ECONOMETRIC MODELS II - CAT I

May 3, 2025

QUESTION ONE

Endogeneity vs. Exogeneity in Econometrics:

(a) **Endogeneity** (2 Marks)

This is a situation where an explanatory variable is correlated with the error term in a regression model, which may be due to omitted variable bias, measurement error, or simultaneity. It leads to biased and inconsistent parameter estimates. For instance, in the model;

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

If X is endogenous, then;

$$Cov(X, \epsilon) \neq 0 \tag{2}$$

(b) **Exogeneity** (2 Marks)

This means the explanatory variables are uncorrelated with the error term. It ensures OLS estimators are unbiased and consistent. In the same model in Equation (1), if X is exogenous, then,

$$Cov(X, \epsilon) = 0. \tag{3}$$

QUESTION TWO

(i) **Coefficients for rprice under OLS and IV:**

(a) **OLS Coefficient**: (2 Marks)

The coefficient of rprice is -1.1506, indicating that a unit increase in cigarette prices is associated with a decrease of approximately 1.15 smoked packets, keeping other factors constant.

(b) **IV Coefficient**: (2 Marks)
The coefficient for rprice is -1.1322, which is slightly less than the OLS estimate. This suggests that after accounting for potential endogeneity using the instrument rtax, the effect of price on consumption is slightly smaller.

(ii) **Strength of the Instrument rtax:** (2 Marks)

(α) **Null Hypothesis (H_0):**
The rtax is weak. This implies the coefficient of the instrument is zero, stated mathematically as:

$$H_0 : \pi = 0 \tag{4}$$

(β) **Alternative Hypothesis (H_1):**
The rtax is strong. This implies the coefficient of the instrument is non-zero, stated mathematically as:

$$H_1 : \pi \neq 0 \tag{5}$$

where π represents the coefficient of the instrument in the first-stage regression model.

(γ) **Conclusions**
The weak instrument test statistic is 90.041 with a p-value of $2.24 * 10^{-15}$, which is highly significant. So, we reject the null hypothesis, and conclude that rtax is a strong instrument for rprice. This means the IV estimator is highly reliable.

(iii) **Endogeneity in the IV model:** (2 marks)

(α) **Null Hypothesis (H_0)**
The OLS estimator is consistent and efficient. This implies that there is no endogeneity and that the difference between the OLS and IV estimators is not statistically significant. Stated mathematically as:

$$H_0 : \hat{\beta}_{OLS} - \hat{\beta}_{IV} = 0 \tag{6}$$

(β) **Alternative Hypothesis** (H_1)
The OLS estimator is inconsistent due to endogeneity, and the IV estimator is consistent. This implies the difference between

the OLS and IV estimators is statistically significant. Stated mathematically as:

$$H_1 : \hat{\beta}_{OLS} - \hat{\beta}_{IV} \neq 0 \tag{7}$$

(γ) **Conclusions**
From the output, the p-value of the Hausman test is 0.874. Therefore, we fail to reject the null hypothesis, and conclude that the OLS estimator is consistent and efficient. This implies that there is no endogeneity and that the difference between the OLS and IV estimators is not statistically significant.

(iv) **Writing an R command** (2 Marks)

```
# Load necessary libraries
library("nlme")
library("ivreg")

# Assuming the dataset is named 'data' and has columns
'rpack', 'rprice'

# OLS Model
ols_model<-lm(rpack~rprice, data=data)
summary(ols_model)

# IV Model
model_iv<-ivreg(rpack~rprice|rtax), data=data)
summary(model_iv, diagnostics = TRUE)
```

QUESTION THREE

To derive the GLS estimator and its variance, we start with the linear regression model:

$$Y = X\beta + \epsilon \tag{8}$$

where:

(a) Y is an $n \times 1$ vector of responses

(b) X is an $n \times p$ matrix of predictors

(c) β is a $p \times 1$ vector of coefficients

(d) ϵ is an $n \times 1$ vector of errors with $\epsilon \sim N(0, \sigma^2 V)$

(e) V is a known $n \times n$ positive definite matrix.

(a) **Transform the Model in Equation (8)**
Transform the model using Cholesky decomposition of V by letting:

$$V = LL^T \tag{9}$$

to handle the correlated errors, where L is a lower triangular matrix. Further, define:

$$Y^* = L^{-1}Y, \quad X^* = L^{-1}X, \quad \epsilon^* = L^{-1}\epsilon \tag{10}$$

And the transformed model is:

$$Y^* = X^*\beta + \epsilon^* \tag{11}$$

where, $\epsilon^* \sim N(0, \sigma^2 I)$. Hence, the transformed errors are homoscedastic and uncorrelated.

(i) **Apply OLS to the Transformed Model in Equation (11) to determine** β^* (3 Marks)
The OLS estimator for β in the transformed model is:

$$\beta^* = (X^{*T}X^*)^{-1}X^{*T}Y^* \tag{12}$$

Substituting $X^* = L^{-1}X$ and $Y^* = L^{-1}Y$, we have:

$$\beta^* = (X^T L^{-T} L^{-1} X)^{-1} X^T L^{-T} L^{-1} Y \tag{13}$$

But $V = LL^T$ and $V^{-1} = L^{-T}L^{-1}$. So, we have:

$$\beta^* = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \tag{14}$$

as the generalized least squares estimate of β, in terms of X, Y, and V only.

(ii) **Derive the variance of the GLS Estimator** (3 Marks)
The variance of the OLS estimator in the transformed model is:

$$Var(\beta^*) = \sigma^2 (X^{*T}X^*)^{-1} \tag{15}$$

Substituting $X^* = L^{-1}X$, we have:

$$Var(\beta^*) = \sigma^2 (X^T L^{-T} L^{-1} X)^{-1} = \sigma^2 (X^T V^{-1} X)^{-1} \tag{16}$$

as the variance of β^*, in terms of X, Y, and V only.

QUESTION FOUR

(i) **Coefficient for price.index under OLS and GLS Models**

(a) **OLS Model:** (2 Marks)
The S.E price.index is -2.34056 with a standard error of 0.05643. This indicates that a unit increase in the price index is associated with a decrease of approximately 2.34 units in revenue, holding other factors constant.

(b) **GLS Model:** (2 Marks)
The coefficient for *price.index* is -1.226674 with an S.E of 0.274401. This suggests a smaller impact of the price index on revenue compared to the OLS model, indicating that accounting for the structure of the error terms (as in GLS) can lead to different estimates.

(ii) **Discussion of Figure 1 (Residual Plots)** (2 Marks)

(1) **Purpose of ACF Plot of Residuals**:
It shows the correlation between the residuals at different lags, and helps to identify whether there is any significant autocorrelation in the residuals.

(2) **Interpretations**:

(a) If the ACF plot shows significant spikes (outside the confidence interval) at certain lags, it suggests that the residuals are autocorrelated, indicating that the model may be misspecified (e.g., missing lagged variables or other time-dependent structures).

(b) If the ACF plot shows no significant spikes, the residuals are likely uncorrelated, which is a good sign that the model is well-specified.

(3) **Conclusion**:
From Figure 1, the ACF plot shows significant spikes (outside the confidence interval) at certain lagsand this suggests that the residuals are autocorrelated, indicating that the model may have been mis-specified or a violation of the OLS assumption.

(4) **Purpose of PACF Plot of Residuals**:
It shows the partial correlation between the residuals at different lags, controlling for the effects of intermediate lags. It

helps identify the specific lags at which autocorrelation occurs.

(5) **Interpretations**:

(a) If it shows significant spikes at certain lags, it suggests residuals have a direct relationship at those lags, which could indicate the need for additional lagged terms in the model.

(b) If it shows no significant spikes, it supports the conclusion that the residuals are uncorrelated.

(6) **Conclusion**:
From Figure 1, the PACF plot shows no significant spikes and this suggests that the residuals are uncorrelated.

(iii) **Fitted Model for Error Terms under the GLS** (2 Marks)
This is an ARMA(1,2) model, and therefore, the fitted model for the error terms under the GLS model is:

$$e_t = 0.9911e_{t-1} + 0.2361\epsilon_{t-1} - 0.2001\epsilon_{t-2} \tag{17}$$

as required

(iv) **R Program for Fitting OLS and GLS Models** (2 Marks)

```
# Load necessary library
library(nlme)

# Assuming the dataset is named 'data' and has columns
'revenue' and 'price.index'

# Fit OLS model
ols_model<-lm(revenue~price.index, data=data)
summary(ols_model)

# Fit GLS model
gls_model<-gls(revenue~price.index, data=data,
correlation=corARMA(p=1, q=2))
summary(gls_model)

# Plot residuals for OLS model
plot(residuals(ols_model), main="Residuals from OLS Model")
abline(h=0, col="red")
```