$.

$$plim(\hat{\beta}_1) = \beta_1 + \frac{plim\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(u_i-\bar{u})\Big)}{plim\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\Big)} = \beta_1 + \frac{Cov(X,u)}{Var(X)}$$

$$= \beta_1 + \frac{Cov\left(\dfrac{\alpha_0 + \alpha_1\beta_0 + \alpha_2 J_i + \alpha_1 u_i + v_i}{1 - \alpha_1\beta_1}\ ,\ u_i\right)}{Var\left(\dfrac{\alpha_0 + \alpha_1\beta_0 + \alpha_2 J_i + \alpha_1 u_i + v_i}{1 - \alpha_1\beta_1}\right)}$$

Since $\frac{\alpha_0+\alpha_1\beta_0}{1-\alpha_1\beta_1}$ is a constant, its covariance with $u_i$ is zero: $Cov(\frac{\alpha_0+\alpha_1\beta_0}{1-\alpha_1\beta_1}, u) = 0$. Similarly, $J_i$ is exogeneous, or at least we assume it is, so $Cov(\frac{\alpha_2}{1-\alpha_1\beta_1}J_i, u_i) = 0$, and if we assume that the disturbance terms in the structural equations, $u_i$ and $v_i$, are independent, then $Cov(\frac{1}{1-\alpha_1\beta_1}v_i, u_i) = 0$. Then,

$$\begin{aligned}
plim(\hat{\beta}_1) &= \beta_1 + \frac{0 + 0 + \dfrac{\alpha_1}{1-\alpha_1\beta_1}Var(u_i) + 0}{\dfrac{1}{(1-\alpha_1\beta_1)^2}\Big(Var(\alpha_0 + \alpha_1\beta_0) + Var(\alpha_2 J_i + \alpha_1 u_i + v_i)\Big)} \\
&= \beta_1 + \frac{\dfrac{\alpha_1}{1-\alpha_1\beta_1}\sigma_u^2}{\dfrac{1}{(1-\alpha_1\beta_1)^2}\begin{pmatrix} Var(\alpha_2 J_i) + Var(\alpha_1 u_i) + Var(v_i) \\ +2Cov(\alpha_2 J_i, \alpha_1, u_i) + 2Cov(\alpha_2 J_i, v_i) + 2Cov(\alpha_1, v_i)\end{pmatrix}} \\
&= \beta_1 + \frac{(1-\alpha_1\beta_1)(\alpha_1\sigma_u^2)}{\alpha_2^2\sigma_J^2 + \alpha_1^2\sigma_u^2 + \sigma_v^2}
\end{aligned}$$

Thus $\hat{\beta}_1^{OLS}$ is an inconsistent and biased estimator of $\beta_1$. Since variances are always positive, and asusming the coefficient for annual price growth rate, $\alpha_1$ is positive, then the direction of the bias depends on $(1 - \alpha_1\beta_1)$.

## QUESTION 2

**Consider the following population regression function (PRF) in which education and ability both positively affect the wage received:**

$$log(wage) = \alpha + \beta_1\ educ + \beta_2\ ability + \varepsilon \tag{1}$$

**(a) If there is no direct measurement of ability and equation (1) is estimated simply using OLS on *educ*, would you expect your estimate $\beta_1$ to be biased upwards or downwards?**

**Answer:** If we estimate equation (1) via OLS on *educ* only, then we are estimating the model

$$log(wage) = \beta_0 + \beta_1\ educ + u$$

where $u = \beta_2\ ability + \varepsilon$ and the estimator $\hat{\beta}_1$ is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\Big(educ_i - \overline{educ}\Big)\Big(log(wage_i) - \overline{log(wage)}\Big)}{\sum_{i=1}^{n}\Big(educ_i - \overline{educ}\Big)^2}$$

By definition, $\hat{\beta}_1$ is an unbiased estimator if and only if $\mathbb{E}(\hat{\beta}_1) = \beta_1$. If we expand this expression for $\hat{\beta}_1$ we get:

$$\hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} \Big(educ_i - \overline{educ}\Big)\Big((\alpha + \beta_1\ educ_i + \beta_2\ ability_i + \varepsilon_i) - (\alpha + \beta_1\ \overline{educ}_i + \beta_2\ \overline{ability}_i + \bar{\varepsilon}_i)\Big)}{\displaystyle\sum_{i=1}^{n} \Big(educ_i - \overline{educ}\Big)^2}$$

$$= \frac{\beta_1 \displaystyle\sum_{i=1}^{n} \Big(educ_i - \overline{educ}\Big)^2 + \beta_2 \sum_{i=1}^{n} \Big((educ_i - \overline{educ})(ability_i - \overline{ability}_i)\Big) + \sum_{i=1}^{n} \Big(educ_i - \overline{educ}\Big)\Big(\varepsilon_i - \bar{\varepsilon}_i\Big)}{\displaystyle\sum_{i=1}^{n} \Big(educ_i - \overline{educ}\Big)^2}$$

$$= \beta_1 + \beta_2 \frac{\displaystyle\sum_{i=1}^{n} \Big((educ_i - \overline{educ})(ability_i - \overline{ability}_i)\Big)}{\displaystyle\sum_{i=1}^{n} \Big(educ_i - \overline{educ}\Big)^2} + \frac{\sum_{i=1}^{n} \Big(educ_i - \overline{educ}\Big)\Big(\varepsilon_i - \bar{\varepsilon}_i\Big)}{\displaystyle\sum_{i=1}^{n} \Big(educ_i - \overline{educ}\Big)^2}$$

All of this analysis is conditional on the sample values of the explanatory variables. When we take expectations, the first two terms are unaffected and the third term is zero. That is,

$$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{\displaystyle\sum_{i=1}^{n} \Big((educ_i - \overline{educ})(ability_i - \overline{ability}_i)\Big)}{\displaystyle\sum_{i=1}^{n} \Big(educ_i - \overline{educ}\Big)^2}$$

To see the intuition behind this notice that because we omit *ability* from the regression model, *educ* will not only have a direct effect on $log(wage)$ but also a proxy effect when it mimics the effect of *ability*. This indirect effect of *educ* on $log(wage)$ depends on two things:

- the extent to which *educ* can mimic *ability*, i.e. the extent to which *educ* can explain *ability*, and
- effect of *ability* on $log(wage)$, which is $\beta_2$.

The extent of *ability* being explained by *educ* is determined by the slope coefficient of

$$ability = \gamma_0 + \gamma_1\ educ + v$$

where $\hat{\gamma}_1$ is given by

$$\hat{\gamma}_1 = \frac{\displaystyle\sum_{i=1}^{n} (educ_i - \overline{educ}_i)(ability_i - \overline{ability}_i)}{\displaystyle\sum_{i=1}^{n} (educ_i - \overline{educ}_i)^2}$$

Since the effect of *ability* on $log(wage)$ is $\beta_2$, we combine these two factors to obtain the indirect effect of *educ* on $log(wage)$:

$$\beta_2 \hat{\gamma}_1 = \beta_2 \frac{\displaystyle\sum_{i=1}^{n} (educ_i - \overline{educ}_i)(ability_i - \overline{ability}_i)}{\displaystyle\sum_{i=1}^{n} (educ_i - \overline{educ}_i)^2}$$

Finally, since the direct effect of *educ* on $Y$ is $\beta_1$, when we regress $log(wage)$ on *educ* only, omitting *ability*, the coefficient of *educ* is then the combination of direct and indirect effects on $log(wage)$:

$$\beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(educ_i - \overline{educ_i})(ability_i - \overline{ability_i})}{\sum_{i=1}^{n}(educ_i - \overline{educ_i})^2} + \text{sampling error}$$

If *educ* and *ability* are nonstochastic, then the expected value of the coefficient will be the sum of the first two terms. The presence of the second term implies that in general the expected value of the coefficient will be different from the true value $\beta_1$ and therefore biased.

To determine the direction of the bias, first notice that $\sum(educ_i - \overline{educ})^2$ will always be positive, which means the direction of the bias will depend on the signs of $\beta_2$ and $\sum(educ_i - \overline{educ_i})(ability_i - \overline{ability_i})$.

Also notice that $\sum(educ_i - \overline{educ_i})(ability_i - \overline{ability_i})$ is the same as the numerator of the sample correlation $r$ between *educ* and *ability*:

$$r_{educ,ability} = \frac{\sum_{i=1}^{n}(educ_i - \overline{educ_i})(ability_i - \overline{ability_i})}{\sqrt{\sum_{i=1}^{n}(educ_i - \overline{educ_i})^2 \sum_{i=1}^{n}(ability_i - \overline{ability_i})^2}}$$

Since the denominator of $r_{educ,ability}$ is always positive, the sign of $\sum(educ_i - \overline{educ_i})(ability_i - \overline{ability_i})$ then is the same as the sign of the correlation coefficient, $r_{educ,ability}$. Therefore, if we assume that $r_{educ,ability} > 0$ and $\beta_1 > 0$ then, there will be upward bias and $\hat{\beta}_1$ will tend to overestimate $\beta_1$.

---

**(b) How would you obtain a reliable estimate of the slope parameter $\beta_1$ using first a proxy variable and then an instrumental variable?**

**Proxy Variable:** Suppose $P$ is an ideal proxy in that there exists a linear relationship between *ability* and $P$ such that

$$ability = \delta_0 + \delta_1 P + v$$

We can rewrite our model using this relationship

$$\begin{aligned} log(wage) &= \alpha + \beta_1 \ educ + \beta_2(\delta_0 + \delta_1 \ P + v) + \varepsilon \\ &= (\alpha + \beta_2\delta_0) + \beta_1 \ educ + \beta_2\delta_2 \ P + (\beta_2 v + \varepsilon) \\ &= \gamma_0 + \beta_1 \ educ + u \end{aligned}$$

The composite error $u$ depends on both the error in the model, $\varepsilon$, and the error in the proxy equation, $v$.

The model is now formally specified correctly in terms of observable variables. If we fit this model we will obtain the following results:

- coefficient of *educ*, i.e. $\beta_1$, its standard error, and its $t$ statistic will be the same as if *ability* has been used instead of $P$;
- $R^2$ will be the same as if *ability* has been used instead of $P$;
- coefficient of $P$ will be an estimate of $\beta_2\delta_2$ which means we cannot obtain an estimate of $\beta_2$, unless we are able to guess the value of $\delta_2$;
- the $t$ statistic for $P$ will be the same as that which would have been obtained for *ability*, so we can assess the significance of *ability*, even though we cannot estimate its coefficient;
- since intercept is now $\alpha + \beta_2\delta_0$ we cannot obtain an estimate of the intercept $\alpha$, though, here, intercept is not a primary interest.

For us to get a consistent estimator of $\beta_1$, the coefficient of *educ*, through the use of proxy variable method, the following two assumptions need to hold:

- The error $\varepsilon$ is uncorelated with *educ* and *ability* as well as $P$. That is, $\mathbb{E}(\varepsilon|educ, ability, P) = 0$. What this means is that $P$ is irrelevant in the population model and is not contained in the error term. It is *ability* that directly affects $log(wage)$ not $P$. $P$ is just a proxy for *ability*.
- The error $v$ is uncorrelated with *educ* and $P$. If $P$ is a good proxy for *ability*, then $v$ is uncorrelated with *educ*. Here, the term 'good' or 'ideal' means that $\mathbb{E}(ability|educ, P) = \mathbb{E}(ability|P) = \delta_0 + \delta_1\ P$. That is, once $P$ is controlled for, the expected value of *ability* does not depend on *educ*. In other words, *ability* has zero correlation with *educ* once $P$ is partialled out. Thus the average level of *ability* only changes with $P$ and not with *educ*.

**Instrumental Variable:** Instrumental variables are especially important when we want to fit models comprising several simultaneous equations. Suppose for this question proxy variable does not have the required properties for a consistent estimate of $\beta_1$. Then we put *ability* in the error term since it is unobserved. This leaves us with:

$$log(wage) = \beta_0 + \beta_1\ educ + \epsilon$$

where $\epsilon$ contains *ability*. If *ability* and *educ* are correlated, then we have a biased and inconsistent estimate of $\beta_1$. However, we can still use this equation as a basis for estimation as long as we can find an instrumental variable $Z$ for *educ*. In order for $Z$ to be used as an instrumental variable, it needs to satisfy the following conditions:

Instrument Relevance: $Z$ is correlated with *educ*, i.e. $Cov(Z, educ) \neq 0$; and

Instrument Exogeneity: $Z$ is uncorrelated with $\epsilon$, i.e. $Cov(Z, \epsilon) = 0$.

Then the estimator of the coefficient for *educ* becomes:

$$\begin{aligned}
\hat{\beta}_1^{IV} &= \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})\Big(log(wage)_i - \overline{log(wage)}\Big)}{\sum_{i=1}^{n}(Z_i - \bar{Z})(educ_i - \overline{educ})} \\
&= \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})\Big((\beta_0 + \beta_1\ educ_i + \epsilon_i) - (\beta_0 + \beta_1\ \overline{educ} + \bar{\epsilon})\Big)}{\sum_{i=1}^{n}(Z_i - \bar{Z})\Big(educ_i - \overline{educ}\Big)} \\
&= \frac{\sum_{i=1}^{n}\Big(\beta_1(Z_i - \bar{Z})(educ_i - \overline{educ}) + (Z_i - \bar{Z})(\epsilon_i - \bar{\epsilon})\Big)}{\sum_{i=1}^{n}(Z_i - \bar{Z})\Big(educ_i - \overline{educ}\Big)}
\end{aligned}$$

$$= \beta_1 + \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(\epsilon_i - \bar{\epsilon}\Big)}{\sum_{i=1}^{n}(Z_i - \bar{Z})\Big(educ_i - \overline{educ}\Big)}$$

Thus the $IV$ estimator is equal to the true value plus an error term. We can't however obtain its expectation because we cannot obtain an expectation for the error term since $educ$ is not distributed independently of $\epsilon$.

As a second best measure, we can investigate whether we can say anything about the error term in large samples by looking at its probability limit:

$$plim\Big(\hat{\beta}_1^{IV}\Big) = \beta_1 + plim\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(\epsilon_i - \bar{\epsilon}\Big)}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})\Big(educ_i - \overline{educ}\Big)}\right)$$

$$= \beta_1 + \frac{Cov(Z, \epsilon)}{Cov(Z, educ)}$$

$$= \beta_1 + \frac{0}{\sigma_{Z,educ}} \qquad \text{since } Cov(Z, \epsilon) = 0$$

$$= \beta_1$$

That is, insofar as $\sigma_{Z,educ} \neq 0$, $\hat{\beta}_1^{IV}$ will tend to the true value of $\beta_1$ in large samples.

---

**(c) Given your answer to (b), evaluate the following statement: "whilst $IQ$ is a good candidate for a proxy variable of *ability*, it cannot be used as an instrument for *education*."**

**Answer:** Even though instrumental variable is a useful method, we cannot test for "instrument exogeneity" assumption. We can only consider economic behavior in order to maintain the $Cov(Z, educ) \neq 0$ assumption. At times there may be an observable proxy for some factor contained in $\epsilon$ and we can check if $Z$ and the proxy variable are more or less uncorrelated. On the other hand, if we have a good proxy, then, we would add that variable to the equation and estimate the expanded form using OLS.

This is exactly where we see a tension between a good proxy vs. a good IV:

good proxy: For $IQ$ to be a good proxy, it needs to be as highly correlated with *ability* as possible;

good IV: For $IQ$ to be a good instrumental variable, it needs to be uncorrelated with *ability* since *ability* is contained in $\epsilon$ and a good IV should not covary with the error term, hence the "instrument exogeneity" condition. That is, a good IV should affect $log(wage)$ only through its influence on *educ* and not in any other way.

Therefore, in this question, it is correct to say that $IQ$ is a good candidate for a proxy variable of *ability*, it is not a good instrumental variable for *educ*.

## QUESTION 3

**Consider the following PRF where $Cov(X_i, u_i) \neq 0$:**

$$Y_i = \alpha + \beta X_i + u_i \tag{2}$$

**Assume that there is an instrument (some variable $Z_i$) that satisfies the assumptions $Cov(Z_i, u_i) = 0$ and $Cov(Z_i, X_i) \neq 0$. By deriving the expression for $Cov(Z_i, Y_i)$, show that the IV estimator for $\beta$ using $Z_i$, is given by the following:**

$$\hat{\beta}_{IV} = \frac{\sum(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum(Z_i - \bar{Z})(X_i - \bar{X})} \tag{3}$$

**Answer:** We have derived this in Question 1(b) above using *educ*. We will do this again here with a slightly different approach. The question is asking us to derive the expression for $Cov(Z_i, Y_i)$, which is

$$Cov(Z_i, Y_i) = Cov(Z_i\ ,\ \alpha + \beta X_i + u_i) = \beta_1 Cov(Z_i, X_i) + Cov(Z_i, u_i)$$

Assuming both instrument relevance, $Cov(Z_i, X_i) \neq 0$ and instrument exogeneity, $Cov(Z_i, u_i) = 0$, assumptions hold true, we can solve this for $\beta_1$ as

$$\beta_1 = \frac{Cov(Z_i, Y_i}{Cov(Z_i, X_i)}.$$

Therefore, $\beta_1$ is the ratio of population covariance between $Z$ and $Y$ to the population covariance betweeen $Z$ and $X$, which shows that $\beta_1$ is identified. Here, *identification* of parameter means that we can write $\beta_1$ in terms of population moments that can be estimated using a sample data.

Given a random sample, we estimate the population quantities by the sample analogs. After canceling the sample sizes in the numerator and the denominator, we get the IV estimator of $\beta_1$:

$$\beta_1 = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})}$$

as desired.

Also note that the IV estimator of $\beta_0$ is $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, where the slope estimator, $\hat{\beta}_1$ is the IV estimator.

## QUESTION 4

**(a) Using the 'Phillips' data in the IV dataset, estimate the following expectations augmented Phillips Curve:**

$$\Delta inf_t = \beta_0 + \beta_1 \; unem_t + e_t \tag{4}$$

**Obtain an estimate of the natural rate of unemployment.**

**Answer:** In STATA we can do this using the following code

```
/* load the data */
quietly cd ..
 import excel using Data/iv.xls, sheet("Phillips") firstrow
/* `firstrow` indicates that the first row contains the variable names */

/* set the time variable */
  tsset year

/* run the regression */
  regress D.inf unem
```

```
(3 vars, 49 obs)


Time variable: year, 1948 to 1996
        Delta: 1 unit

      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(1, 46)        =      5.56
       Model |  33.3829986         1  33.3829986   Prob > F        =    0.0227
    Residual |  276.305126        46  6.00663318   R-squared       =    0.1078
-------------+----------------------------------   Adj R-squared   =    0.0884
       Total |  309.688125        47  6.58910904   Root MSE        =    2.4508

------------------------------------------------------------------------------
       D.inf | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        unem |  -.5425869   .2301559    -2.36   0.023    -1.005867    -.079307
       _cons |   3.030581    1.37681     2.20   0.033      .259206    5.801955
------------------------------------------------------------------------------
```

We see that the $t$ statistic for both the intercept and the slope coefficient are significant.

To get the natural rate, we can set the change in inflation, $\Delta inf_t$ equal to zero and then rearrange for unemployment to obtain the natural rate. That is,

$$\begin{aligned} \Delta inf_t &= 3.030581 + -0.54255769 \; unem_t \\ 0 &= 3.030581 + -0.54255769 \; unem_t \\ \frac{-3.030581}{-0.54255769} &= unem_t \\ 5.585 &= unem_t \end{aligned}$$

Therefore we estimate that the natural rate of unemployment is about 5.585.

And in R we can use the following code to obtain the same results:

```
# Load the data
phillips_df <- read_excel("../Data/iv.xls", sheet = "Phillips")

# create the Delta variable of first differences
phillips_df <- phillips_df %>%
  mutate(delta_inf = inf - lag(inf))

# Run the regression
SQ4a_lm <- lm(delta_inf ~ unem, data = phillips_df)
summary(SQ4a_lm)

# Obtain the natural rate
-SQ4a_lm$coefficients[1]/SQ5a_lm$coefficients[2]
```

---

**(b) It is suspected that $unem_t$ is related to $e_t$. Why might this be and what implications follow if this is correct? Explain carefully why this problem might be alleviated by using $unem_{t-1}$ as an instrument to construct an instrumental variable (IV) for $unem_t$? Test your assumptions where possible.**

**Answer:** If there are supply-side shocks that occur every now and again and influence both price expectations and unemployment, then unemployment is not exogeneous. This endogeneity means we will get biased estimates. Notice that since these shocks are random and correctly put in the error term, they are not the same as omitted variable bias, but it nevertheless causes the same problems. As such, it can be helped by the use of IV estimation.

The second part of the question asks why lagged unemployment can be used as an IV for unemployment. For this, recall that IV has two criteria - instrument relevance and instrument exogeneity. The former requires the IV to be related to *unem*, while the latter requires that it should not be related to $e$. Since the shocks are included in this error term, it also means that the IV should not be related to the shocks. Given that by definition 'shock' means it is unpredictable before it occurs, then using unemployment rate the year before a shock happens means it should not be related to the shock due it latter's unpredictability.

We cannot test the instrument exogeneity assumption but we can test the instrument relevance assumption. For this, we can run a regression of *unem* on its lagged values and check if the lagged variable is significant. In R,

```
SQ4b_lm <- lm(unem ~ lag(unem,1), data = phillips_df)
summary(SQ4b_lm)
```

and in STATA:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("Phillips") firstrow
quietly tsset year

regress unem L.unem
```

```
      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(1, 46)        =     57.13
       Model |  62.8162744         1  62.8162744   Prob > F        =    0.0000
    Residual |  50.5768506        46  1.09949675   R-squared       =    0.5540
-------------+----------------------------------   Adj R-squared   =    0.5443
       Total |  113.393125        47  2.41261968   Root MSE        =    1.0486

------------------------------------------------------------------------------
        unem | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        unem |
         L1. |   .7323538   .0968906     7.56   0.000     .5373231    .9273845
             |
       _cons |   1.571741   .5771181     2.72   0.009     .4100628     2.73342
------------------------------------------------------------------------------
```

From the output we see that the $t$ statistic for the lagged unemployment variable is significant at 7.56, which means the instrument relevance, i.e. identification, assumption seems to hold.

---

**(c) Estimate the equation (4) on page 14 by IV. Compare these results to those obtained using the 2SLS option in Stata. Compare your results to those obtained in part (a).**

**Answer:** There are multiple ways of doing this. For illustration, we will do the 2-stage least squares regression manually and then use specific STATA commands for IV regression.

For manual calculation, notice that we have already completed the first stage in part (b) above. We will now put the fitted values from the reduced form regression into a new variable called $unemf$ using the 'predict' command. For the second stage, we will then estimate the first equation again, but this time using these fitted values $unemf$ instead of $unem$.

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("Phillips") firstrow
quietly tsset year

quietly regress unem L.unem
quietly predict unemf

regress D.inf unemf
```

```
      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(1, 46)        =      0.18
       Model |   1.2021291         1   1.2021291   Prob > F        =    0.6740
    Residual |  308.485996        46   6.7062173   R-squared       =    0.0039
-------------+----------------------------------   Adj R-squared   =   -0.0178
       Total |  309.688125        47  6.58910904   Root MSE        =    2.5896
```

```
------------------------------------------------------------------------------
       D.inf | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       unemf |  -.1383374   .3267403    -0.42   0.674    -.7960315    .5193568
       _cons |   .6935128   1.925594     0.36   0.720    -3.182506    4.569532
------------------------------------------------------------------------------
```

The coefficient for $unemf$ is $-0.1383374$ compared to $-0.5425869$ for $unem$. Similarly the intercept changed to 0.6935128 from 3.030581. This will also change the natural rate to

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("Phillips") firstrow
quietly tsset year
quietly regress unem L.unem
quietly predict unemf
quietly regress D.inf unemf

display -_b[_cons]/_b[unemf]
```

```
5.0132
```

To obtain the same results using STATA commands instead of doing it manually, we can either use `ivreg` command or `ivregress 2sls` command. Note that `ivreg` is not short for `ivregress 2sls` - they are different commands. However, if we add the `,small` option at the end of `ivregress 2sls` command, then it will give the same result as `ivreg`.

The use of `ivreg` is limited in that we cannot make use of the post estimation options. This is not important for this question but it will be for the next one when we need to do overidentification test using the `estat overid` and endogeneity test using the `estat endog` command. These commands only work if we use `ivregress 2sls` and not `ivreg`.

To estimate the model with the IV method we reference the variable that we suspect is endogeneous, i.e. *unem* within a parenthesis and set it equal to the instrument or instruments we are using, i.e. lagged values of *unem*. If you want to obtain the reduced form equation, use the `first` option at the end of the IV regression:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("Phillips") firstrow
quietly tsset year

ivregress 2sls D.inf (unem = L.unem), small first
```

```
First-stage regressions
-----------------------

                                                Number of obs =     48
                                                F(1, 46)      =  57.13
                                                Prob > F      = 0.0000
                                                R-squared     = 0.5540
                                                Adj R-squared = 0.5443
                                                Root MSE      = 1.0486

------------------------------------------------------------------------------
        unem | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
```

```
-------------+----------------------------------------------------------------
        unem |
         L1. |   .7323538   .0968906     7.56   0.000     .5373231    .9273845
             |
       _cons |   1.571741   .5771181     2.72   0.009     .4100628     2.73342
------------------------------------------------------------------------------
```

```
Instrumental-variables 2SLS regression

      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(  1,    46)   =      0.19
       Model |  14.8525524         1  14.8525524   Prob > F        =    0.6670
    Residual |  294.835573        46  6.40946897   R-squared       =    0.0480
-------------+----------------------------------   Adj R-squared   =    0.0273
       Total |  309.688125        47  6.58910904   Root MSE        =    2.5317

------------------------------------------------------------------------------
       D.inf | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        unem |  -.1383373   .3194294    -0.43   0.667    -.7813154    .5046408
       _cons |   .6935127   1.882508     0.37   0.714     -3.09578    4.482805
------------------------------------------------------------------------------
Endogenous: unem
Exogenous:  L.unem
```

Notice that while the coefficients are the same as the ones we obtained manually, the standard errors and $t$ statistics are slightly different. Using STATA's commands give more correct information so use the STATA commands when possible.

In R we can use the `ivreg()` function from the `library` package. To fit the model with this function we extend the original regression formula by adding a second part after the | separator to specify the instrumental variables. If there are multiple variables, then we use three parts using the | separator. The first part is the exogeneous variables, the second part is the endogeneous variables, and the third part is the instrumental variables. Since we only have one endogeneous variable, we use one | separator.

```
SQ4c_lm <- ivreg(delta_inf ~ unem | lag(unem,1), data = phillips_df)
summary(SQ4c_lm)
```

---

**(d) Estimate the equation (4) on page 14 one more time, but this time add $unem_{t-1}$ as a second regressor. What implications follow from the results above?**

**Answer:** If we add $unem_{t-1}$ as a second regressor, we get the following

with R:

```
SQ4d_lm <- update(SQ4a_lm, ~ . + lag(unem,1))
summary(SQ4d_lm)
```

and with STATA:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("Phillips") firstrow
quietly tsset year

regress D.inf unem L.unem
```

```
      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(2, 45)        =      5.01
       Model |  56.3977489         2  28.1988745   Prob > F        =    0.0109
    Residual |  253.290376        45  5.62867502   R-squared       =    0.1821
-------------+----------------------------------   Adj R-squared   =    0.1458
       Total |  309.688125        47  6.58910904   Root MSE        =    2.3725

------------------------------------------------------------------------------
       D.inf | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        unem |
         --. |  -1.044663   .3336009    -3.13   0.003     -1.71657   -.3727568
         L1. |   .6637514   .3282505     2.02   0.049     .0026209    1.324882
             |
       _cons |   2.118023   1.407123     1.51   0.139    -.7160676    4.952114
------------------------------------------------------------------------------
```

Notice that both *unem* and its lagged variable are significant suggesting that lagged unemployment seems to be a regressor it its own right, and so has a direct impact on the "change in inflation".

Since this seems to be the case, the results in parts (b) and (c) are likely to be misleading. Also, this means we cannot use $unem_{t-1}$ as an instrument because if it is a regressor in its own right, then previously it would have been in the error term, and therefore $Cov(unem_{t-1}) \neq 0$ which would have violated the instrument exogeneity assumtion.

## QUESTION 5

**(a) Using the 'regional' data from iv.xls, estimate the following equation using OLS**

$$log(wage) = \alpha + \beta_1 \ educ + \varepsilon \tag{5}$$

**Answer:**

In R:

```r
regional_df <- read_excel("../Data/iv.xls", sheet = "regional")
SQ5a_lm <- lm(lwage ~ educ, data = regional_df)
summary(SQ5a_lm)
```

and in STATA:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow
regress lwage educ
```

```
      Source |       SS           df       MS      Number of obs   =     2,220
-------------+----------------------------------   F(1, 2218)      =    183.37
       Model |  32.7582544         1  32.7582544   Prob > F        =    0.0000
    Residual |  396.241249     2,218  .178647993   R-squared       =    0.0764
-------------+----------------------------------   Adj R-squared   =    0.0759
       Total |  428.999504     2,219  .193330105   Root MSE        =    .42267

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .0469534   .0034674    13.54   0.000     .0401536    .0537531
       _cons |   5.645567   .0480961   117.38   0.000     5.551249    5.739885
------------------------------------------------------------------------------
```

---

**(b) Now estimate equation (5) on page 19 again using $fatheduc$ as an instrument. Do this by (i) running a reduced form equation of *educ* against $fatheduc$ and substituting the fitted values into equation (5); (ii) by using the IV formula derived in Question 3 above; (iii) by using ivreg command in STATA. Verify that the estimates of $\beta_1$ are the same in each case. Are these results what you expected (i.e. is the change in the estimate of $\beta_1$ roughly what you expected)?**

**Answer (i):** This part of the question is asking us to manually conduct the 2-stage least squares estimation using the IV method.

In STATA:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow

quietly regress educ fatheduc

/* put the fitted values into a new variable educf */
quietly predict educf

/* run the regression again with educf */
regress lwage educf
```

```
      Source |       SS           df       MS      Number of obs   =     2,220
-------------+----------------------------------   F(1, 2218)      =     81.89
       Model |  15.2756231         1  15.2756231   Prob > F        =    0.0000
    Residual |  413.723881     2,218  .186530154   R-squared       =    0.0356
-------------+----------------------------------   Adj R-squared   =    0.0352
       Total |  428.999504     2,219  .193330105   Root MSE        =    .43189

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       educf |   .0683401   .0075518     9.05   0.000     .0535307    .0831494
       _cons |    5.35412   .1033194    51.82   0.000     5.151508    5.556733
------------------------------------------------------------------------------
```

**Answer (ii):** Next we use the formula from Question 3 to derive the coefficient for $educf$. The formula we use for this is

$$\hat{\beta}^{IV} = \frac{\sum(fatheduc_i - \overline{fatheduc})(lwage_i - \overline{lwage})}{\sum(fatheduc_i - \overline{fatheduc})(educ_i - \overline{educ})}$$

```
sum((regional_df$fatheduc - mean(regional_df$fatheduc))
    *(regional_df$lwage - mean(regional_df$lwage))) /
  sum((regional_df$fatheduc - mean(regional_df$fatheduc))
    *(regional_df$educ - mean(regional_df$educ)))
```

```
Error in eval(expr, envir, enclos): object 'regional_df' not found
```

which gives us the same coefficient $\hat{\beta}_1 = 0.06834005$.

**Answer (iii):** Next we use the commands that do this automatically.

In STATA we can again use `ivreg` or `ivregress 2sls ,small` command but the question is asking specifically for us to use `ivreg`:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow

ivreg lwage (educ=fatheduc)
```

```
Instrumental variables 2SLS regression

      Source |       SS           df       MS      Number of obs   =     2,220
-------------+----------------------------------   F(1, 2218)      =     84.06
       Model |  25.9619248         1  25.9619248   Prob > F        =    0.0000
    Residual |  403.037579     2,218  .181712164   R-squared       =    0.0605
-------------+----------------------------------   Adj R-squared   =    0.0601
       Total |  428.999504     2,219  .193330105   Root MSE        =    .42628

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .0683401   .0074536     9.17   0.000     .0537232    .0829569
```

```
       _cons |    5.35412   .1019763    52.50   0.000     5.154141    5.554099
------------------------------------------------------------------------------
Endogenous: educ
Exogenous:  fatheduc
```

which gives us the same coefficients with slightly different $t$ statistics, though giving significant coefficients in either approach. We would accept the results from this STATA command as more correct, though.

Notice that the result would not be what we would expect if *ability* is the variable omitted since *ability* should be positively correlated with *educ* and therefore causing and upward bias.

In R we can obtain the same results via:

```r
SQ5b_lm <- ivreg(lwage ~ educ | fatheduc, data = regional_df)
summary(SQ5b_lm)
```

---

**(c) It is suggested that equation (5) is problematic because it ignores experience and that the following specification is likely to give better results:**

$$log(wage) = \alpha + \beta_1\ educ + \beta_2\ exper + \beta_3\ exper^2 + \varepsilon \tag{6}$$

**What is the reasoning behind this new specification? Estimate equation (6) using both OLS and IV methods of estimation (using the same instrument as in part (b)). Discuss your results.**

**Answer:** We will start with OLS and then run the regression with IV method.

OLS in R:

```r
SQ5c_lm <- lm(lwage ~ educ + exper + I(exper^2), data = regional_df)
# or lm(lwage ~ educ + poly(exper, 2, raw=T), data = regional_df)
summary(SQ5c_lm)
```

and OLS in STATA:

```stata
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow

regress lwage educ exper expersq
```

```
      Source |       SS           df       MS      Number of obs   =     2,220
-------------+----------------------------------   F(3, 2216)      =    171.48
       Model |  80.8281504         3  26.9427168   Prob > F        =    0.0000
    Residual |  348.171353     2,216  .157117037   R-squared       =    0.1884
-------------+----------------------------------   Adj R-squared   =    0.1873
       Total |  428.999504     2,219  .193330105   Root MSE        =    .39638
```

```
------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .0897809   .0041614    21.57   0.000     .0816202    .0979416
       exper |   .0932152   .0083042    11.23   0.000     .0769304    .1095001
     expersq |  -.0025751   .0004171    -6.17   0.000    -.003393    -.0017571
       _cons |    4.50583   .0796664    56.56   0.000     4.349602    4.662059
------------------------------------------------------------------------------
```

Next we look at the results of regression with IV method.

In R:

```
SQ5c_iv_lm <- ivreg(lwage ~ poly(exper,2, raw=T) | educ | fatheduc, data = regional_df)
summary(SQ5c_iv_lm)
```

and in STATA:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow

ivregress 2sls lwage exper expersq (educ = fatheduc), small
```

```
Instrumental-variables 2SLS regression

      Source |       SS        df       MS        Number of obs   =      2,220
-------------+-------------------------------     F(  3,  2216)   =      58.58
       Model |  51.6039647     3  17.2013216      Prob > F        =     0.0000
    Residual |  377.395539  2216  .170304846      R-squared       =     0.1203
-------------+-------------------------------     Adj R-squared   =     0.1191
       Total |  428.999504  2219  .193330105      Root MSE        =     .41268

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .1465357   .0128246    11.43   0.000     .1213863    .1716851
       exper |   .1179213   .0101172    11.66   0.000     .0980811    .1377614
     expersq |  -.0026499   .0004345    -6.10   0.000    -.003502    -.0017977
       _cons |   3.533836   .2227414    15.87   0.000     3.097033     3.97064
------------------------------------------------------------------------------
Endogenous: educ
Exogenous:  exper expersq fatheduc
```

Again an unexpected result from the IV, we would have expected it to fall. This suggests that the bias is in the opposite direction. One possibility is that it is being caused by experience which would be negatively related to wage. Taking this out of the error term should increase the estimate of $\beta_1$ as observed. This suggests either that there is more in the error term that is negatively related to wage, or that $fatheduc$ is not a very good instrument.

---

**(d) It is now suggested that as well as *fatheduc*, *motheduc*, and *nearc4* should be used as instruments for *educ*. Explain why these seem plausible instruments. Can you find any support for the suggestion?**

**Answer:** To check if instrument relevance condition is being met, we can regress the endogeneous variable *educ* on these variables and check if any of them are significant.

In R:

```
SQ5d_lm <- lm(educ ~ fatheduc + motheduc + nearc4 + exper + I(exper^2), data = regional_df)
summary(SQ5d_lm)
```

and in STATA:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow

regress educ fatheduc motheduc nearc4 exper expersq
```

```
      Source |       SS           df       MS      Number of obs   =     2,220
-------------+----------------------------------   F(5, 2214)      =    405.40
       Model |  7101.83502         5    1420.367   Prob > F        =    0.0000
    Residual |  7757.08885     2,214   3.5036535   R-squared       =    0.4780
-------------+----------------------------------   Adj R-squared   =    0.4768
       Total |  14858.9239     2,219  6.69622527   Root MSE        =    1.8718

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    fatheduc |   .1245994   .0142697     8.73   0.000     .0966159    .1525829
    motheduc |   .1386471   .0169403     8.18   0.000     .1054265    .1718676
      nearc4 |   .3221091   .0866504     3.72   0.000     .1521845    .4920337
       exper |  -.3773007   .0384127    -9.82   0.000    -.4526294   -.3019719
     expersq |   .0023973   .0019706     1.22   0.224     -.001467    .0062617
       _cons |   13.60584    .249862    54.45   0.000     13.11585    14.09582
------------------------------------------------------------------------------
```

it appears all three additional variables are significant so the identification condition is being met for these as well.

---

**(e) Estimate equation (6) on page 22 using all three instruments. Discuss your results.**

**Answer:** We are now treating experience as exogenous and regress with the three instruments.

In R:

```r
SQ5e_lm <- ivreg(lwage ~ exper + expersq | educ | fatheduc + motheduc + nearc4,
                 data = regional_df)
summary(SQ5e_lm)
```

and in STATA:

```stata
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow

ivregress 2sls lwage exper expersq (educ = fatheduc motheduc nearc4), small
```

```
Instrumental-variables 2SLS regression

      Source |       SS           df       MS      Number of obs   =     2,220
-------------+----------------------------------   F(  3,  2216)   =     75.67
       Model |  40.9137684         3  13.6379228   Prob > F        =    0.0000
    Residual |  388.085735     2,216  .175128942   R-squared       =    0.0954
-------------+----------------------------------   Adj R-squared   =    0.0941
       Total |  428.999504     2,219  .193330105   Root MSE        =    .41848

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .1561088   .0115375    13.53   0.000     .1334833    .1787343
       exper |   .1220886   .0099213    12.31   0.000     .1026325    .1415446
     expersq |  -.0026625   .0004406    -6.04   0.000    -.0035265   -.0017985
       _cons |   3.369886   .2011369    16.75   0.000      2.97545    3.764323
------------------------------------------------------------------------------
Endogenous: educ
Exogenous:  exper expersq fatheduc motheduc nearc4
```

We see that the estimate of $\beta_1$ increased again which is a surprising result as we would have expected the error term to contain *ability* which would be positively related to *educ*.

---

**(f) Use the Over-Identifying Restrictions Test to see if either $fatheduc$ or $motheduc$ and $nearc4$ might be endogeneous.**

**Answer:** Overidentification refers to having more than one potential instruments for an endogeneous variable. For this we will conduct a test that is similar to Breusch-Godfrey test though here we do not have an autoregressive error term. Here we have the model

$$log(wage) = \alpha + \beta_1\ educ + \beta_2\ exper + \beta_3\ exper^2 + u$$

where we use instrumental variables $fatheduc, motheduc$, and $nearc4$ for the endogeneous variable $educ$. For them to be a good IV they also need to be uncorrelated with the error term. So the coefficients in

$$u = \gamma_1 fatheduc + \gamma_2 motheduc + \gamma_3 nearc4 + v$$

where $v$ is the white noise term. The null hypothesis is that these variables satisfy instrumental exogeneity assumption, i.e. the coefficients are jointly zero:

$$\mathbb{H}_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0.$$

The $R^2$ from this regression multiplied by the sample size asymptotically follows a $\chi^2_3$ distribution.

In R:

```
SQ5f_iv_lm <-
  ivreg(lwage ~ poly(exper,2, raw=T) | educ | fatheduc + motheduc + nearc4,
        data = regional_df)
SQ5f_res_lm <-
  lm(SQ5f_iv_lm$residuals ~ fatheduc + motheduc + nearc4 + exper + I(exper^2),
     data = regional_df)
summary(SQ5f_res_lm)

# calculate n times R-squared
length(regional_df$lwage) * summary(SQ5f_res_lm)$r.squared
```

In STATA:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow

quietly ivreg 2sls lwage (educ = fatheduc motheduc nearc4) exper expersq
predict U, r
reg U fatheduc motheduc nearc4 exper expersq

/* calculate n times R-squared */
display e(r2)*e(N)
```

```
2sls invalid name
r(198);

r(198);
```

We can also obtain this same $\chi^2$ statistic using `estat overid` command:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow
quietly ivregress 2sls lwage (educ = fatheduc motheduc nearc4) exper expersq, small

estat overid
```

```
  Tests of overidentifying restrictions:

  Sargan (score) chi2(2) =  8.53794  (p = 0.0140)
  Basmann chi2(2)        =  8.54774  (p = 0.0139)
```

The Sargan score gives us the $chi^2$ statistic we are interested in, which is the same as the one we obtained manually.

At $\alpha = 0.05$ the $\chi_2$ with 3 degrees of freedom is

```
qchisq(p = 0.05, df=3, lower.tail=FALSE)
```

```
[1] 7.814728
```

Since our statistic of 8.5379406 exceeds this critical value of 7.814728 we reject the null hypothesis. This means at least one of the variables is correlated with the error term and thus not a good candidate for being an instrumental variable. To find out which of these candidates are not exogeneous, we can run the same auxiliary error regression on two of the three candidates at a time.

Lets start by keeping $fatheduc$ and $motheduc$ and removing $nearc4$:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow
quietly ivregress 2sls lwage (educ = fatheduc motheduc) exper expersq, small
estat overid
```

```
  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  .319686  (p = 0.5718)
  Basmann chi2(1)        =  .319012  (p = 0.5722)
```

We get a $\chi^2$ statistic of 0.319686 which is below the critical value of 7.814728, thus we fail to reject the null hypothesis. Therefore both of these seem to be good candidates for being an instrument.

Lets now try the same by keeping $fatheduc$ and $near4c$ and removing $motheduc$:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow
quietly ivregress 2sls lwage (educ = fatheduc nearc4) exper expersq, small
estat overid
```

```
  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  8.46808  (p = 0.0036)
  Basmann chi2(1)        =  8.48136  (p = 0.0036)
```

We get a $\chi^2$ statistic of 8.46808 which exceeds the critical value of 7.814728, thus we reject the null hypothesis. This makes $nearc4$ a suspect, i.e. it appears $nearc4$ is endogeneous.

**(g) Now regress $IQ$ on $motheduc$, $fatheduc$, and $nearc4$. What do these results appear to suggest about your choice of instruments?**

**Answer:** What we are doing in this question is to check if these instruments are related to $IQ$ and thus to *ability*.

Before we run the commands though, note that IQ is coded as "strings" in STATA and as "character" in R. We need to convert it to numerical data first. For this we use the 'real' command in STATA, while we use the 'transform()' function in R in combination with 'as.numeric()' and 'as.character()' functions. This is because the latter first converts the column into actual "character" structure, and then we convert it to numeric data.

```
regional_df$IQ
regional_df <- regional_df %>%
  transform(IQ = as.numeric(as.character(IQ)))
summary(lm(IQ ~ fatheduc + motheduc + nearc4 + exper + expersq, data=regional_df))
```

and in STATA:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow

gen iq = real(IQ)
reg iq fatheduc motheduc nearc4 exper expersq
```

```
(601 missing values generated)

      Source |       SS           df       MS      Number of obs   =     1,619
-------------+----------------------------------   F(5, 1613)      =     77.69
       Model |  70649.5496         5  14129.9099   Prob > F        =    0.0000
    Residual |  293372.009     1,613  181.879733   R-squared       =    0.1941
-------------+----------------------------------   Adj R-squared   =    0.1916
       Total |  364021.559     1,618  224.982422   Root MSE        =    13.486

------------------------------------------------------------------------------
          iq | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    fatheduc |    .745289   .1240387     6.01   0.000     .5019951    .9885829
    motheduc |   .6739114   .1462515     4.61   0.000     .3870484    .9607744
      nearc4 |   1.476281   .7401082     1.99   0.046     .0246061    2.927956
       exper |  -2.215204   .3784058    -5.85   0.000    -2.957423   -1.472985
     expersq |   .0633378    .020892     3.03   0.002     .0223594    .1043161
       _cons |   100.4036   2.322438    43.23   0.000     95.84833     104.959
------------------------------------------------------------------------------
```

From the $t$ statistics, it looks like all instruments are related to $IQ$ including $nearc4$, and so by implication to *ability*. This would seem to explain our results in that the instruments are not passing the exogeneity condition since *ability* is left in the error term.

It is also a bit odd that the main offender above, $nearc4$ is least related to IQ with a $t$ statistic at the cusp of being rejected at $\alpha = 0.05$.

**(h) Repeat the regression from (g) but now adding the regional dummies** ($reg661, \ldots, 669$)**. Explain your results and their implications for the use of fatheduc, motheduc, and nearc4 as instruments (N.B. that the regional dummies are exhaustive, so you don't need to include a constant term).**

**Answer:** In this question we are adding the regional dummies into the regression. Since the dummies are exhaustive we can either leave all of them in the equation and take the intercept out, or leave one of the regions out and keep the intercept. However, regional dummies will be significant only if we regress without the intercept. This is because if we leave the intercept, the other dummies are not significantly different from the constant term; i.e. they are all pretty much the same.

In R, it is probably easiest to do this by creating a new dataframe that is a subet of the original dataframe with only the relevant variables for this question and regress it that way. Also since, all the regions start with *reg* we can use a shortcut to get all the variables that begin with *reg* instead of typing them one by one.

In R regressing on 0 as the first variable removes the intercept:

```
regional_df_subset <- regional_df %>%
  select(matches(c("lwage","IQ","nearc4","educ","fatheduc","motheduc","^exp", "^reg")))

SQ5h_lm <- lm(IQ ~ 0 + . - lwage - educ, data=regional_df_subset)
summary(SQ5h_lm)
```

In STATA we remove the intercept via the `noco` or `noconstant` option :

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow
gen iq = real(IQ)

reg iq nearc4 fatheduc motheduc exp* reg**, noconstant
```

```
(601 missing values generated)

      Source |       SS           df       MS      Number of obs   =     1,619
-------------+----------------------------------   F(14, 1605)     =   7070.35
       Model |  17544502.9        14  1253178.78   Prob > F        =    0.0000
    Residual |    284477.1     1,605  177.244299   R-squared       =    0.9840
-------------+----------------------------------   Adj R-squared   =    0.9839
       Total |    17828980     1,619   11012.341   Root MSE        =    13.313

------------------------------------------------------------------------------
          iq | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      nearc4 |    .233572   .7660108     0.30   0.760    -1.268915    1.736059
    fatheduc |   .6996141    .123311     5.67   0.000     .4577466    .9414816
    motheduc |   .6154042   .1448345     4.25   0.000     .3313196    .8994889
       exper |    -2.2364   .3738661    -5.98   0.000    -2.969717   -1.503083
     expersq |   .0632576   .0206387     3.07   0.002      .022776    .1037392
      reg661 |   105.3192    2.87969    36.57   0.000     99.67085    110.9675
      reg662 |   105.4357   2.452782    42.99   0.000     100.6247    110.2467
      reg663 |   103.6808   2.427093    42.72   0.000     98.92018    108.4414
      reg664 |   104.3312   2.594447    40.21   0.000     99.24237    109.4201
```

```
      reg665 |   100.5692   2.398629    41.93   0.000     95.86441     105.274
      reg666 |   98.26051   2.573879    38.18   0.000     93.21199     103.309
      reg667 |   98.73799   2.462225    40.10   0.000     93.90848    103.5675
      reg668 |    99.9522   2.963687    33.73   0.000     94.13909    105.7653
      reg669 |   102.3201    2.56142    39.95   0.000     97.29602    107.3442
------------------------------------------------------------------------------
```

From the regression output we see that $nearc4$ has a $t$ statistic of 0.3 and is not significant. Therefore, when controlling for regional factors, $nearc4$ is not related to $IQ$. On the other hand both $fatheduc$ and $motheduc$ are significant, and thus seem related to $IQ$. So it appears that the best thing to do is to only use $nearc4$ and to ensure the regional dummies are controlling for regional factors in the structural equation.

---

**(i) Estimate equation (6) on page 22 once more, this time using the dummies from part (h) and only $nearc4$ as an instrument, also use the `first` option so that you can check that the identification condition holds. Discuss these results.**

**Answer:**

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow

ivreg lwage exper expersq reg* (educ=nearc4), noco first
```

```
First-stage regressions
-----------------------

      Source |       SS           df       MS      Number of obs   =     2,220
-------------+----------------------------------   F(12, 2208)     =   8765.12
       Model |  418348.898        12  34862.4082   Prob > F        =    0.0000
    Residual |  8782.10207     2,208   3.9774013   R-squared       =    0.9794
-------------+----------------------------------   Adj R-squared   =    0.9793
       Total |      427131     2,220  192.401351   Root MSE        =    1.9943

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       exper |  -.4380766   .0408229   -10.73   0.000    -.5181318   -.3580213
     expersq |   .0018092   .0021014     0.86   0.389    -.0023118    .0059302
      reg661 |   17.03128   .2832607    60.13   0.000     16.47579    17.58676
      reg662 |   16.97511   .2166233    78.36   0.000      16.5503    17.39992
      reg663 |   16.90308   .2088767    80.92   0.000     16.49347     17.3127
      reg664 |    17.1388    .244935    69.97   0.000     16.65847    17.61913
      reg665 |   16.55829   .2111376    78.42   0.000     16.14424    16.97234
      reg666 |   16.43495   .2330453    70.52   0.000     15.97794    16.89196
      reg667 |   16.55231   .2250317    73.56   0.000     16.11102    16.99361
```

```
     reg668 |   17.50338   .3033482    57.70   0.000      16.9085    18.09826
     reg669 |   17.31334   .2358749    73.40   0.000     16.85078     17.7759
     nearc4 |   .3631534   .0969822     3.74   0.000     .1729675    .5533394
------------------------------------------------------------------------------
```

```
Instrumental variables 2SLS regression

      Source |       SS           df       MS      Number of obs   =     2,220
-------------+----------------------------------   F(12, 2208)     =         .
       Model |  87668.7585        12  7305.72987   Prob > F        =         .
    Residual |  464.763217     2,208  .210490587   R-squared       =         .
-------------+----------------------------------   Adj R-squared   =         .
       Total |  88133.5217     2,220  39.6997845   Root MSE        =    .45879

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .2079933   .0614354     3.39   0.001     .0875161    .3284705
       exper |    .145927   .0284331     5.13   0.000     .0901687    .2016854
     expersq |  -.0027441   .0004935    -5.56   0.000    -.0037118   -.0017763
      reg661 |   2.425281   1.067259     2.27   0.023     .3323456    4.518217
      reg662 |   2.529025   1.064072     2.38   0.018     .4423386    4.615712
      reg663 |   2.562769   1.055728     2.43   0.015     .4924449    4.633094
      reg664 |   2.411707   1.067903     2.26   0.024     .3175083    4.505906
      reg665 |   2.410029   1.032755     2.33   0.020      .384757    4.435301
      reg666 |   2.426581   1.018729     2.38   0.017     .4288147    4.424347
      reg667 |   2.431834   1.031192     2.36   0.018      .409627    4.454041
      reg668 |    2.30266   1.091624     2.11   0.035     .1619433    4.443377
      reg669 |   2.490722   1.083201     2.30   0.022     .3665226    4.614922
------------------------------------------------------------------------------
Endogenous:  educ
Exogenous:   exper expersq reg661 reg662 reg663 reg664 reg665 reg666 reg667
             reg668 reg669 nearc4
```

All the variables are significant. Interestingly, the coefficient on *educ* has now increased to 0.208 which is still not what we would expect. What is causing the problem is not the omission of *ability* because if it was then our estimates would be falling.

---

**(j) Using the residuals from the equation estimated in part (i), test the hypothesis that *educ* is endogeneous. Confirm your results using the STATA 'endogeneity test'. Discuss your results.**

**Answer:** This is essentially a test to see if any of this is required in the first place. We should test this because 2SLS estimator is less efficient with larger standard errors than OLS when the explanatory variables are exogenous.

### Testing for Endogeneity[a]

[a]Wooldridge (2021, $7^{th}$ ed) Section 15.5a: Testing for Endogeneity

Suppose there is a single suspected endogeneous variable in

$$Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 Z_1 + \beta_3 Z_2 + u$$

which would be *educ* in this question where the model with regional dummies is

$$log(wage) = \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 reg661 + \beta_5 reg662 + \beta_6 reg663 + \beta_7 reg664 + \beta_8 reg665 + \beta_9 reg666 + \beta_{10} reg6$$

which is without an intercept because all the regional dummy variables are present in the model.
If $Y_2$, which is *educ* in our case, is uncorrelated with $u$ then we should estimate the model by OLS and not 2SLS.
To test this, Hausman (1978)[a] suggested directly comparing the OLS and 2SLS estimates and determining whether the differences are statistically significant. If all variables are exogenous, then both OLS and 2SLS are consistent. If $Y_2$, or in our case *educ*, is endogeneous while the other variables are exogeneous, then 2SLS and OLS must differ significantly.
A regression test, is therefore, the most straight forward way of checking if the difference between 2SLS and OLS are statistically significant. This approach is based on estimating the reduced form for $Y_2$

$$Y_2 = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \alpha_3 Z_3 + \alpha_4 Z_4 + v$$

Here $Z_3$ and $Z_4$ are two additional exogeneous variables that do not appear in the main model.
If $v$ is uncorrelated with $u$ then $Y_2$ is uncorrelated with $u_1$ as well since each $Z_j$ is uncorrelated with $u$.
In this question, the reduced form is

$$educ = \alpha_1 nearc4 + \alpha_2 exper + \alpha_3 exper^2 + \alpha_4 reg661 + \alpha_5 reg662 + \alpha_6 reg663 + \alpha_7 reg664 + \alpha_8 reg665 + \alpha_9 reg666 + \alpha_{10} reg66$$

If $v$ is uncorrelated with $u$, then *educ* is also uncorrelated with $u$, and thus exogeneous, since all the other regressors are exogeneous.
This is what we want to test.
This means, we want to test if $\eta_1$ in the following relationship is zero or not

$$u = \eta_1 v + e$$

where $e$ is uncorrelated with $v$ and has zero mean.
Then $u$ and $v$ are uncorrelated if and only if $\eta_1 = 0$. The most straightforward way of testing is to include $v$ as an additional regressor to our model and do a $t$ test. However, $v$ is not observed, so we need to estimate the reduced form and use the residuals $\hat{v}$ instead. Therefore we estimate by OLS the following,

$$Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 Z_1 + \beta_3 Z_2 + \eta_1 \hat{v} + u$$

which, in this question is

$$log(wage) = \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 reg661 + \beta_5 reg662 + \beta_6 reg663 + \beta_7 reg664 + \beta_8 reg665 + \beta_9 reg666 + \beta_{10} reg6$$

and test $\mathbb{H}_0 : \eta_1 = 0$ using a $t$-statistic. If we reject the null hypothesis then we conclude that $Y_2$, or in our case *educ* is endogeneous because $v$ and $u$ are correlated. If that is the case, then we should use 2SLS.
Also note that all the coefficients from this last regression, with the exception the coefficient for $v$, $\eta_1$, will always be identical to the 2SLS estimates, even though we use OLS to estimate them. This can be used as a check to see whether we have done a proper regression in testing for endogeneity.

This point also gives us a useful interpretation of 2SLS. When we add $\hat{v}$ as an explanatory variable and applying OLS, we clear up the endogeneity of $Y_2$. So when we start by estimating the structural equation, i.e. original model without $\hat{v}$, by OLS, we can quantify the importance of allowing $Y_2$ to be endogeneous by seeing how much $\hat{\beta}_1$ changes when $\hat{v}_2$ is added to the equation. Irrespective of the outcome of the statistical tests, we can see whether the change in $\hat{\beta}_1$ is expected and practically significant.
A caution, however. If we go ahead with 2SLS estimates in the end, the standard errors should not come from the model with $\hat{v}$ included, which are only valid under the null hypothesis that $\eta_1 = 0$, but from the built-in 2SLS routines instead.
Finally, we can test for endogeneity of multiple explanatory variables, where we test for joint significance of the residuals in the structural equation using an $F$-test.

[a]Hausman J A (1978) "Specification Tests in Econometrics", *Econometrica*, 46:1251:1271

We can do this test in R as follows:

```
SQ5j_rf_lm <- lm(educ ~ 0 + . - IQ - fatheduc - motheduc - lwage,
                 data = regional_df_subset)
SQ5j_lm <- lm(lwage ~ 0 + . + SQ5j_rf_lm$residuals - IQ - fatheduc - motheduc,
              data = regional_df_subset)
summary(SQ5j_lm)
```

and in STATA:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow

/* obtain the residuals v from reduced form */
quietly regress educ nearc4 exper expersq reg*, noconstant
predict v, r

/* run the model with v included */
reg lwage educ exper expersq reg* v, noconstant
```

```
      Source |       SS           df       MS      Number of obs   =     2,220
-------------+----------------------------------   F(13, 2207)     =  45118.60
       Model |  87803.1419        13  6754.08784   Prob > F        =    0.0000
    Residual |  330.379754     2,207  .149696309   R-squared       =    0.9963
-------------+----------------------------------   Adj R-squared   =    0.9962
       Total |  88133.5217     2,220  39.6997845   Root MSE        =    .38691

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .2079934   .0518093     4.01   0.000     .1063933    .3095935
       exper |    .145927    .023978     6.09   0.000     .0989052    .1929488
     expersq |  -.0027441   .0004161    -6.59   0.000    -.0035601    -.001928
      reg661 |   2.425281   .9000339     2.69   0.007     .6602787    4.190283
      reg662 |   2.529025   .8973466     2.82   0.005     .7692926    4.288757
      reg663 |   2.562769   .8903102     2.88   0.004     .8168352    4.308702
      reg664 |   2.411707   .9005772     2.68   0.007     .6456394    4.177774
      reg665 |   2.410028   .8709362     2.77   0.006     .7020881    4.117969
      reg666 |    2.42658   .8591078     2.82   0.005     .7418361    4.111325
```

```
      reg667 |   2.431833   .8696181     2.80   0.005     .7264779    4.137189
      reg668 |   2.302659   .9205814     2.50   0.012      .497363    4.107956
      reg669 |   2.490722   .9134785     2.73   0.006     .6993544    4.282089
           v |  -.1237012   .0519736    -2.38   0.017    -.2256234    -.021779
------------------------------------------------------------------------------
```

The $t$ statistic for $v$ is $-2.38$ with a $p$ value of 0.017 which means we can reject the null hypothesis and conclude that *educ* is endogeneous. Accordingly we should estimate our model using 2SLS. Also notice that the relationship to the error term $v$ is negative.

We can also use the `estat endog` function in STATA for this:

```
quietly cd ..
quietly import excel using Data/iv.xls, sheet("regional") firstrow

quietly ivregress 2sls lwage (educ=nearc4) exper expersq reg*, small noconstant
estat endog
```

```
  Tests of endogeneity
  H0: Variables are exogenous

  Durbin (score) chi2(1)          =  5.68355  (p = 0.0171)
  Wu-Hausman F(1,2207)            =  5.66477  (p = 0.0174)
```

Which gives us an $F$-statistic of 5.66477 which is the square of the $t$-statistic of $-2.38$ we observed above.