01_OLS

QuantFit Estimator Standard Operating Procedure

SOP: Ordinary Least Squares (OLS)

The workhorse linear regression estimator

=> Use OLS as the baseline whenever residuals are well-behaved and X is exogenous.

1. Purpose

OLS estimates the linear conditional mean of Y given X by minimising the sum of squared residuals. It is BLUE (Best Linear Unbiased Estimator) under the Gauss-Markov assumptions: linearity in parameters, strict exogeneity, no perfect collinearity, homoskedasticity, and no autocorrelation of the error term.

2. When to use this estimator

Cross-sectional data with a single equation and exogenous regressors.

Time series where residuals are not autocorrelated and the data is stationary.

As the first benchmark before moving to richer estimators (FE, IV, ARDL).

When R2 and parsimony are the goal and residual diagnostics are clean.

3. Required data structure

Variables in numeric columns; categorical predictors must be dummy-encoded.

Single dependent variable Y; one or more X variables; no panel grouping required.

Sample size n >= k + 30 as a soft rule of thumb (k = number of regressors).

Missing data must be handled before estimation (listwise deletion or imputation).

4. Mathematical formulation

The model Y = Xbeta + epsilon is estimated by minimising epsilon'epsilon. The closed-form estimator and standard errors are:

beta? = (X'X)^-1 X'Y

Var(beta?) = sigma?2 (X'X)^-1

sigma?2 = epsilon?'epsilon? / (n - k)

t_j = beta?_j / sqrt(Var(beta?_j))

5. Pre-estimation diagnostics

Stationarity check (ADF/PP/KPSS) for time series - non-stationary regressors invalidate inference.

VIF < 5 for all regressors to rule out multicollinearity.

Examine the dependent variable’s distribution; log-transform if heavy-tailed.

Identify and address obvious outliers via winsorising or robust regression where needed.

6. Estimation procedure

Build the design matrix X (with intercept) and response vector Y.

Solve beta? = (X’X)^-1 X’Y via QR or Cholesky decomposition.

Compute fitted values ? = Xbeta? and residuals epsilon? = Y - ?.

Estimate residual variance sigma?2 and the variance-covariance matrix.

Derive standard errors, t-statistics, and p-values for each coefficient.

Report R2, adjusted R2, F-statistic, AIC, BIC, and Durbin-Watson.

7. Output produced

8. Output interpretation

beta?_j is the expected change in Y per unit change in X_j, holding others constant.

p < 0.05 => reject H0: beta_j = 0 at 5% level. Stars: *** 1%, ** 5%, * 10%.

R2 near 1 means the model explains most variation; near 0 means little.

F-statistic with p < 0.05 indicates at least one regressor is jointly significant.

Durbin-Watson far from 2 (< 1.5 or > 2.5) signals serial correlation - switch to HAC SE.

9. Post-estimation diagnostics

Heteroskedasticity (Breusch-Pagan, White) - if rejected, use HC0-HC3 robust SE.

Autocorrelation (Durbin-Watson, Breusch-Godfrey) - if rejected, use HAC SE or ARDL.

Normality of residuals (Jarque-Bera, Shapiro-Wilk) - if rejected for small n, treat inference cautiously.

Specification (RESET) - if rejected, consider non-linear terms.

Influence (Cook’s D, leverage) - investigate top-3 influential points.

10. Common pitfalls

Spurious regression: applying OLS to non-stationary data produces inflated R2 and t-stats.

Endogeneity: if X is correlated with epsilon, beta? is biased and inconsistent - use 2SLS / GMM.

Omitted variable bias: confounders soak up explanatory power and bias coefficients.

Multicollinearity inflates SEs without biasing beta? - but makes individual t-tests unreliable.

Autocorrelation produces under-stated SEs; reported significance is misleading.

11. Reporting checklist

Sample period and N reported.

Coefficient table with beta?, SE, t, p, stars, and 95% CI.

R2, adjusted R2, F-statistic, AIC, BIC, Durbin-Watson.

All Card A diagnostics (BG, BP, JB, RESET) with verdict.

Robust SE method named (classical / HC0 / HC1 / HC3 / HAC) with reason.

Plain-language interpretation paragraph naming the dependent variable.

12. References

Greene, W. H. (2018). Econometric Analysis, 8th ed. Pearson.

Wooldridge, J. M. (2020). Introductory Econometrics, 7th ed. Cengage.

Davidson, R., MacKinnon, J. G. (2004). Econometric Theory and Methods. Oxford.

Field | Meaning

coefficients | beta?_j keyed by variable name

standardErrors | SE(beta?_j) - classical or robust depending on SE type

tStatistics | beta?_j / SE(beta?_j)

pValues | Two-sided p-value of the t-test

rSquared / adjustedRSquared | Goodness of fit

fStatistic | Joint significance of all regressors

residuals / fitted | epsilon? and ? at each observation

durbinWatson | Autocorrelation diagnostic; DW ~ 2 indicates no AR(1)

aic / bic | Information criteria for model comparison