Overview of Regression Models *Cross-Sectional Data*

	OLS *(Ordinary Least Squares)*	**IV** *(Instrument Variables)*	**2SLS** *(2 Stage Least Squares)*	**GLS** *(General Least Squares)*
Estimator: $\hat{\beta}$	$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$ *simple model with one x:* $\hat{\beta}_{OLS} = \frac{\hat{c}ov(x_i, y_i)}{\hat{v}ar(x_i)}$	$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$ *simple model:* $\hat{\beta}_{IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$	$\hat{\beta}_{2SLS} = (\hat{X}'X)^{-1}(\hat{X}'Y) = (\hat{X}'\hat{X})^{-1}(\hat{X}'Y)$	$\hat{\beta}_{GLS} = (X'\Omega^{'-\frac{1}{2}} \cdot \Omega^{-\frac{1}{2}}X)^{-1} \cdot X'\Omega^{'-\frac{1}{2}} \cdot \Omega^{-\frac{1}{2}}y$ $\hat{\beta}_{GLS} = \left(\sum_{i=1}^{N} X_i'\Omega^{-1}X_i\right)^{-1} \sum_{i=1}^{N} X_i'\Omega^{-1}y_i$
Population: β	$\beta_{OLS} = E[(X'X)]^{-1}E[X'y]$	$\beta_{IV} = E(Z'X)^{-1}E(Z'y)$, *simple*: $\beta_{IV} = \frac{cov(z, y)}{cov(z, x)}$	$\beta_{2SLS} = E[(\hat{X}'X)^{-1}(\hat{X}'Y)] = E[(\hat{X}'\hat{X})^{-1}(\hat{X}'Y)]$	$\beta_{GLS} = E(x_i'\Omega^{-1}x_i)^{-1}E(x_i'\Omega^{-1}y_i)$
Model: y	$y_i = x_i\beta + u_i$	$y_i = x_i\beta + u_i$	$y_i = x_i'\beta + u_i$	$y_i = x_i'\beta + u_i$
Why OLS fails		$E(x_i \cdot u_i) \neq 0$ or $cov(x_K, u) \neq 0$, x_i endogenous	$E(x_i \cdot u_i) \neq 0$ or $cov(x_K, u) \neq 0$, x_i endogenous	Heteroskedasticity and autocorrelation possible
Assumptions	- **OLS.1**: Linearity: Observations are IID and satisfy $y_i = x_i'\beta + u_i$ - **OLS.2**: Strict Exogenity: $E(u_i \mid X) = 0$ - **OLS.3**: Variables have finite second moments: $E(y_i^2) < \infty$ - **OLS.4**: Invertibility (no multicollinearity) $E(x_i x_i') = Q_{xx}$ is positive definite - **OLS.5**: Homoskedastic: $\Omega = E(uu' \mid X) = \sigma^2$	- **IV.1**: exogenity (exclusion restriction): $cov(z_1, u_i) = 0$ *(not testable)* - **IV.2**: relevance: $\theta_1 \neq 0$, where $x_i = \delta_0 + \delta_j x_j + \theta_i z_i + r_k$, $E(r_k) = 0$ *(test by first stage)*	- **2SLS.1**: exogenity $E(z'u) = 0$ - **2SLS.2**: a) rank $E(z'z) = L$ b) rank $E(z'x) = K, L \geq K$ - **2SLS.3**: Homoscedasticity: $E(u^2 \mid z) = \sigma^2$	- **GLS.1**: $E(X_i \otimes u_i) = 0$ $(cor(X_i, u_i) = 0) \rightarrow$ implies: $E(u_i) = 0$, *alternatively and simpler*: $E(X_i'\Omega^{-1}u_i) = 0$ - **GLS.2**: Ω is positive definite *and* $E(X_i'\Omega^{-1}X_i)$ is nonsingular - **GLS.3**: System homoscedasticity assumption: $E(X_i'\Omega^{-1}u_u u_i'\Omega^{-1}X_i) = E(X_i'\Omega^{-1}X_i)$
Structure			1st Stage: $z \equiv (1, x_2, \ldots, x_{K-1}, z_1, \ldots, z_M)$ $\hat{x}_k = \hat{\delta}_0 + \hat{\delta}_1 x_1 + \ldots + \hat{\delta}_{K-1}x_{K-1} + \hat{\theta}_1 z_1 + \ldots + \hat{\theta}_M z_M + r_k$ Test: $H_0 : \theta_i = 0 \ \forall_i$ 2nd Stage: $\hat{x}_i \equiv (1, x_{i,2}, \ldots, x_{i,K-1}, \hat{x}_{i,K})$ $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{K-1}\hat{x}_{K-1} + \beta_K \hat{x}_K + e$	We estimate a model where the error vector has a scalar var-cov matrix: $\Omega^{-\frac{1}{2}}y_i = \Omega^{-\frac{1}{2}} \cdot X_i\beta + \Omega^{-\frac{1}{2}} \cdot u_i$, where $\Omega = E(u_i u_i')$ thus $y = X\beta + u$ must hold.
Specials		unique solution only if rank $E(z'x) = K$	- rank $E(z'x) = K, M \geq K$ - $\hat{X} = Z(Z'Z)^{-1}Z'X = PX$	$Var(\tilde{u} \mid \tilde{X}) = Var(\Omega^{-\frac{1}{2}} \cdot u) = \Omega^{-1} \cdot Var(u)$ $= Var(u)^{-1} \cdot Var(u) = I_n$ $\longrightarrow$ since this error term is homoscedastic, GLS must be BLUE
Properties	- **unbiased** if OLS.1 - OLS.4 hold: $E(\hat{\beta}_{OLS}) = \beta$ - **consistent** if OLS.1, OLS.3, OLS.4 & $E(u_i x_i) = 0$ hold - *asymptotically* **efficient** if OLS.1 - OLS.5 hold - **BLUE** if OLS.1 - OLS.5 holds	- **<u>not</u> unbiased** - **consistent** if IV.1 and IV.2 hold: $\hat{\beta}_{IV} \xrightarrow{p} \beta$ as $N \rightarrow \infty$	- **consistent** if 2SLS.1 - 2SLS.2 hold: $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$ as $N \rightarrow \infty$ - *asymptotically* **normal** if 2SLS.1 - 2SLS.2 hold - *asymptotically* **efficient** if 2SLS.1 - 2SLS.3 hold - in the class of all instrument variables estimators using instruments linear in z	- **consistent** if GLS.1 - GLS.2 hold: $\hat{\beta}_{GLS} \xrightarrow{p} \beta$ as $N \rightarrow \infty$ - *asymptotically* **normal** if GLS.1 - GLS.2 hold - *asymptotically* **efficient** if GLS.1 - GLS.3 hold (no estimator with smaller variance) - **BLUE** even if homoscedasticity does <u>not</u> hold
Variance	<u>Heteroscedastic Robust Variance conditional on x:</u> $V_{\hat{\beta}} = Var(\hat{\beta} \mid X) = (X'X)^{-1}(X'\Omega X)(X'X)^{-1}$ $\hat{V}_{\hat{\beta}} = \hat{V}ar(\hat{\beta} \mid X) = (X'X)^{-1}(X'\hat{\Omega}X)(X'X)^{-1}$ $se(\hat{\beta}_j) = \sqrt{[\hat{V}_{\hat{\beta}}]_{jj}}$ <u>If OLS.5 holds (homoskedastic):</u> $V_{\hat{\beta}} = Var(\hat{\beta} \mid X) = \sigma^2(X'X)^{-1}$ $\hat{V}_{\hat{\beta}} = \hat{V}ar(\hat{\beta} \mid X) = \hat{s}^2(X'X)^{-1}$, $\hat{s}^2 = \frac{1}{N-K}\sum_{i=1}^{N}\hat{u}_i^2$			$Var(\hat{\beta} \mid X) = (X'\Omega^{-1}X)^{-1}$ - GLS gives more weight to those observations which provide more useful information - Ω is higher where the variance of the error is higher, therefore Ω^{-1} puts less weight on this observations Feasible GLS: $\Omega = Var(u \mid X) = E(v_i v_i')$ Ω is symmetric: $\Omega = \Omega'$
Asymptotic Distribution $\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V)$	<u>Heteroscedastic Robust Asymptotic Variance (OLS.1 - OLS.4):</u> $V_{\beta} = AVar(\sqrt{N}(\hat{\beta} - \beta)) =$ $[E(x_i x_i')]^{-1}E(x_i x_i' u_i^2)[E(x_i x_i')]^{-1}$ $\hat{V}_{\beta} = A\hat{V}ar(\sqrt{N}(\hat{\beta} - \beta)) =$ $\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i' \hat{u}_i^2\right)\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1}$ $\rightarrow$ It holds that: $n \cdot \hat{V}_{\hat{\beta}} \xrightarrow{p} V_{\beta} \mid n\hat{V}_{\hat{\beta}} = \hat{V}_{\beta} \mid \hat{V}_{\hat{\beta}} \xrightarrow{p} V_{\hat{\beta}} \approx \frac{V_{\beta}}{n}$		<u>Robust Asymptotic Variance (2SLS.1 - 2SLS.2):</u> $A\hat{V}ar(\sqrt{N}(\hat{\beta} - \beta)) = (\hat{X}'\hat{X})^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2 \hat{x}_i'\hat{x}_i\right)(\hat{X}'\hat{X})^{-1}$ <u>If additionally 2SLS3 holds:</u> $AVar(\sqrt{N}(\hat{\beta} - \beta)) = \sigma^2[E(x'z)^{-1}E(z'z)E(z'x)^{-1}]$ $A\hat{V}ar(\sqrt{N}(\hat{\beta} - \beta) = \hat{\sigma}^2(\hat{X}'\hat{X})^{-1}$ where $\hat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^{N}\hat{u}_i^2$	<u>Robust Asymptotic Variance (GLS.1 - GLS.2):</u> $A\hat{V}ar(\sqrt{N}(\hat{\beta}_{GLS} - \beta)) =$ $E(X_i'\Omega^{-1}X_i)^{-1}E(X_i'\Omega^{-1}u_u u_i'\Omega^{-1}X_i)E(X_i'\Omega^{-1}X_i)^{-1}$
Notes	special case of GLS		2SLS equals IV if there is only one instrument for x_k	GLS is not feasible as Variance is unknown. In FGLS we use the estimated variance More efficient when there is autocorrelation or heteroskedasticity (different u are correlated)

Overview of Regression Models *Panel Data*

	POLS *(Pooled OLS)*	**RE** *(Random Effects)*	**FE** *(Fixed Effects)*	**FD** *(First Differences)*
Estimator: $\hat{\beta}$	$\hat{\beta}_{POLS} = \left(\sum_{i=1}^N \sum_{t=1}^T x_{it}'x_{it}\right)^{-1}\left(\sum_{i=1}^N \sum_{t=1}^T x_{it}'y_{it}\right)$	$\hat{\beta}_{RE} = \left(\sum_{i=1}^N X_i'\hat{\Omega}^{-1}X_i\right)^{-1}\left(\sum_{i=1}^N X_i'\hat{\Omega}^{-1}y_i\right)$	$\hat{\beta}_{FE} = \left(\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}'\ddot{x}_{it}\right)^{-1}\left(\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}'\ddot{y}_{it}\right)$	$\hat{\beta}_{FD} = (\Delta X'\Delta X)^{-1}\Delta X'\Delta y$
Model: y	$y_{it} = x_{it}\beta + v_{it}$, where $v_{it} = c_i + u_{it}$	$y_{it} = x_{it}\beta + v_{it}$, where $v_{it} = c_i + u_{it}$	$y_{it} = x_{it}\beta + c_i + u_{it}$	$y_{it} = x_{it}\beta + c_i + u_{it}$
Why POLS fails			c_i arbitrarily correlated with the x_{it}	c_i arbitrarily correlated with the x_{it}
Assumptions	- **POLS.1**: contemporaneous (only for the same time) exogenity $E(x_t'v_t) = 0,\ t = 1,...,T \rightarrow E(x_t'c) = 0$ - **POLS.2**: rank $\left[\sum_{t=1}^T E(x_t'x_t)\right] = K$ - **POLS.3**: a) homoscedasticity: $E(u_t^2x_t'x_t) = \sigma^2E(x_t'x_t)$ b) no serial correlation: $E(u_tu_sx_t'x_s) = 0$	- **RE.1**: unrelated effects: a) strict exogenity: $E(u_{it} \mid x_{i1}, x_{i2}, \dots, x_{iT}, c_i) = 0$ b) orthogonality between c_i and x_{it}: $E(c_{it} \mid x_{i1}, x_{i2}, \dots, x_{iT}) = E(c_i) = 0$ - **RE.2**: $\Omega = E(v_iv_i') = Var(v_i)$ is nonsingular and rank $E(X_i'\Omega^{-1}X_i) = K$ - **RE.3**: a) homoscedasticity on u: $E(u_iu_i' \mid x_i, c_i) = \sigma_u^2I_T$ b) homo on the unobserved effect c_i: $E(c_i^2 \mid x_i) = \sigma_c^2$	- **FE.1**: strict exogenity: $E(u_{it} \mid x_i, c_i) = 0$ - **FE.2**: rank $\left[\sum_{t=1}^T E(\ddot{x}_{it}'\ddot{x}_{it})\right] = K$ (rules out elements without time variation (c)) - **FE.3**: $E(u_iu_i' \mid x_i, c_i) = \sigma_u^2I_T$ (idiosyncratic errors have constant variance across t (homoscedasticity) & no serially correlation (serial uncorrelated))	- **FD.1**: strict exogenity: $E(u_{it} \mid x_i, c_i) = 0$ - **FD.2**: rank $\left[\sum_{t=2}^T E(\Delta x_{it}'\Delta x_{it})\right] = K$ - **FD.3**: $E(e_ie_i' \mid x_{i1}, \dots, x_{iT}, c_i) = \sigma_e^2I_{T-1},\ e_{it} = \Delta u_{it}$ (homoscedasticity and no series correlation of first differences Δu_{it})
Structure	The Pooled OLS model applies the Ordinary Least Squares (OLS) methodology to panel data	- The individual-specific effect is a random variable c that is uncorrelated with the explanatory variables, thus its in the error term. - RE is asymptotically equivalent to GLS under RE.1-RE.3	- Average the original equation across t to get a cross sectional equation: $\bar{y}_i = \bar{x}_i\beta + c_i + \bar{u}_i$ (<u>between equation</u>) - $y_{it} - \bar{y} = (x_{it} - \bar{x}_i)\beta + u_{it} - \bar{u} \rightarrow \ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{u}_{it}$	- The FD estimator is the POLS estimator from the regression Δy_{it} on Δx_{it} - FD explicitly lose the first time period Δy_{it} - $\Delta x_{it} = x_{i,t} - x_{i,t-1}$
Specials	- Appropriate if there is no reason to belief that there is individual or time specific effects in the data, i.e. each observation is indecent of each other - Inference should be made robust to serial correlation and heteroskedasticity	More efficient than POLS if $Var(c_i) > 0$	Removes c_i	Removes c_i by differencing adjacent observations
Properties	- **unbiased** if POLS.1 hold: $E(\hat{\beta}_{POLS}) = \beta$ - **consistent** if POLS.1 - POLS.2 hold: $\hat{\beta}_{POLS} \xrightarrow{p} \beta$ as $N \rightarrow \infty$ - *asymptomatically* **normal** if POLS.1 and POLS.2 hold	- **unbiased** if RE.1 hold: $E(\hat{\beta}_{RE}) = \beta$ - **consistent** if RE.1 and RE.2 hold $\hat{\beta}_{RE} \xrightarrow{p} \beta$ as $N \rightarrow \infty$ - *asymptotically* **normal** if RE.1 and RE.2 hold - *asymptotically* **efficient** if RE.1 - RE.3 hold - in the class of estimators consistent under $E(v_i \mid x_i) = 0$	- **unbiased** if FE.1 hold: $E(\hat{\beta}_{FE}) = \beta$ - **consistent** if FE.1 and FE.2 hold: $\hat{\beta}_{FE} \xrightarrow{p} \beta$ as $N \rightarrow \infty$ - *asymptotically* **normal** if FE.1 - FE.2 hold - *asymptotically* **efficient** if FE.1 - FE.3 hold - in the class of all estimators using the strict exogenity assumption FE.1	- **unbiased** if FD.1 hold: $E(\hat{\beta}_{FD}) = \beta$ - **consistent** if FD.1 and FD.2 hold: $\hat{\beta}_{FD} \xrightarrow{p} \beta$ as $N \rightarrow \infty$ - *asymptotically* **normal** if FD.1 - FD.2 hold - *asymptotically* **efficient** if FD.1 - FD.3 hold - in the class of all estimators using strict the exogenity assumption FD.1, *If FE.3 is violated but FD.3 isn't FD is the better estimator*
Asymptotic Distribution $\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0,V)$	<u>Heteroscedastic Robust Asymptotic Variance:</u> $A\hat{V}ar(\sqrt{N}(\hat{\beta}_{POLS} - \beta)) = \left(\sum_{i=1}^N \sum_{t=1}^T x_{it}'x_{it}\right)^{-1}\left(\sum_{i=1}^N \sum_{t=1}^T \sum_{r=1}^T \hat{v}_{it}\hat{v}_{it}x_{it}'x_{ir}\right)\left(\sum_{i=1}^N \sum_{t=1}^T x_{it}'x_{it}\right)^{-1}$ <u>Homoscedastic Asymptotic Variance (POLS.3):</u> $AVar(\sqrt{N}(\hat{\beta}_{POLS} - \beta)) = \sigma^2[E(X_i'X_i')^{-1}]/N$ $A\hat{V}ar(\sqrt{N}(\hat{\beta}_{POLS} - \beta)) = \hat{\sigma}^2(X'X)^{-1} = \hat{\sigma}^2\left(\sum_{i=1}^N \sum_{t=1}^T x_{it}'x_{it}\right)^{-1}$ $\rightarrow \hat{\sigma}^2$ is the usual OLS variance estimator from the pooled regression	<u>Heteroscedastic Robust Asymptotic Variance:</u> $A\hat{V}ar(\sqrt{N}(\hat{\beta}_{RE} - \beta)) = \left(\sum_{i=1}^N X_i'\hat{\Omega}^{-1}X_i\right)^{-1}\left(\sum_{i=1}^N X_i'\hat{\Omega}^{-1}\hat{v}_i\hat{v}_i\hat{\Omega}^{-1}X_i\right)\left(\sum_{i=1}^N X_i'\hat{\Omega}^{-1}X_i\right)^{-1}$ *serial correlation not changing variances* $\Omega \neq E(v_iv_i')$ <u>If additionally RE.3 holds:</u> $A\hat{V}ar(\sqrt{N}(\hat{\beta}_{RE} - \beta)) = \left(\sum_{i=1}^N X_i'\hat{\Omega}^{-1}X_i\right)^{-1}$ $\hat{\Omega} = \begin{pmatrix} \hat{\sigma}_v^2 & \dots & \hat{\sigma}_c^2 & \hat{\sigma}_c^2 \\ \hat{\sigma}_c^2 & \hat{\sigma}_v^2 & & \hat{\sigma}_c^2 \\ \vdots & & \ddots & \vdots \\ \hat{\sigma}_c^2 & \dots & \hat{\sigma}_c^2 & \hat{\sigma}_v^2 \end{pmatrix} = \hat{\sigma}_u^2I_T + \hat{\sigma}_c^2j_Tj_T'$	<u>Heteroscedastic Robust Asymptotic Variance:</u> $A\hat{V}ar(\sqrt{N}(\hat{\beta}_{FE} - \beta)) = \left(\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}'\ddot{x}_{it}\right)^{-1}\left(\sum_{i=1}^N \sum_{t=1}^T \sum_{r=1}^T \hat{v}_{it}\hat{v}_{it}\ddot{x}_{it}'\ddot{x}_{ir}\right)\left(\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}'\ddot{x}_{it}\right)^{-1}$ <u>If additionally FE.3 holds:</u> $Avar(\sqrt{N}(\hat{\beta}_{FE} - \beta)) = \sigma_u^2[E(\ddot{X}_t'\ddot{X}_i)]^{-1}$ $A\hat{v}ar(\sqrt{N}(\hat{\beta}_{FE} - \beta)) = \hat{\sigma}_u^2[E(\ddot{X}_t'\ddot{X}_i)]^{-1}$ $\hat{\sigma}_u^2 = \frac{\sum\sum \hat{u}_{it}^2}{N \cdot (T-1) - K} = \frac{\text{SSR}_{\text{FE}}}{N(T-1) - K}$	<u>Heteroscedastic Robust Asymptotic Variance:</u> $A\hat{V}ar(\sqrt{N}(\hat{\beta}_{FD} - \beta)) = (\Delta X'\Delta X)^{-1}\left(\sum_{i=1}^N \Delta X_i'\hat{e}_i\hat{e}_i'\Delta X_i\right)(\Delta X'\Delta X)^{-1}$ <u>If additionally FD.3 holds:</u> $A\hat{V}ar(\sqrt{N}(\hat{\beta}_{FD} - \beta)) = \hat{\sigma}_e^2(\Delta X'\Delta X)^{-1}$, $\hat{\sigma}_e^2 = \frac{\sum_{i=1}^N \sum_{t=2}^T \hat{e}_{it}^2}{N(T-1) - K} = \frac{\text{SSR}_{\text{FD}}}{N(T-1) - K}$
Notes	OLS for panel data but we ignore the fact that it is panel data	POLS: contemporaneous exogenity, RE: strict exogenity	FE uses within variance, between is lost due to demeaning RE uses within and between variance	FE & FD are the same when T=2 If u_{it} follows a random walk FD is more efficient than FE

Jacob J., Osburg P., Sarrazin J. (2023). Overview of Regression Models