Regression Model Summary Metrics

Metric: Definition, Importance

Dep. Variable: The variable that the model aims to predict., Identifies the focus of the regression analysis.

Model: Type of regression used., Determines the mathematical approach and assumptions (e.g., OLS for linear relationships).

Method: Statistical method used for estimation., Indicates the computational technique (usually "Least Squares").

No. Observations: Number of observations in the dataset., Reflects the sample size used for the analysis.

DF Residuals: Degrees of freedom of the residuals, calculated as number of observations less the number of parameters., Shows the number of data points that contributed to the fit of the model beyond the number required to estimate the model.

DF Model: Number of parameters in the model excluding the constant., Indicates the number of independent variables used.

R-squared: The proportion of variance in the dependent variable that is predictable from the independent variables., Higher values indicate a model that explains more of the variability around

the mean. A key measure of model fit.

Adj. R-squared: Adjusted R-squared, which considers the number of predictors and the complexity of the model., More reliable than R-squared for models with multiple predictors as it adjusts for the number of variables.

F-statistic: A measure of how well the model fits as a whole., A higher value indicates a better model fit. Tests whether at least one explanatory variable has a non-zero coefficient.

Prob (F-statistic): The probability that you would observe the given or an even more extreme value of F if the null hypothesis were true., Low values (typically <0.05) suggest that the model coefficients are statistically significantly different from zero.

Log-Likelihood: Logarithm of the likelihood function., Higher values generally indicate a better model fit to the data.

AIC: Akaike's Information Criterion, which penalizes more complex models., Helps in model selection. Lower AIC values indicate a better model, taking into account both the goodness of fit and the simplicity of the model.

BIC: Bayesian Information Criterion, similar to AIC but with a higher penalty for models with more parameters., Useful for model selection among a set of models. Lower BIC values indicate a better model.

coef: Estimated effect of each predictor., Represents the change in the dependent variable for a one unit change in an independent variable, holding other variables constant.

std err: Standard error of the estimated coefficient., Measures the accuracy of the coefficients' estimates. Smaller values indicate more stable estimates.

t: t-statistic for the coefficients., Used to determine the statistical significance of each coefficient.

P > |t|: P-value for the hypothesis test of the coefficients., A value less than 0.05 typically indicates statistical significance, suggesting the coefficient has a real effect assuming the model is correctly specified.

[0.025 0.975]: 95% confidence interval for the coefficients., Provides a range within which the true parameter value is expected to fall with 95% confidence.

Omnibus: Test for the skewness and kurtosis of the residuals., Tests normality of residuals. Non-significant values ($p > 0.05$) suggest residuals are normally distributed (a good sign).

Prob(Omnibus): P-value of the Omnibus test., Lower values indicate a departure from normality in the residuals.

Skew: Measure of data symmetry or asymmetry., Indicates the symmetry of the residuals distribution around their mean. Ideally close to zero.

Kurtosis: Measure of the tailedness of the data distribution., Higher values indicate more outliers. Normal distribution has a kurtosis of 3.

Durbin-Watson: Tests for autocorrelation in the residuals., Values around 2 suggest no

autocorrelation; values towards 0 indicate positive autocorrelation; values towards 4 indicate negative autocorrelation.

Jarque-Bera (JB): A goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution., Low JB value (and high p-value) supports the assumption that residuals are normally distributed.

Cond. No.: Condition Number, which measures the sensitivity of the function's output to its input., High values (>30) may indicate multicollinearity, suggesting that some independent variables are highly correlated and may not provide unique or reliable information.