MGRECON 548Q: Empirical Economic Analysis

Fuqua School of Business, Duke University
Fall 2020

Help Section Week#3[1]

Decide whether the following statements are true or false. Explain your reasoning.

Asumptions of Odinary Least Square (OLS) Regression:

A1 (Linearity): The regression model is linear in the coefficients and the error term

$$Y = \alpha + \beta_1 X1 + \beta_2 X2 + \epsilon$$

(i) Non-linearity in Regressors:

$$\ln(Y) = \alpha + \beta_1 \ln(X_1) + \beta_2 X_2^2 + \epsilon$$

$\Longrightarrow$ problem : A1 ✓ and OLS ✓, but need to pay attention to model interpretation

$\Longrightarrow$ solution : Redefine regressors

(ii) Non-linearity in Parameters:

$$Y = \alpha + \ln(\beta_1) X_1 + \beta_2^2 X_2 + \epsilon$$

$\Longrightarrow$ problem : A1 $\times$ and OLS $\times$

$\Longrightarrow$ solution: Nonlinear Least Squares (beyond our scope)

[1] Please watch for typos and errors

A2 (Orthogonal): The error term has zero expectation and is (weakly) orthogonal with X

$$\mathrm{E}[\epsilon] = 0 \text{ and } \mathrm{E}[\epsilon \mathrm{X}] = 0$$

- By A2, we have $\mathrm{cov}(\epsilon, \mathrm{X}) = 0 \implies \rho(\epsilon, \mathrm{X}) = 0$

A3 (No perfect multicollinearity): No independent variable is a perfect linear function of other explanatory variables

- Does not rule out predictors are perfect non-linear relationship:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

- Does not rule out predictors are imperfect linear relationship:

$$\mathrm{Y} = \alpha + \beta_1 \mathrm{age} + \beta_2 \mathrm{Year\ of\ education} + \epsilon$$

  where age = 5 + Year of education + year of work

From A1 $\sim$ A3, we have the OLS estimators are unbiased.

(i) Include Irrelevant Regressors:

$$Y = \alpha + \beta_1 X1 + \beta_2 X2 + \beta_3 Z + \epsilon$$

$\implies$ A1 $\sim$ A3 $\checkmark$and OLS $\checkmark$, estimator is unbiased, std errors are incorrect so as inference

Spurious Regression Fit (Z is irrelevant):

- $$\begin{aligned} \mathrm{M0}: &\ \mathrm{Y} = \alpha + \beta_1 \mathrm{X1} + \beta_2 \mathrm{X2} + \beta_3 \mathrm{Z} + \epsilon, \quad \beta_3 = 0 \\ \mathrm{M1}: &\ \mathrm{Y} = \alpha + \beta_1 \mathrm{X1} + \beta_2 \mathrm{X2} + \beta_3 \mathrm{Z} + \epsilon \end{aligned}$$
- R^2 will never decrease if you add more regressors
- Look at R^2_{adj} and drop irrelevant regressors

Insufficient Degrees of Freedom:

- df = N-K-1
- Drop irrelevant regressors (decrease K) or collect more data (increase N)

Detect: (1) Theory; (2) Drop and check for change in overall goodness of fit

(ii) Highly Multicollinearity:

$$Y = \alpha + \beta X_1 + \gamma Z + \epsilon$$

$$\rho(\mathrm{X}_1, \mathrm{Z}) > 0.8 \text{ but } \rho(\mathrm{X}_1, \mathrm{Z}) \neq 1$$

$\implies$: A1 $\sim$ A3 $\checkmark$ and OLS $\checkmark$, estimator is unbiased, std errors are incorrect so as inference

- Recall(Lecture 5, p. 38):

$$t_{stat} = \frac{\hat{\beta}_k - 0}{se(\hat{\beta_k})} \ (H_0 : \beta = 0)$$

$$se(\hat{\beta}) = \hat{\sigma}_e \times \frac{1}{\sqrt{N}} \times \frac{1}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}} \times \frac{1}{\sqrt{(1 - R_k^2)}}$$
$$= \frac{\hat{\sigma}_e}{\sqrt{SST_k(1 - R_k^2)}}$$

$$\text{Type I error} = \Pr(|t_{stat}| > t_{crit}|\beta = 0)$$
$$\text{Type II error} = \Pr(|t_{stat}| < t_{crit}|\beta \neq 0)$$

- $R_k \uparrow \implies se(\hat{\beta}) \uparrow \implies t_{stat} \downarrow \implies$ Type II error $\uparrow \implies$ less likely to reject H_0
  $\implies$ t test is not reliable anymore, look at F test instand
- Detect:
  (1) Check correlations among regressors

  (2) Use Variance Inflation Factor (VIF): $\mathrm{VIF} = \frac{1}{1 - R_k^2}$

  (3) Watch out for low t-stats combined with high F-stats for overall goodness of fit

  (4) Drop regressors and detect changes in estimates

(iii) Perfect Multicollinearity:

$$y = \alpha + \beta X_1 + \gamma Z + \epsilon$$
$$\rho(X_1, Z) = 1$$

$\implies$: OLS $\times$

- Solution: Drop a regressor; combine regressors; joint hypothesis test

(iv) Omitted Variable Bias (OVB):

$$\text{(long) } Y = \alpha^{long} + \beta^{long} X_1 + \gamma^{long} X_2 + \epsilon_{long}$$
$$\text{(short) } Y = \alpha^{short} + \beta^{short} X_1 + \epsilon_{short}$$

$\implies$ A2 $\times$ and OLS $\checkmark$, estimator is biased

- $$\hat{\beta}^{short} = \frac{cov(X_1, Y)}{var(X_1)}$$
$$= \frac{cov(X_1, \alpha^{long})}{var(X_1)} + \frac{cov(X_1, \beta^{long} X1)}{var(X_1)} + \frac{cov(X_1, \gamma^{long} X_2)}{var(X_1)} + \frac{cov(X_1, \epsilon_{long})}{var(X_1)}$$
$$= \hat{\beta}^{long} + \hat{\gamma}^{long} \frac{cov(X_1, X_2)}{var(X_1)}$$

- If $\hat{\gamma}^{long}$ and $\frac{cov(X_1, X_2)}{var(X_1)}$ not equal to zero, then we will have a biased estimator $\hat{\beta}^{short}$

- Estimate biased part: $\hat{\gamma}^{\text{long}} \frac{\text{cov}(\text{X}_1, \text{X}_2)}{\text{var}(\text{X}_1)}$, i.e. $\hat{\gamma}^{\text{long}}\hat{\pi}_1$:

$$Y = \alpha + \beta X_1 + \gamma^{long} X_2 + \epsilon$$

$$X_2 = \pi_0 + \pi_1 X_1 + \epsilon_{X_2}$$

- Two-scale regression: Regress Y on "pure" X_1, i.e $\hat{u}$:

$$\begin{aligned} X_1 &= \lambda_0 + \lambda_1 X_2 + u \\ Y &= \alpha + \theta u + \epsilon \\ &= \alpha + \theta(X_1 - \lambda_0 - \lambda_1 X_2) + \epsilon \\ &= (\alpha - \theta\lambda_0) + \theta X_1 + (-\theta\lambda_1) X_2 + \epsilon \\ &= \tilde{\alpha} + \beta X_1 + \tilde{\gamma} X_2 + \epsilon \end{aligned}$$

  where $\tilde{\alpha} = (\alpha - \theta\lambda_0)$, $\beta = \theta$ and $\tilde{\gamma} = (-\theta\lambda_1)$

- Detect:

  (1) The single best technique for detecting omitted relevant regressors is via theory or domain expertise

  (2) Using proxies to detect: Consider some Z as a proxy for X_2. If we include Z in the "short" model and β changes it might be because:

$$\text{cov}(\text{Z}, \text{X}_1) \neq 0 \implies \text{cov}(\text{X}_2, \text{X}_1) \neq 0 \implies \text{OVB exists}$$

- Solution[2]: Controls or proxies; Instrumental Variables

[2]You will not be examined this on the midterm

A4 (No heteroscedasticity): The error term has a constant variance

$$Var(\epsilon_i|X) = \sigma^2$$

(i) Heteroscedasticity:

$$V(\epsilon_i|X) = \sigma^2_{\epsilon_i}, \quad \sigma^2_{\epsilon_i} \neq \sigma^2_{\epsilon_j}$$

$\implies$ A1 $\sim$ A3 $\checkmark$, A4 $\times$ and OLS $\checkmark$, but std errors are incorrect

- Coefficient estimates: Remain unbiased
- Inference: Standard Errors may be distorted. The usual OLS tstatistics don't follow tdistribution
- Detect:

  (1) Plot Var(residual) vs combinations of X's

  (2) Breusch Pagan (BP) test

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$
$$\hat{\epsilon}^2 = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots + \delta_k X_k + u$$

  H_0 : $\delta_0 = \delta_1 = \delta_2 = \cdots = \delta_k = 0$ (Homoscedasticity)
  H_1 : at least one of the $\delta \neq 0$ (Heteroscedasticity)

- Solution: GLS, FLGS, Hetero. robust standard errors