DegreeLive

Bachelor of Commerce (B2001) | Level 2 • Part B Major | Monash University

Introductory econometrics

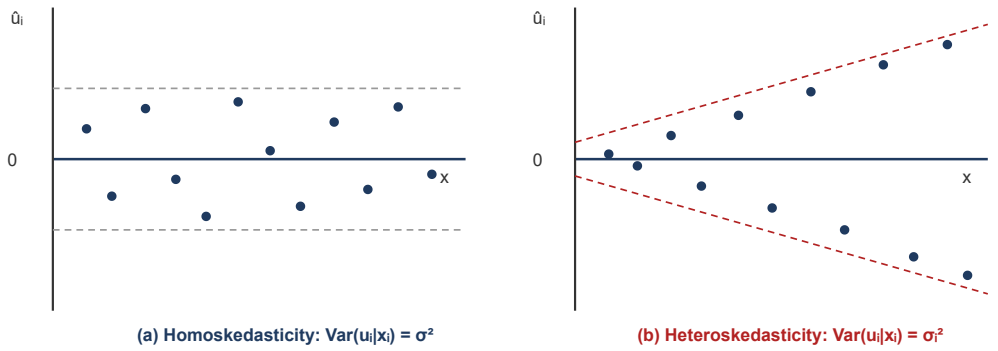
Course Code: ETC2410 • Module 3 Notes

Module 3: Model Diagnostics: Heteroskedasticity, Autocorrelation & Specification Issues

3.1 Heteroskedasticity: Causes and Consequences for OLS

Assumptions SLR.5 and MLR.5 require the conditional variance of the population error term to be constant across all observations: $\text{Var}(u_i \mid \mathbf{x}_i) = \sigma^2$ (**homoskedasticity**). In cross-sectional economic datasets—such as ABS household expenditure surveys, Australian corporate financial statements, or regional property valuations—this assumption frequently fails. When the conditional error variance depends on one or more explanatory variables, $\text{Var}(u_i \mid \mathbf{x}_i) = \sigma_i^2 = \sigma^2 h(\mathbf{x}_i)$, the error process exhibits **heteroskedasticity**.

For example, high-income Australian households possess greater discretionary flexibility in how much they spend on travel, dining, and housing than low-income households whose budgets are tightly constrained by essential living costs. Consequently, the spread of consumption residuals widens systematically as disposable income rises.



Under heteroskedasticity, what happens to Ordinary Least Squares?

- **Unbiasedness and Consistency Remain Valid:** Unbiasedness of $\hat{\boldsymbol{\beta}}$ relies solely on MLR.1–MLR.4 (zero conditional mean $E(u_i \mid \mathbf{x}_i) = 0$). Because heteroskedasticity concerns the second conditional moment rather than the first, OLS point estimates remain unbiased and consistent, and R^2 and $\bar{R}^2$ remain valid descriptive goodness-of-fit measures.
- **Loss of Efficiency (Not BLUE):** Because MLR.5 fails, OLS no longer achieves the minimum variance among linear unbiased estimators. Weighted Least Squares (WLS) places smaller weights on noisier observations and achieves strictly lower sampling variance.
- **Invalid Classical Standard Errors and Tests:** The classical variance formula $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is biased under heteroskedasticity. Consequently, conventional OLS standard errors, t-statistics, F-statistics, and confidence intervals are unreliable—often understating true sampling uncertainty and producing spurious statistical significance.

3.2 Diagnostic Tests for Heteroskedasticity

Because the squared OLS residual $\hat{u}_i^2$ estimates the conditional error variance $E(u_i^2 \mid \mathbf{x}_i) = \text{Var}(u_i \mid \mathbf{x}_i)$, formal heteroskedasticity tests examine whether $\hat{u}_i^2$ varies systematically with the explanatory variables.

The Breusch-Pagan (BP) Test

The Breusch-Pagan test evaluates the null hypothesis of homoskedasticity, $H_0 : \text{Var}(u_i \mid x_{i1}, \dots, x_{ik}) = \sigma^2$, by estimating an auxiliary regression of squared OLS residuals on all k regressors:

$$\hat{u}_i^2 = \delta_0 + \delta_1 x_{i1} + \delta_2 x_{i2} + \cdots + \delta_k x_{ik} + v_i$$

Where:

- $\hat{u}_i^2$ is the squared residual from the primary OLS regression of y_i on $x_{i1}, \dots, x_{ik}$.
- $R^2_{\hat{u}^2}$ is the R-squared from this auxiliary regression.
- v_i is the auxiliary error term.

Under $H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0$, the Lagrange Multiplier (LM) and F-statistic versions of the Breusch-Pagan test are:

$$LM_{BP} = n \cdot R^2_{\hat{u}^2} \overset{a}{\sim} \chi^2_k, \qquad F_{BP} = \frac{R^2_{\hat{u}^2}/k}{(1 - R^2_{\hat{u}^2})/(n-k-1)} \sim F_{k,\, n-k-1}$$

Where n is the sample size and k is the number of regressors in the auxiliary equation. Rejecting H_0 indicates linear dependence of error variance on the regressors.

The White Test

Because heteroskedasticity may take a quadratic or interactive form missed by a purely linear BP auxiliary equation, White (1980) proposed regressing $\hat{u}_i^2$ on all regressors, their squares (x_{ij}^2), and all pairwise cross-products ($x_{ij}x_{il}$). When k is large, this consumes many degrees of freedom ($p = k(k+3)/2$ auxiliary regressors). A parsimonious alternative—the **special White test**—regresses $\hat{u}_i^2$ on the OLS fitted values $\hat{y}_i$ and their squares $\hat{y}_i^2$:

$$\hat{u}_i^2 = \delta_0 + \delta_1 \hat{y}_i + \delta_2 \hat{y}_i^2 + v_i \quad \Longrightarrow \quad LM_W = n \cdot R^2_{\hat{u}^2} \overset{a}{\sim} \chi^2_2$$

Where $\hat{y}_i$ is a linear combination of all x_{ij}, so $\hat{y}_i$ and $\hat{y}_i^2$ compactly span linear, quadratic, and cross-product combinations of the regressors using only 2 degrees of freedom.

Worked Numerical Test (Australian Household Health Expenditure): An analyst estimates out-of-pocket medical expenditure (*medexp*, in AUD) on household income (*inc*), age (*age*), and private health insurance status (*phi*) across $n = 400$ Australian households ($k = 3$). The Breusch-Pagan auxiliary regression yields $R^2_{\hat{u}^2} = 0.045$. Computing both statistics:

$$LM_{BP} = 400 \times 0.045 = 18.00, \qquad F_{BP} = \frac{0.045/3}{(1 - 0.045)/396} = \frac{0.015}{0.002412} = 6.22$$

DegreeLive

The 1% critical value for χ^2_3 is 11.34, and for $F_{3,396}$ is 3.83. Since 18.00 > 11.34 and 6.22 > 3.83, we strongly reject homoskedasticity at the 1% level. Next, running the special White auxiliary regression of $\hat{u}_i^2$ on $\hat{y}_i$ and $\hat{y}_i^2$ produces $R^2_{\hat{u}^2} = 0.052$, giving $LM_W = 400 \times 0.052 = 20.80 > \chi^2_{2,0.01} = 9.21$, confirming heteroskedasticity.

The Goldfeld-Quandt Test

When heteroskedasticity is suspected to scale monotonically with a single ordered variable x_j (for example, firm size or household income), the **Goldfeld-Quandt test** orders the n observations from lowest to highest x_j, omits c central observations (typically 20% of the sample), and estimates separate OLS regressions on the bottom $n_1 = (n-c)/2$ and top $n_2 = (n-c)/2$ sub-samples. Under homoskedasticity and normality:

$$GQ = \frac{SSR_2/(n_2 - k - 1)}{SSR_1/(n_1 - k - 1)} \sim F_{n_2-k-1,\, n_1-k-1}$$

Where SSR_1 and SSR_2 are the residual sums of squares from the low-variance and high-variance sub-samples. Large values ($GQ > F_{crit}$) reject homoskedasticity.

Diagnostic Test	Auxiliary Equation / Procedure	Null Distribution	Key Advantage & Limitation
Breusch-Pagan (BP)	Regress $\hat{u}_i^2$ on $x_{i1}, \ldots, x_{ik}$	$nR^2_{\hat{u}^2} \sim \chi^2_k$ (or $F_{k,n-k-1}$)	High power against linear variance; misses purely quadratic forms
Full White Test	Regress $\hat{u}_i^2$ on x_{ij}, x_{ij}^2, and $x_{ij}x_{il}$	$nR^2_{\hat{u}^2} \sim \chi^2_p$	Detects general non-linear variance; consumes many degrees of freedom
Special White Test	Regress $\hat{u}_i^2$ on $\hat{y}_i$ and $\hat{y}_i^2$	$nR^2_{\hat{u}^2} \sim \chi^2_2$	Parsimonious ($df = 2$) regardless of k; good general screening test
Goldfeld-Quandt	Ratio of upper to lower sub-sample SSR	$F_{n_2-k-1,\, n_1-k-1}$	Exact finite-sample test under normality; requires sorting by a single x_j

3.3 Remedying Heteroskedasticity: Robust Standard Errors, WLS & FGLS

Consistent with ULO4, econometricians address heteroskedasticity in two distinct ways: adjusting OLS standard errors for valid inference, or transforming the model to restore efficiency via Weighted Least Squares (WLS) and Feasible Generalised Least Squares (FGLS).

White Heteroskedasticity-Consistent (HC) Robust Standard Errors

Without changing the OLS coefficient estimates $\hat{\boldsymbol{\beta}}$, White (1980) proved that even when $\text{Var}(u_i \mid \mathbf{x}_i) = \sigma_i^2$ takes an unknown form, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is consistently estimated by the "sandwich" estimator:

$$\widehat{\text{Var}}_{HC}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} \hat{u}_i^2\, \mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1}$$

Where:

- $\mathbf{x}_i'$ is the $1 \times (k+1)$ row vector of regressors (including 1) for observation i.
- $\hat{u}_i^2$ is the squared OLS residual serving as a noisy individual estimate of σ_i^2.
- $(\mathbf{X}'\mathbf{X})^{-1}$ acts as the "bread" and $\sum \hat{u}_i^2 \mathbf{x}_i\mathbf{x}_i'$ acts as the "meat" of the sandwich matrix.

Taking square roots of the diagonal entries gives White heteroskedasticity-robust standard errors ($HC0$), or with small-sample degrees-of-freedom scaling $n/(n-k-1)$ ($HC1$, the standard default in EViews and Stata). In simple regression, the robust variance of $\hat{\beta}_1$ simplifies transparently to $\sum(x_i - \bar{x})^2\hat{u}_i^2/[\sum(x_i-\bar{x})^2]^2$. Under heteroskedasticity, classical F-tests are invalid and must be replaced by heteroskedasticity-robust Wald F-tests.

Worked Numerical Comparison (Classical vs. White Robust SEs): Consider $n = 6$ Australian retail outlets recording store floor area ($x_i = 1, 2, 3, 4, 5, 6$, in 100 m²) and monthly profit ($y_i = 3.0, 5.0, 7.0, 9.0, 10.2, 16.8$, in AUD 10,000s). Here $\bar{x} = 3.5$, $\bar{y} = 8.50$, and $\sum_{i=1}^{6}(x_i - \bar{x})^2 = 17.5$. OLS yields $\hat{y}_i = -0.1600 + 2.4743x_i$, with residuals $\hat{u} = [0.686, 0.211, -0.263, -0.737, -2.011, 2.114]'$ (note larger error spread at $x = 5, 6$) and $SSR = 9.6434$.

Under classical homoskedastic assumptions, $\hat{\sigma}^2 = 9.6434/(6-2) = 2.4109$, giving $SE_{OLS}(\hat{\beta}_1) = \sqrt{2.4109/17.5} = 0.3712$. Computing the White sandwich numerator yields $\sum_{i=1}^{6}(x_i - \bar{x})^2\hat{u}_i^2 = 40.2344$, so:

$$SE_{HC0}(\hat{\beta}_1) = \sqrt{\frac{40.2344}{17.5^2}} = 0.3625, \qquad SE_{HC1}(\hat{\beta}_1) = 0.3625 \times \sqrt{\frac{6}{4}} = 0.4439$$

Where SE_{HC1} applies the finite-sample degrees-of-freedom adjustment $n/(n-k-1) = 6/4$. Because heteroskedasticity concentrates large residuals at extreme values of x_i ($x = 5, 6$), the $HC1$ robust standard error (0.4439) exceeds the biased classical OLS standard error (0.3712) by nearly 20%.

Generalised Least Squares (GLS) in Matrix Notation

Both heteroskedasticity and serial correlation are special cases of non-spherical disturbances where $\text{Var}(\mathbf{u} \mid \mathbf{X}) = \sigma^2\boldsymbol{\Omega}$, with $\boldsymbol{\Omega} \neq \mathbf{I}_n$ a known symmetric positive-definite $n \times n$ matrix. Factoring $\boldsymbol{\Omega}^{-1} = \mathbf{P}'\mathbf{P}$ and pre-multiplying $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ by $\mathbf{P}$ yields transformed errors $\mathbf{Pu}$ with covariance $\sigma^2\mathbf{P}\boldsymbol{\Omega}\mathbf{P}' = \sigma^2\mathbf{I}_n$. Applying OLS to the transformed model yields the **GLS estimator**, which is BLUE under non-spherical errors:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}, \qquad \text{Var}(\hat{\boldsymbol{\beta}}_{GLS} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$$

DegreeLive

Where $\mathbf{\Omega} = \text{diag}(h_1, h_2, \dots, h_n)$ recovers Weighted Least Squares (WLS) for pure heteroskedasticity, whereas $\mathbf{\Omega}_{ts} = \rho^{|t-s|}/(1-\rho^2)$ recovers Prais-Winsten GLS for AR(1) serial correlation.

Weighted Least Squares (WLS) and Feasible GLS (FGLS)

When the sample size is moderate or efficiency gains are desired, Generalised Least Squares (GLS) transforms the model so that the transformed errors become homoskedastic. Suppose $\text{Var}(u_i \mid \mathbf{x}_i) = \sigma^2 h_i$, where $h_i = h(\mathbf{x}_i) > 0$. Dividing every term in the population equation by $\sqrt{h_i}$ gives the **WLS transformed model**:

$$\frac{y_i}{\sqrt{h_i}} = \beta_0 \frac{1}{\sqrt{h_i}} + \beta_1 \frac{x_{i1}}{\sqrt{h_i}} + \cdots + \beta_k \frac{x_{ik}}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}}$$

Where:

- $u_i^* = u_i / \sqrt{h_i}$ is the transformed error term, with conditional variance $\text{Var}(u_i^* \mid \mathbf{x}_i) = \sigma^2$.
- Minimising the weighted sum $\sum_{i=1}^{n} \hat{u}_i^2 / h_i$ assigns smaller weight $w_i = 1/h_i$ to observations with larger error variance, making WLS BLUE when h_i is known.

Because the exact skedastic function h_i is rarely known a priori, **Feasible Generalised Least Squares (FGLS)** estimates $\hat{h}_i$ from the data using an exponential variance specification:

$$\text{Var}(u_i \mid \mathbf{x}_i) = \sigma^2 \exp(\delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik})$$

Where the exponential function guarantees strictly positive variance weights $\hat{h}_i > 0$. The four FGLS steps are:

1. Estimate the primary regression by OLS and save residuals $\hat{u}_i$.
2. Regress $\ln(\hat{u}_i^2)$ on $x_{i1}, \dots, x_{ik}$ (and a constant) to obtain fitted values $\hat{g}_i = \widehat{\ln(u_i^2)}$.
3. Exponentiate to form estimated variance weights $\hat{h}_i = \exp(\hat{g}_i)$.
4. Estimate the original regression by WLS using weights $w_i = 1/\hat{h}_i$ (equivalently, OLS on variables divided by $\sqrt{\hat{h}_i}$).

Worked WLS Example (Australian Regional Savings): Suppose household savings sav_i depends on income inc_i, and error variance is proportional to income: $\text{Var}(u_i \mid inc_i) = \sigma^2 inc_i$, so $h_i = inc_i$. Dividing the model $sav_i = \beta_0 + \beta_1 inc_i + u_i$ by $\sqrt{inc_i}$ gives:

$$\frac{sav_i}{\sqrt{inc_i}} = \beta_0 \left(\frac{1}{\sqrt{inc_i}}\right) + \beta_1 \sqrt{inc_i} + u_i^*$$

Suppose OLS on unscaled data for $n = 200$ Melbourne households yields $\hat{\beta}_1^{OLS} = 0.145$ with $SE_{OLS} = 0.032$ and $SE_{HC} = 0.044$. Estimating the transformed equation without a constant yields $\hat{\beta}_1^{WLS} = 0.152$ with $SE_{WLS} = 0.029$. Both $\hat{\beta}_1^{OLS}$ and $\hat{\beta}_1^{WLS}$ estimate the same population MPC out of income, while WLS

DegreeLive

achieves a noticeably smaller standard error (0.029 < 0.044). Note that if the WLS weighting function $\hat{h}_i$ is slightly misspecified, FGLS remains consistent under MLR.4, and reporting White robust standard errors on top of FGLS provides double protection.

3.4 Serial Correlation (Autocorrelation) in Time-Series Data

In time-series regressions ($t = 1, 2, \dots, T$), observations follow a natural chronological ordering, and shocks in period t (such as an RBA cash rate surprise, drought, or commodity price spike) often persist into period $t + 1$. When errors across distinct time periods are correlated, $\text{Cov}(u_t, u_s \mid \mathbf{X}) \neq 0$ for $t \neq s$, the model exhibits **serial correlation (autocorrelation)**.

The most common dynamic error structure is the first-order autoregressive process, **AR(1)**:

$$u_t = \rho u_{t-1} + e_t, \qquad |\rho| < 1$$

Where:

- $\rho = \text{Corr}(u_t, u_{t-1})$ is the first-order autocorrelation coefficient ($-1 < \rho < 1$ for stationarity).
- e_t is a white-noise innovation with $E(e_t) = 0$, $\text{Var}(e_t) = \sigma_e^2$, and $\text{Cov}(e_t, e_s) = 0$ for $t \neq s$.

Consequences for OLS: If all regressors are strictly exogenous ($E(u_t \mid \mathbf{X}) = 0$ across all time periods), OLS remains unbiased and consistent under serial correlation, though it is no longer BLUE and conventional OLS standard errors are severely biased (typically understated when $\rho > 0$ and regressors are positively autocorrelated). However, if the regression includes a **lagged dependent variable** (y_{t-1}) alongside autocorrelated errors ($\rho \neq 0$), then $\text{Cov}(y_{t-1}, u_t) \neq 0$, rendering OLS *biased and inconsistent*.

Testing for Serial Correlation: Durbin-Watson and Breusch-Godfrey

The classical test for AR(1) serial correlation ($H_0 : \rho = 0$) in models with strictly exogenous regressors and an intercept is the **Durbin-Watson (DW) statistic**:

$$DW = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_t^2} \approx 2(1 - \hat{\rho})$$

Where $\hat{\rho} = \sum_{t=2}^{T} \hat{u}_t \hat{u}_{t-1} / \sum_{t=1}^{T} \hat{u}_{t-1}^2$. Thus $DW \in [0, 4]$: values near 2 indicate no serial correlation ($\hat{\rho} \approx 0$), values close to 0 indicate strong positive autocorrelation ($\hat{\rho} \to 1$), and values close to 4 indicate negative autocorrelation ($\hat{\rho} \to -1$). Because the exact null distribution depends on $\mathbf{X}$, the DW test uses lower (d_L) and upper (d_U) bounds: reject H_0 against $\rho > 0$ if $DW < d_L$, do not reject if $DW > d_U$, and treat $[d_L, d_U]$ as an inconclusive zone.

Crucially, the Durbin-Watson test is **invalid when lagged dependent variables (y_{t-1}) appear as regressors** and tests only AR(1). The more general **Breusch-Godfrey (BG) LM test** accommodates both lagged dependent variables and higher-order $AR(p)$ serial correlation ($u_t = \rho_1 u_{t-1} + \cdots + \rho_p u_{t-p} + e_t$). After running OLS and saving residuals $\hat{u}_t$ (setting pre-sample residuals $\hat{u}_0, \dots, \hat{u}_{1-p}$ to 0), estimate the auxiliary regression:

Degree Live

$$\hat{u}_t = \gamma_0 + \gamma_1 x_{t1} + \cdots + \gamma_k x_{tk} + \rho_1 \hat{u}_{t-1} + \cdots + \rho_p \hat{u}_{t-p} + \text{error}_t$$

Where $x_{t1}, \ldots, x_{tk}$ include any lagged dependent variables from the original model. Under $H_0 : \rho_1 = \cdots = \rho_p = 0$, the statistic $LM_{BG} = (T-p)R^2_{\hat{u}}$ (or $T \cdot R^2_{\hat{u}}$ when zeros replace missing initial lags) is asymptotically χ^2_p.

Worked Autocorrelation Example (Australian Quarterly Retail Growth): An economist models quarterly Australian retail sales growth on disposable income growth and consumer sentiment over $T = 80$ quarters ($k = 2$). The DW statistic equals 1.18, implying $\hat{\rho} \approx 1 - 1.18/2 = 0.41$. At $\alpha = 5\%$ ($T = 80, k = 2$), $d_L = 1.59$ and $d_U = 1.69$; since $1.18 < 1.59$, we reject H_0 and confirm positive AR(1) serial correlation. Re-checking with a Breusch-Godfrey AR(4) test ($p = 4$ lags for quarterly data) gives $R^2_{\hat{u}} = 0.21$, so $LM_{BG} = 80 \times 0.21 = 16.80 > \chi^2_{4,0.01} = 13.28$, confirming significant autocorrelation.

3.5 Remedying Serial Correlation: FGLS and HAC (Newey-West) SEs

When regressors are strictly exogenous and errors follow an AR(1) process $u_t = \rho u_{t-1} + e_t$, **quasi-differencing** eliminates the autocorrelation for $t = 2, \ldots, T$:

$$y_t - \rho y_{t-1} = (1-\rho)\beta_0 + \beta_1(x_{t1} - \rho x_{t-1,1}) + \cdots + \beta_k(x_{tk} - \rho x_{t-1,k}) + e_t$$

Where e_t is serially uncorrelated white noise. Because ρ is unknown, the **Cochrane-Orcutt** procedure estimates $\hat{\rho}$ by regressing $\hat{u}_t$ on $\hat{u}_{t-1}$, transforms $t = 2, \ldots, T$, runs OLS on the quasi-differenced data, and iterates until $\hat{\rho}$ converges. The **Prais-Winsten** FGLS procedure retains the first observation by scaling $t = 1$ using the unconditional variance identity $\text{Var}(u_1) = \sigma_e^2/(1-\rho^2)$:

$$\sqrt{1-\hat{\rho}^2}\, y_1 = \beta_0 \sqrt{1-\hat{\rho}^2} + \beta_1 \sqrt{1-\hat{\rho}^2}\, x_{11} + \cdots + \beta_k \sqrt{1-\hat{\rho}^2}\, x_{1k} + \sqrt{1-\hat{\rho}^2}\, u_1$$

Where the transformed first-period disturbance $\sqrt{1-\hat{\rho}^2}\, u_1$ has variance σ_e^2, matching the variance of e_t for $t \geq 2$ and ensuring higher finite-sample efficiency than Cochrane-Orcutt when T is moderate.

Method	Requires Strict Exogeneity?	Handles Higher-Order $AR(p)$?	Efficiency & Coefficient Behaviour
Cochrane-Orcutt FGLS	Yes	AR(1) only (unless extended)	Changes $\hat{\boldsymbol{\beta}}$; drops $t = 1$; asymptotically efficient under AR(1)
Prais-Winsten FGLS	Yes	AR(1) only	Changes $\hat{\boldsymbol{\beta}}$; retains $t = 1$; BLUE-approximating FGLS
Newey-West HAC SEs	No (Contemporaneous exog. suffices)	Yes (arbitrary lags up to L)	Keeps OLS $\hat{\boldsymbol{\beta}}$ unchanged; adjusts SEs for both HC and AC

DegreeLive

When strict exogeneity is uncertain or autocorrelation has an unknown higher-order structure, practitioners keep OLS estimates (assuming contemporaneous exogeneity) and compute **Newey-West Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors**:

$$\underline{\text{Var}}_{HAC}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\hat{\boldsymbol{\Omega}}_{NW}(\mathbf{X}'\mathbf{X})^{-1}, \qquad \hat{\boldsymbol{\Omega}}_{NW} = \sum_{t=1}^{T} \hat{u}_t^2 \mathbf{x}_t \mathbf{x}_t' + \sum_{l=1}^{L} w_l \sum_{t=l+1}^{T} \hat{u}_t \hat{u}_{t-l}(\mathbf{x}_t \mathbf{x}_{t-l}' + \mathbf{x}_{t-l}\mathbf{x}_t')$$

Where:

- $w_l = 1 - \frac{l}{L+1}$ are Bartlett kernel weights ensuring the estimated covariance matrix remains positive semi-definite.
- L is the truncation lag length (commonly chosen via $L \approx 0.75\, T^{1/3}$ or $L = \lfloor 4(T/100)^{2/9} \rfloor$).

Notice that setting $L = 0$ reduces Newey-West HAC directly to White's HC covariance matrix.

3.6 Specification Issues: RESET, Measurement Error & Endogeneity

Even with valid standard errors, OLS estimates are biased and inconsistent if the regression equation is misspecified—through omitted nonlinearities, measurement error in regressors, or correlation between an explanatory variable and the error term.

Ramsey's RESET Test for Functional Form Misspecification

If an analyst specifies a linear equation when the true relationship contains omitted quadratic, cubic, or interaction terms, $E(u_i \mid \mathbf{x}_i) \neq 0$. Ramsey's **Regression Equation Specification Error Test (RESET)** detects omitted nonlinearities by augmenting the estimated regression with powers of the OLS fitted values $\hat{y}_i^2$ and $\hat{y}_i^3$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + \text{error}_i$$

Where $\hat{y}_i$ are the fitted values from the original OLS model. Testing $H_0 : \gamma_1 = \gamma_2 = 0$ via an F-test ($F \sim F_{2,n-k-3}$) evaluates whether nonlinear combinations of the included regressors explain additional variation in y. Rejection signals functional form misspecification (such as needing logarithms or polynomials), though it does not identify which specific variable is omitted.

Classical Measurement Error

Survey and macroeconomic variables frequently contain reporting or measurement noise. Suppose y_i^* is the true dependent variable and we observe $y_i = y_i^* + e_i$, where e_i has mean zero and is uncorrelated with $\mathbf{x}_i$ and u_i. Substituting into $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ gives $y_i = \mathbf{x}_i'\boldsymbol{\beta} + (u_i + e_i)$, so **measurement error in the dependent variable** leaves OLS unbiased and consistent, inflating only the error variance $\text{Var}(u_i + e_i) = \sigma_u^2 + \sigma_e^2$ and standard errors.

By contrast, **classical measurement error in an explanatory variable** causes bias and inconsistency. Suppose $y_i = \beta_0 + \beta_1 x_i^* + u_i$, but we observe $x_i = x_i^* + v_i$, where v_i is random noise with variance σ_v^2 uncorrelated with x_i^* and u_i. Substituting $x_i^* = x_i - v_i$ gives:

$$y_i = \beta_0 + \beta_1 x_i + (u_i - \beta_1 v_i)$$

Because $x_i = x_i^* + v_i$ correlates positively with v_i, $\text{Cov}(x_i, u_i - \beta_1 v_i) = -\beta_1 \sigma_v^2 \neq 0$. As $n \to \infty$, the OLS slope converges in probability to:

$$\text{plim}\,\hat{\beta}_1 = \beta_1 \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_v^2}\right)$$

Where $\sigma_{x^*}^2$ is the variance of the true signal x_i^* and σ_v^2 is the variance of measurement noise. Because the reliability ratio $\lambda = \sigma_{x^*}^2/(\sigma_{x^*}^2 + \sigma_v^2) < 1$, OLS suffers from **attenuation bias**—systematic shrinkage toward zero. For instance, if true schooling variance is 9.0 and survey reporting error variance is 1.0, OLS estimates only 9/(9+1) = 90% of the true return to education.

Endogeneity and Instrumental Variables (IV / 2SLS) Intuition

More broadly, whenever $\text{Cov}(x_{ij}, u_i) \neq 0$, regressor x_{ij} is **endogenous** and OLS is inconsistent. Endogeneity arises from three primary sources: (1) *omitted variables* correlated with both x_j and y; (2) *measurement error* in x_j; and (3) *simultaneity* (reverse causality, such as price and quantity jointly determined in market equilibrium, or RBA monetary policy responding to expected inflation while simultaneously affecting inflation).

When exogeneity fails, consistent estimation requires an **instrumental variable (z_i)** satisfying two foundational conditions:

1. **Instrument Relevance:** z_i is correlated with the endogenous regressor x_i after controlling for other exogenous covariates ($\text{Cov}(z_i, x_i) \neq 0$), testable via a first-stage F-statistic (commonly requiring $F > 10$ to avoid weak-instrument bias).
2. **Instrument Exogeneity (Exclusion Restriction):** z_i is uncorrelated with the structural error term u_i ($\text{Cov}(z_i, u_i) = 0$), meaning z_i affects y_i solely through its effect on x_i.

In simple regression with one endogenous regressor and one instrument, the IV slope estimator is:

$$\hat{\beta}_{1,IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

Where $\text{Avar}(\hat{\beta}_{1,IV}) = \sigma^2/[SST_x \cdot R_{x,z}^2]$ is always at least as large as OLS asymptotic variance because $R_{x,z}^2 \leq 1$. With multiple exogenous controls $\mathbf{x}_{i,exog}$ and $m \geq 1$ instruments $\mathbf{z}_i$, let $\mathbf{W}_i = [\mathbf{x}_{i,exog}', \mathbf{z}_i']'$ denote all exogenous variables:

1. **Stage 1 (Reduced Form):** Regress the endogenous variable x_i on $\mathbf{W}_i$ by OLS and save fitted values $\hat{x}_i$ and residuals $\hat{v}_i = x_i - \hat{x}_i$. Check instrument strength via the first-stage F-statistic on the excluded instruments (Stock-Yogo rule of thumb: $F > 10$).
2. **Stage 2 (Structural Equation):** Replace x_i with $\hat{x}_i$ in the structural model and estimate y_i on $\hat{x}_i$ and $\mathbf{x}_{i,exog}$ by OLS. (Note: standard errors must be computed using residuals $y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{2SLS}$ rather than $y_i -$

$\hat{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}_{2SLS}$, which econometric software such as EViews performs automatically.)

Testing for Endogeneity: The Hausman Control-Function Test

Because IV/2SLS is less efficient than OLS when x_i is actually exogenous ($\text{Cov}(x_i, u_i) = 0$), analysts test whether OLS and 2SLS differ significantly using the **regression-based Hausman endogeneity test**:

$$y_i = \beta_0 + \beta_1 x_i + \mathbf{x}_{i,exog}'\boldsymbol{\gamma} + \delta\hat{v}_i + \text{error}_i$$

Where $\hat{v}_i$ is the residual from the Stage 1 reduced-form regression of x_i on $\mathbf{x}_{i,exog}$ and the excluded instrument(s) z_i. Because $\hat{x}_i$ is uncorrelated with u_i when z_i is valid, any endogeneity in x_i enters through $\hat{v}_i$. Testing $H_0 : \delta = 0$ with a standard t-test (or heteroskedasticity-robust t-test) directly tests $H_0 : \text{Cov}(x_i, u_i) = 0$. Rejecting H_0 confirms endogeneity and mandates 2SLS over OLS.

Worked Numerical IV Example (Australian Wage Return to Schooling): Using HILDA survey data ($n = 2,000$), an analyst regresses $\ln(wage_i)$ on years of education ($educ_i$) and experience controls, using distance to the nearest university campus at age 17 ($nearc_i = 1$ if within 20 km) as an instrument for $educ_i$. Suppose OLS yields $\hat{\beta}_{educ}^{OLS} = 0.088$ ($SE = 0.005$). In Stage 1, $nearc_i$ increases schooling by $+0.75$ years with $F = 24.6 > 10$ (strong instrument). In Stage 2, $\hat{\beta}_{educ}^{2SLS} = 0.068$ ($SE = 0.016$). Running the Hausman control-function regression yields $\hat{\delta} = +0.022$ ($SE = 0.009$, $t = 2.44$, $p = 0.015$). Because $p < 0.05$, we reject exogeneity of $educ_i$: unobserved ability biased the OLS estimate upward (0.088 vs. 0.068), and 2SLS provides the consistent causal estimate (6.8% per year of education).

Testing Overidentifying Restrictions (Sargan Test)

When an equation is **just-identified** ($m = 1$ instrument for 1 endogenous regressor), the exclusion restriction $\text{Cov}(z_i, u_i) = 0$ cannot be tested statistically and must be justified from institutional knowledge. When an equation is **overidentified** ($m > r$ excluded instruments for r endogenous regressors), we can test whether the extra $m - r$ instruments satisfy exogeneity using the **Sargan test**:

1. Estimate the structural equation by 2SLS and save the structural 2SLS residuals $\hat{u}_{i,2SLS} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{2SLS}$.
2. Regress $\hat{u}_{i,2SLS}$ on all exogenous regressors and all m excluded instruments ($\mathbf{W}_i$) to obtain $R_{\hat{u}}^2$.
3. Compute $S = n \cdot R_{\hat{u}}^2 \overset{a}{\sim} \chi_{m-r}^2$ under the null hypothesis that all instruments are valid and exogenous.

Rejecting H_0 indicates that at least one instrument is correlated with the structural error u_i or that the structural functional form is misspecified.

Proxy Variables for Unobserved Confounders

When an unobservable variable q_i^* (such as innate worker ability in a wage regression or corporate management quality in an ASX firm profitability study) is omitted, an alternative to instrumental variables is incorporating a **proxy variable** q_i (such as an ATAR score, IQ test, or prior academic GPA) satisfying $q_i^* = \delta_0 + \delta_1 q_i + r_i$. Substituting into $y_i = \beta_0 + \beta_1 x_i + \beta_2 q_i^* + u_i$ gives:

$$y_i = (\beta_0 + \beta_2\delta_0) + \beta_1 x_i + \beta_2\delta_1 q_i + (u_i + \beta_2 r_i)$$

Degreelive

Where OLS consistently estimates β_1 provided two conditions hold: (1) $E(u_i \mid x_i, q_i^*, q_i) = 0$ (q_i is redundant once q_i^* is known), and (2) $E(q_i^* \mid x_i, q_i) = \delta_0 + \delta_1 q_i$ (the proxy error r_i is uncorrelated with x_i after controlling for q_i).

Clustered Standard Errors and Non-Nested Model Testing

Even in cross-sectional datasets, observations frequently share common group shocks—for instance, Australian students sampled within the same school, or households sampled within the same ABS Statistical Area Level 2 (SA2). Ignoring positive intra-cluster error correlation ($\rho_u > 0$) when regressors also correlate within groups ($\rho_x > 0$) causes White HC standard errors to severely understate true sampling uncertainty by the **Moulton factor** $[1 + \rho_x \rho_u (\bar{m} - 1)]$, where $\bar{m}$ is average cluster size. Replacing individual outer products in the White sandwich with cluster-summed score vectors $\mathbf{s}_g = \sum_{i \in g} \mathbf{x}_i \hat{u}_i$ (summing $\mathbf{s}_g \mathbf{s}_g^{'}$ across $g = 1, \dots, G$) yields **Cluster-Robust Standard Errors**, valid whenever the number of clusters G is reasonably large ($G \geq 40$).

Finally, when comparing two competing *non-nested* specifications (for example, Model A: $y_i = \mathbf{X}_i^{'}\boldsymbol{\beta} + u_i$ vs. Model B: $y_i = \mathbf{Z}_i^{'}\boldsymbol{\gamma} + v_i$ where neither set of regressors is a subset of the other), the **Davidson-MacKinnon J-test** tests Model A by saving fitted values $\hat{y}_i^{(B)}$ from Model B and estimating $y_i = \mathbf{X}_i^{'}\boldsymbol{\beta} + \theta \hat{y}_i^{(B)} + \text{error}_i$. A significant t-statistic on $\hat{\theta}$ rejects Model A in favour of incorporating information from Model B; reversing the roles tests Model B.

DegreeLive