

Trend Analysis of Gender Data

Workshop on Gender Statistics and Analysis Module II – Topic 2

Providenciales, Turks and Caicos Islands
16-19 February 2026



UNITED NATIONS

ECLAC

UNITED NATIONS ECONOMIC COMMISSION FOR LATIN AMERICA AND THE CARIBBEAN, SUBREGIONAL HEADQUARTERS FOR THE CARIBBEAN

The topics covered are:

General Considerations

- Cross-sectional vs. time series data
- Statistical softwares

Trend lines

- Cross-sectional vs. time series data
- Line graphs
- Scatter plots

Correlation and regression analyses

- Correlation vs. causation
- Types of correlations
- Time-series Pearson correlations
- Types of regressions
- Time-series regressions



The background is a solid orange color with a gradient. There are two large, overlapping, semi-transparent shapes: a yellow-orange circle on the left and a lighter orange circle on the right. The text "General Considerations" is centered in white, bold font.

General Considerations

Point-in-time vs. Trend Analyses

	Cross-Sectional	Time-Series
Characteristics	<ul style="list-style-type: none"> • Observations measured at one point in time • Compares different units (people, countries, firms, regions) • Focus: differences between units • Assumes observations are independent 	<ul style="list-style-type: none"> • Observations measured across multiple time periods • Follows one unit over time • Focus: change and dynamics • Observations often correlated over time (autocorrelation)
Typical questions	<ul style="list-style-type: none"> • <i>Why do countries differ in income?</i> • <i>What predicts individual health outcomes?</i> 	<ul style="list-style-type: none"> • <i>How does unemployment evolve over years?</i> • <i>Do policies have delayed effects?</i>
Methods	<ul style="list-style-type: none"> • Linear regressions • Logistic regressions • Descriptive stats (tables, graphs) 	<ul style="list-style-type: none"> • Time-series regressions • ARIMA models • Panel regressions • Descriptive statistics (trendlines)





Trendlines

Line graphs

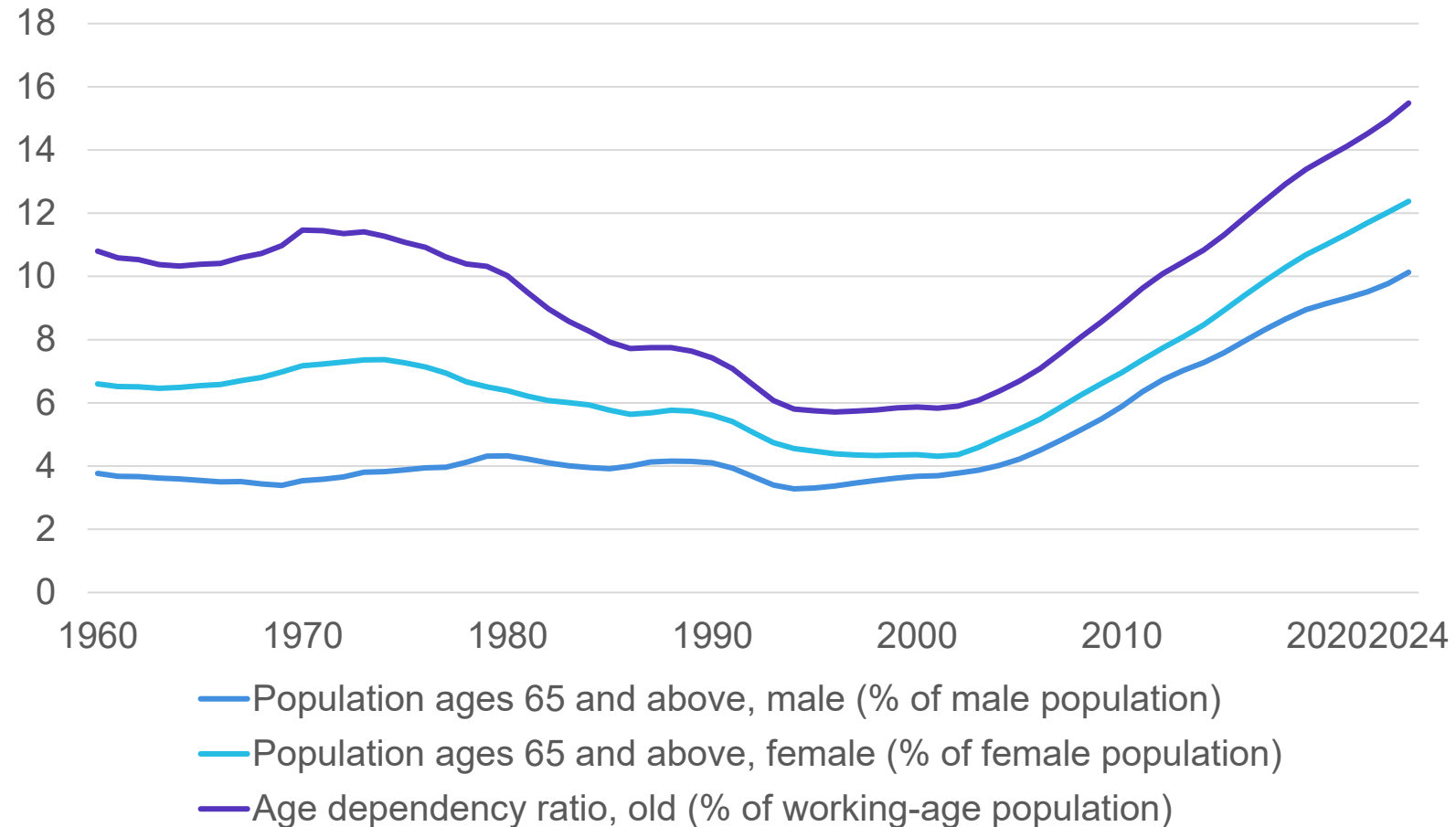
Use when you want to show...

- Trends and changes over time
- Direction, pace, and turning points
- Comparisons of trends across groups

Avoid when...

- You have very few time points
- The data are irregular or sparse

Ageing indicators in the Turks and Caicos Islands by gender, 1960 to 2020



Source: World Bank, World Development Indicators



Scatterplots

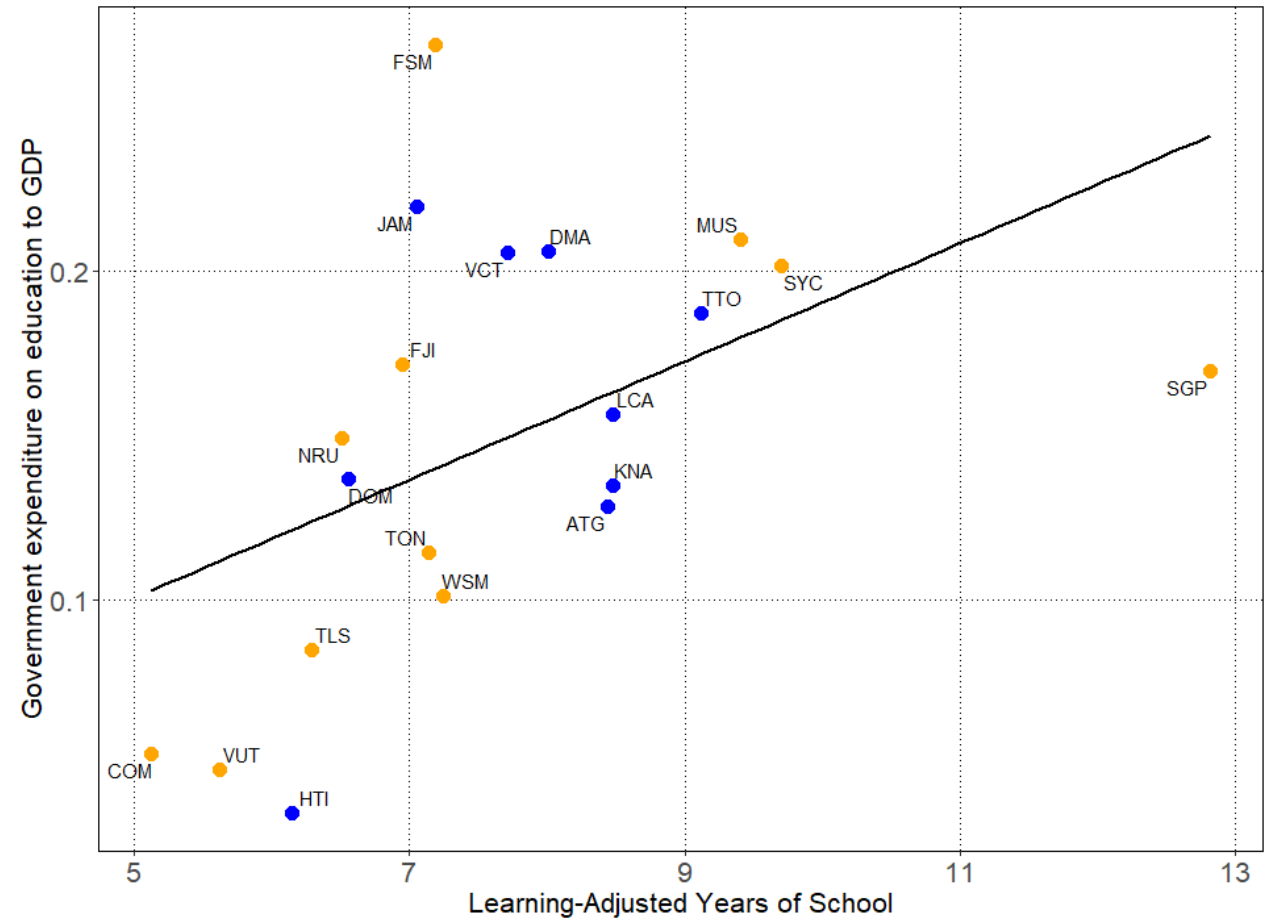
Use when you want to show...

- Patterns or variability over time or across categories
- Outliers or unusual observations
- Visualize correlation or regression lines

Avoid when...

- Communicating key messages to non-technical audiences

Learning-Adjusted Years of School and government expenditures on education to GDP (%), scaled for population under 18 years old (2020)



Source: ECLAC



Correlation and Regression Analyses

Correlation vs. Causation

Correlation

A relationship between two variables = B happens *when* A happens. However, B may not happen because of A. A confounding variable (a third variable) may be causing both or A can also lead to B (reverse causation)

Causation

According to J.S. Mill, the cause of an effect is when A *invariably* leads to B. Statistical inference (and theoretical logic) can help point toward conditioned relationships (through regressions), but the gold standard is randomized control trials (experiments; not possible with observational data)



Correlation vs. Regression

	Correlation	Regression
Objective	Measures the strength and direction of an association	Estimates the impact of one or more variable variable on an outcome
Directionality	<i>Symmetrical</i> : The relationship between X and Y is the same as Y and X	<i>Asymmetrical</i> : The independent variables condition (not cause) the effect on the outcome
Output	A standardized coefficient between -1 and 1	An unstandardized estimate representing the change in the outcome's unit (e.g., dollars, years, births) for a one-unit change in the independent variable
Complexity	Limited to the association between two variables	Can involve multiple predictors (independent or condition variable) simultaneously (Multivariate Regression)



Types of correlations

Pearson

Degree of the relationship between linearly related variables

Assumptions: Variables need to be continuous, linearly related and normally distributed

Kendall

Degree of dependence between two non-parametric variables

Used for continuous but non-linear data or ordinal variables

Spearman

Degree of association between two ordinal variables

More appropriate test for ordinal (ranked) variables



A warning about trend data analyses: *Autocorrelation*

- Measures how a time series relates to its own past values (in technical terms, correlation between residuals across the unit of time)
 - *Very common in time series!*
- Evaluates correlation between observations at different lags
- Autocorrelated time-series inflate the statistical significance of a correlation (can lead to spurious results)
- Durbin-Watson test for autocorrelation: measure from 0 to 4; results from 1.5 to 2.5 mean no autocorrelation (0 to 1.4 = positive autocorrelation; 2.6 to 4 = negative autocorrelation)



Pearson's Correlation

Net migration and population of older women (1960-2024)

		Pearson's r	p
Net migration	- Population ages 65 and above, female (% of female population)	0.240	.054

Shapiro-Wilk Test for Bivariate Normality: 0.905 ($p < .001$)

Durbin-Watson Autocorrelation: 0.02



Pearson's Correlation

Life expectancy by gender and urbanization (1960-2024)

		Pearson's r	p
Mortality rate, adult, female (per 1,000 female adults)	- Fertility rate, total (births per woman)	0.986	< .001
<i>Shapiro-Wilk Test for Bivariate Normality:</i> 0.019 (p < .001)			
Durbin-Watson Autocorrelation: 0.05			



Types of (Multivariate) Regressions

Equation:

$$Y = a + bX_1 + bX_2 + \dots bX_5 + e$$

Linear

Predicts the estimate relationship between independent variables and continuous variable

Logistic

Predicts a categorical outcome (usually binary)

The estimate is not a value change, but the likelihood of an event occurring

GLM

Generalized Linear Model

A GLM is an overarching framework for linear regressions that are not normally distributed, thus combining linear and logistic regressions. Allows for the analysis of counts (Poisson), proportions, or skewed continuous data without needing complex data transformations.



Linear time series regressions

Core idea

Estimate relationships between a dependent variable and one or more predictors using historical data

Statistical model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \dots + \beta_k X_{kt} + \varepsilon_t$$

Interpretation

Coefficients quantify the expected change in the outcome variable. These coefficients estimates represent the change in the outcome's unit (e.g., dollars, years, births) for a one-unit change in the independent variable

Assumptions

- Linear relation between X and Y
- Independence of errors: lack of autocorrelation
- Homoscedasticity: constant variance of residual errors across independent variables
- No severe multicollinearity: independent variables are not highly correlated with one another



Linear Regression

Y=Dependency ratio, older persons

X=net migration controlled by population by gender (1960-2024)

Model Summary - Age dependency ratio, old (% of working-age population)

Model	R	R ²	Adjusted R ²	RMSE	Durbin-Watson		
					Autocorrelation	Statistic	p
M ₁	0.940	0.884	0.878	0.904	0.935	0.066	< .001

Note. M₁ includes Net migration, Population, male, Population, female

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
M ₁	(Intercept)	7.863	0.192		40.919	< .001
	Net migration	-4.900×10 ⁻⁴	4.310×10 ⁻⁴	-0.088	-1.137	.260
	Population, male	-0.006	4.003×10 ⁻⁴	-16.811	-16.044	< .001
	Population, female	0.007	3.976×10 ⁻⁴	17.250	16.943	< .001

Source: World Bank, World Development Indicators



Linear Regression

Y=Female mortality rate

X=Fertility rate controlled by urbanization rate(1960-2024)

Model Summary - Mortality rate, adult, female (per 1,000 female adults)

Model	R	R ²	Adjusted R ²	RMSE	Durbin-Watson		
					Autocorrelation	Statistic	p
M ₁	0.988	0.977	0.976	9.490	0.968	0.060	< .001

Note. M₁ includes Fertility rate, total (births per woman), Urban population (% of total population)

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
M ₁	(Intercept)	227.503	54.391		4.183	< .001
	Fertility rate, total (births per woman)	19.252	5.777	0.457	3.333	.001
	Urban population (% of total population)	-1.894	0.486	-0.534	-3.895	< .001

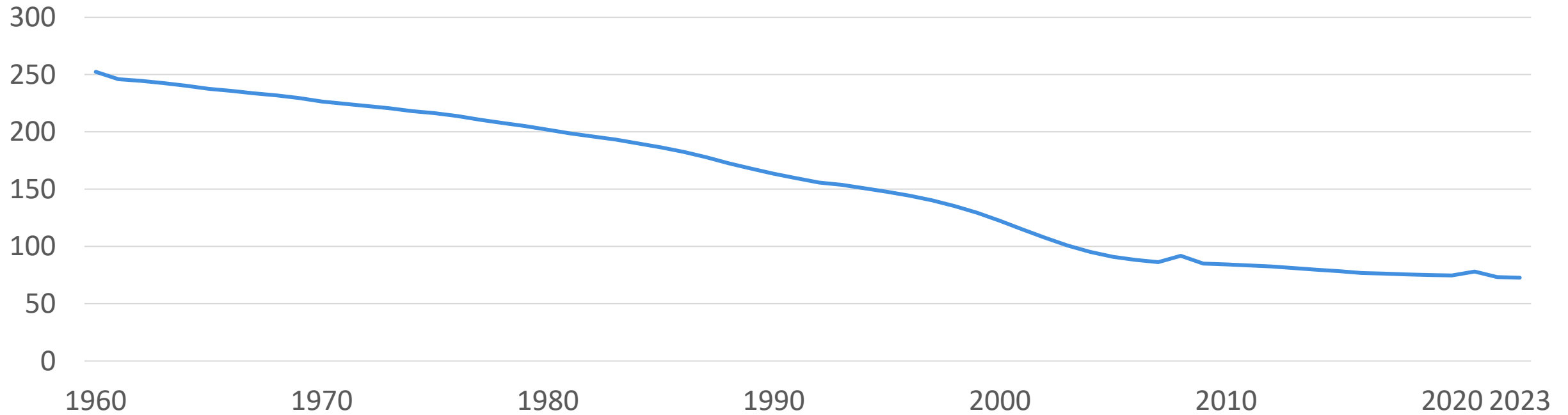
Source: World Bank, World Development Indicators



Autocorrelation is a common problem in time-series

Observations are “too related” across time

Mortality rate, adult, female (per 1,000 female adults)



Source: World Bank, World Development Indicators

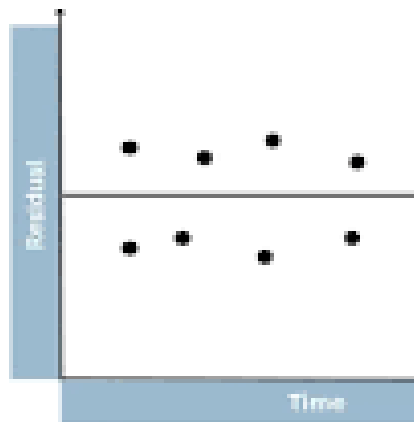


UNITED NATIONS

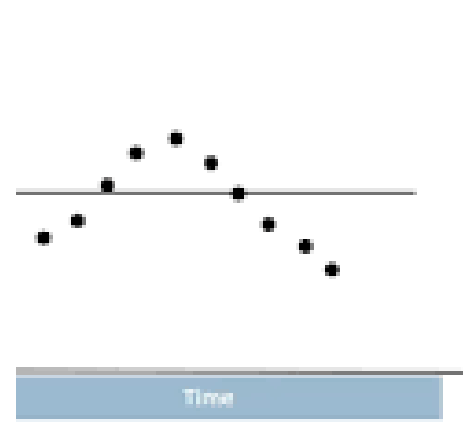
ECLAC

Autocorrelation, Residuals and Time

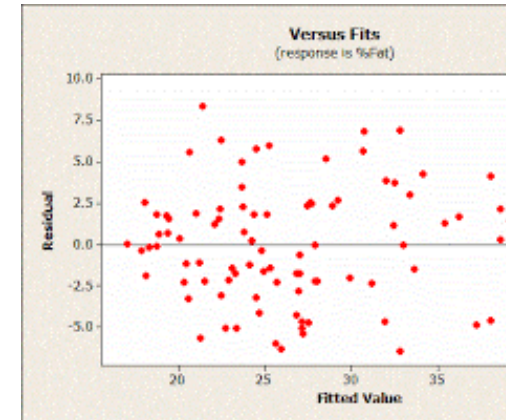
Negatively
Autocorrelated



Positively
Autocorrelated

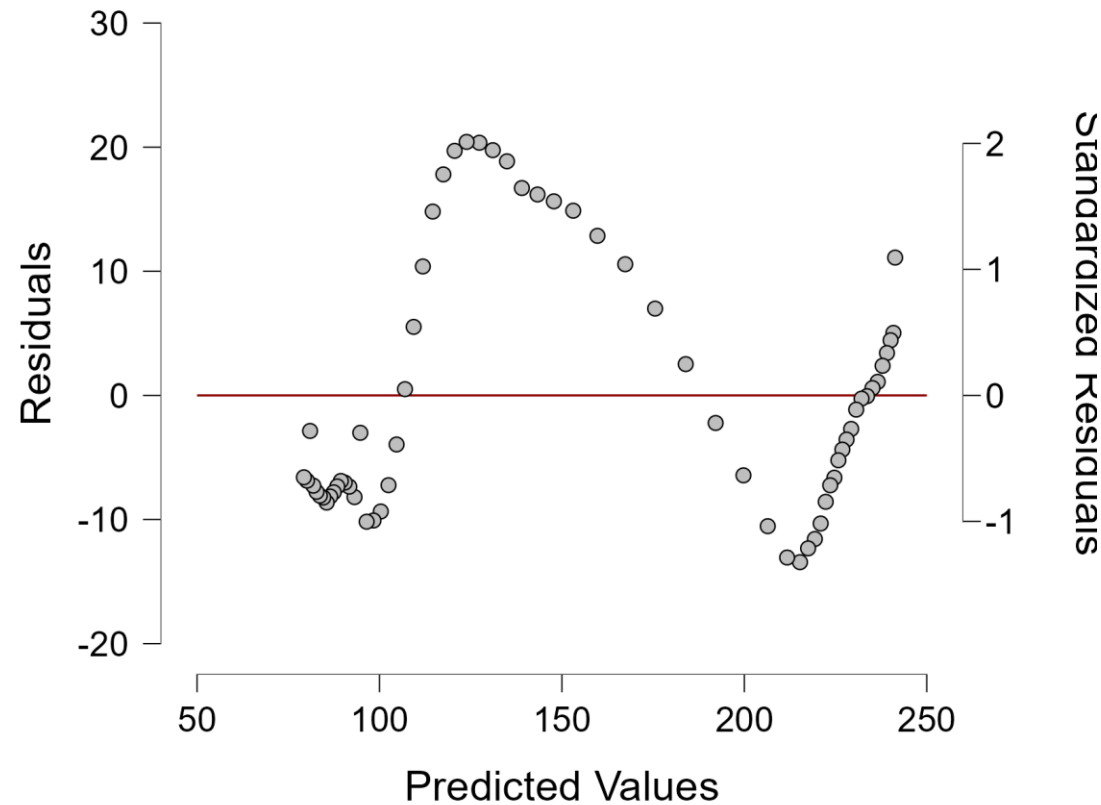


Not
Autocorrelated



Female mortality regression model:

Residuals vs. predicted values



Correcting for autocorrelation

Autocorrelation can be corrected. This requires the Cochrane-Orcutt procedure that transforms the variables and reestimates the model. However, not all statistical programmes support this. The ones that do: SPSS, Stata, and R

An ARIMA model (Autoregressive Integrated Moving Average) is also a possibility



UNITED NATIONS

ECLAC



Thank you!

Feel free to ask any questions



UNITED NATIONS

ECLAC

**Please complete the
workshop questionnaire
if you have not already
done so**

**It takes about 5-10
minutes to complete**

