Introducing Different Regression for Different Analysis

Dr. Wing Wahyu Winarno, MAFIS
STIE YKPN (YKPN School of Business) Yogyakarta
wing@stieykpn.ac.id

This presentation was delivered at the Konferensi Ilmiah Akuntansi XI and 1st International Conference held by Faculty of Economics and Business Atma Jaya Catholic University of Indonesia and Institute of Indonesia Chartered Accountant, Educator Accountant Compartment (IAI KAPd) in Jakarta, 7-8 March 2024.

Ordinary Least Squares (OLS) Regression

Regression analysis (especially multiple regression) is a popular statistical analysis tool among researchers. This analysis requires the presence of a dependent variable and several independent variables.

So far, what is widely known is OLS multiple regression with various assumptions that must be met. If the assumptions are not met, then the regression model is considered inappropriate.

But actually, there are different types of regression that can be used for various conditions. In this exposure will be shown different types of regression for various data conditions.

Ordinary Least Squares (OLS) Regression

Ordinary Least Squares (OLS) regression is a fundamental statistical method used for estimating the relationships between a dependent variable and one or more independent variables.

It is the most common method of linear regression, providing a way to model the linear relationship between the explanatory (independent) variables and the response (dependent) variable.

The primary objective of OLS regression is to find the best-fitting line through the data points that minimizes the sum of the squared differences (residuals) between the observed values and the values predicted by the linear model.
This line is known as the regression line.

Ordinary Least Squares (OLS) Regression

The regession line is represented by the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

where

- Y is the dependent variable
- $X_1, X_2, \ldots, X_k$ are the independent variables
- β_0 is the intercept
- $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients of the independent variables
- ε is the error term, representing the difference between the observed and predicted values

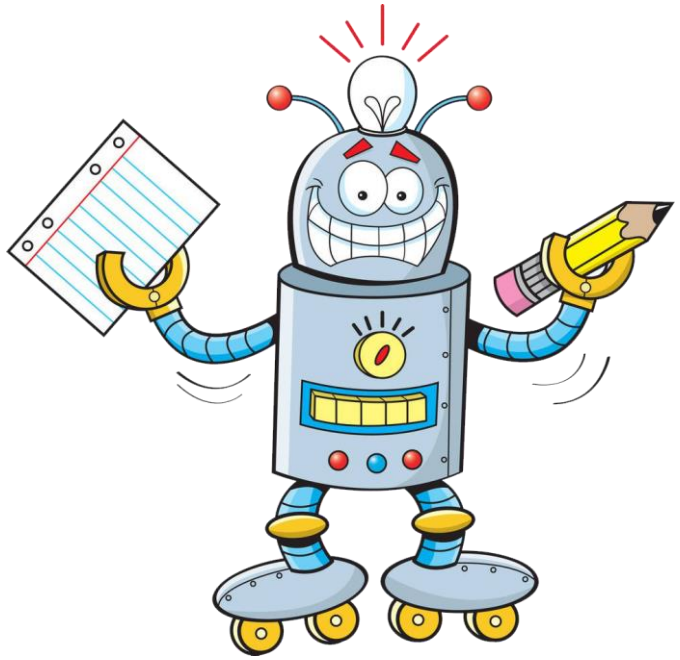
Ordinary Least Squares (OLS) Regression

It's crucial to assess the model's fit and check whether the assumptions of the OLS model are met. This involves:

- Analyzing residual plots to detect non-linearity, heteroscedasticity, and outliers.
- Conducting statistical tests for normality of residuals.
- Checking for multicollinearity through variance inflation factor (VIF) analysis.

Key assumption in Regression Models

Linearity	The relationship between the independent variables and the dependent variable is linear.
No perfect multicollinearity	The independent variables are not perfectly linearly related.
Autocorrelation (serial correlation)	Specifically refers to the correlation of a variable with itself across observations ordered in time or space.
Homoscedasticity	The variance of the residuals is constant across all levels of the independent variables.
Normality	The residuals are normally distributed (particularly important for hypothesis testing).
Independence	The residuals are independent across observations.



Thank you

Regression with Categorical Data

Logistic Regression	It estimates the probability that a given input (or set of inputs) belongs to a particular category (usually denoted as 1) versus the probability that it belongs to the other category (denoted as 0).
Probit Regression	Probit regression models the probability that an observation falls into one of two categories. It uses the cumulative distribution function of the standard normal distribution (the probit function) to model the probability.
Poisson Regression	Poisson regression can be adapted for dichotomous data, especially when dealing with rates or counts per unit of exposure. It models the log of the expected count as a linear combination of the independent variables.
Negative Binomial Regression	This is an extension of Poisson regression used for count data that exhibits overdispersion, where the variance exceeds the mean. It can also be adapted for dichotomous outcomes, particularly in counts with high variability.
Generalized Estimating Equations (GEE)	GEE is used for correlated or clustered categorical data, extending generalized linear models to accommodate correlated observations. It's particularly useful in longitudinal data analysis and the responses are binary.
Generalized Linear Models (GLMs)	GLMs are a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

Regression with Nonlinearity

Robust Regression	Robust regression methods are designed to be less sensitive to outliers than ordinary least squares (OLS) regression, thereby providing more reliable estimates when the data contain outliers or are not normally distributed. Also: Huber regression and Quantile regression
Quantile Regression	Unlike OLS regression, which estimates the mean of the dependent variable given the independent variables, quantile regression estimates the conditional median or other quantiles of the dependent variable.
Generalized Linear Models (GLMs)	GLMs extend linear regression to models where the dependent variable is not necessarily normally distributed. GLMs are useful when dealing with data that exhibit characteristics such as skewness, kurtosis, or heteroscedasticity.
Nonparametric Regression	Nonparametric regression, such as kernel smoothing and spline models, do not assume a specific functional form for the relationship between the IVs and DVs. These methods are flexible and can model complex relationships without assuming normality in the residuals.
Transformation of Variables	Transforming the dependent and/or independent variables can sometimes resolve issues of nonnormality. Common transformations include logarithmic, square root, and Box-Cox transformations.

Regression with Multicollinearity

Ridge Regression (L2 Regularization)	Ridge regression is particularly effective at dealing with multicollinearity by adding a penalty term to the loss function, which is proportional to the square of the coefficients. This helps to reduce the variance of the coefficients, leading to more stable estimates.
Lasso Regression (L1 Regularization)	Lasso regression also modifies the loss function to include a penalty term, but this time the penalty is the absolute value of the magnitude of coefficients. This can lead to some coefficients being shrunk to zero, effectively performing variable selection.
Elastic Net Regression	Elastic Net is a middle ground between Ridge and Lasso regression. It combines the penalties of Ridge and Lasso, thus benefiting from both the variable selection feature of Lasso and the ability to handle multicollinearity of Ridge.
Principal Component Regression (PCR)	PCR can be effective in handling multicol by reducing the dimensionality of the data before regression, using PCA. By regressing on the principal components instead of the original correlated variables, PCR mitigates the multicollinearity issue.
Partial Least Squares Regression (PLS)	PLS is similar to PCR in that it projects the predictors and the response variable into a new space. However, unlike PCR, which only considers the IVs, PLS takes into account the DV as well, aiming to find the multidimensional direction in the predictor space that explains the maximum multidimensional variance direction in the response space.

Regression with Autocorrelation

Autoregressive Integrated Moving Average (ARIMA)	ARIMA models are specifically designed for time series data to predict future points in the series. They incorporate three main components: autoregression (AR), differencing (I) to make the series stationary, and moving average (MA).
Generalized Least Squares (GLS)	GLS is an extension of OLS that allows for modeling of hetero directly through the specification of the covariance matrix of the errors. By assuming a particular form of hetero, GLS adjusts the estimation process to account for it, leading to more efficient estimates.
Durbin-Watson statistic	While not a regression method itself, the Durbin-Watson statistic is a widely used test for autocorrelation in the residuals of a regression analysis. It provides a measure of the extent of autocorrelation.
Cochrane-Orcutt or Prais-Winsten	These methods are modifications of the OLS to correct for autocorrelation, specifically for the first-order autocorrelation. They involve iteratively estimating the parameters of the regression model and the autocorrelation coefficient.
Vector Autoregression (VAR)	VAR is a system of equations model where all the variables are treated as endogenous. This method is particularly useful for multivariate time series data, where each variable is modeled as a function of past values of itself and past values of all the other variables.

Regression with Heteroscedasticity

Weighted Least Squares (WLS)	WLS is a variation of OLS that assigns weights to each data point based on the inverse of the variance of its error term. By doing so, WLS helps to ensure that observations with smaller variances have a larger influence on the estimation of the regression coefficients.
Generalized Least Squares (GLS)	GLS extends the WLS to more complex forms of hetero & also addresses issues of correlation between error terms. GLS uses a known covariance matrix of the error terms to transform the original equation into one where the transformed errors have a constant variance.
Robust Regression	This method is designed to be less sensitive to outliers and violations of assumptions like hetero. Techniques such as Huber regression and quantile regression provide alternative fitting procedures that are not as affected by the presence of heteroscedastic errors.
Quantile Regression	This regression models the relationship between the Ivs & specific quantiles (percentiles) of the DV, rather than the mean. This approach is inherently robust to outliers & does not assume homoscedasticity of errors, making it suitable for data with heteroscedasticity.
Heteroscedasticity-Consistent Standard Errors (HCSE)	HCSE (robust standard errors) is a common approach to deal with hetero. This technique adjusts the standard errors of the regression coefficients to reflect the presence of heteroscedasticity, allowing for more accurate hypothesis testing.
Transformation of Variables	Applying transformations to the DV and/or Ivs can sometimes reduce or eliminate hetero. Common transformations include logarithmic, square root, or Box-Cox transformation, which can stabilize the variance of the errors across levels of the IVs.

Regression with Nonnormality of Residuals

Polynomial Regression	Polynomial regression extends linear regression by considering polynomial terms of the IVs. By including squared, cubed, or higher-order terms of the predictors, polynomial regression can model curves in the data.
Generalized Additive Models (GAMs)	GAMs extend linear models by allowing non-linear functions of each of the Ivs while maintaining additivity. Unlike polynomial regression, GAMs use smooth functions, such as splines, to model the non-linearities.
Nonparametric Regression	Nonparametric regression methods do not assume a predefined form for the relationship between the IVs and DVs. These techniques estimate the relationship by closely following the observed data, making them highly flexible for modeling nonlinear trends.
Decision Trees and Random Forests	Decision trees model data by splitting it into subsets based on the value of input features, making them inherently capable of capturing non-linear relationships. Random forests improve upon single decision trees by averaging multiple trees to reduce overfittng.
Support Vector Regression (SVR)	SVR applies the principles of support vector machines (SVMs) to regression problems. By using kernel functions, SVR can model nonlinear relationships in a high-dimensional space where the data might be linearly separable.

Regression with Multicollinearity & Hetero

Ridge Regression (L2 Regularization)	Ridge regression is particularly effective at dealing with multicollinearity by adding a penalty term to the loss function, which is proportional to the square of the coefficients. This helps to reduce the variance of the coefficients, leading to more stable estimates.
Generalized Least Squares (GLS)	GLS is an extension of OLS that allows for modeling of hetero directly through the specification of the covariance matrix of the errors. By assuming a particular form of hetero, GLS adjusts the estimation process to account for it, leading to more efficient estimates.
Elastic Net Regression	Elastic Net combines the penalties of L1 (lasso) and L2 (ridge) regularization, making it capable of addressing multicol by shrinking correlated predictors towards each other. The L1 penalty can also help in variable selection, reducing the model complexity.
Weighted Least Squares (WLS)	WLS is a variation of OLS that assigns weights to each observation, which is particularly useful for handling hetero by giving less weight to observations with higher variance. When combined with techniques to address multicol ,WLS can effectively handle both issues.
Quantile Regression	Quantile regression estimates the conditional median or other quantiles of the dependent variable, rather than the mean. This approach is inherently robust to outliers and can be less sensitive to heteroscedasticity since it does not assume a constant variance of errors.
Principal Component Regression (PCR)	PCR can be effective in handling multicol by reducing the dimensionality of the data before regression, using PCA. By regressing on the principal components instead of the original correlated variables, PCR mitigates the multicollinearity issue.