Regressions

Noé Debrois

June 16, 2024

Multivariate Regressions

$y_t = \alpha + \beta x_t + u_t$ (error), $t \in [\![1, T]\!]$.

- **Generalizing the Simple Model**: Extend the simple one-dimensional regression model to accommodate multiple independent variables. For example, the number of cars sold may depend on factors like car prices, public transport prices, petrol prices, and public concern about global warming.

rows. cl.
y is $T \times 1$
x is $T \times (r+1)$
β is $(r+1) \times 1$
u is $T \times 1$

- **Matrix Notation**: Represent the multiple regression model using vectorial notation: $y = X\beta + u$, where y is the dependent variable, X is the matrix of independent variables, β is the vector of coefficients, and u is the error term.

it comes from : $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \cdots + \beta_{r+1} x_{(r+1)t} + u_t$ (error), for $t \in [\![1, T]\!]$.

Multiple Regression and the Constant Term

- **Inclusion of the Constant Term**: The regression equation includes a constant term represented as a column of ones in the X matrix. The general form of the equation is $y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \cdots + \beta_{r+1} x_{r+1,t} + u_t$.

Ordinary Least Squares (OLS) Estimator

- **Parameter Estimation**: The OLS method is used to estimate the parameters β by minimizing the sum of squared residuals. The OLS estimator is BLUE (Best Linear Unbiased Estimator), if certain assumptions hold.

↳ the MLE

$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$

$\hat{u} = Y - X\beta$ in matrix form.
OLS: min. the RSS:
$S(\beta) = \sum_{t=1}^{T} (y_t - \beta_1 - \beta_2 x_{2t} - \cdots - \beta_{r+1} x_{r+1,t})^2$
$= \sum_{t=1}^{T} \hat{u}_t^2 = \hat{u}'\hat{u}$.

- **Assumptions of the CLRM (Classical Linear Regression Model)**:

1,2,3&5 ⇕ $u \sim \mathcal{N}(0, \sigma^2 I)$

1. The error term has zero mean $E(u_t) = 0$.
2. Homoscedasticity $Var(u_t) = \sigma^2$.
3. No autocorrelation $Cov(u_i, u_j) = 0$.
4. The independent variables are non-stochastic.
5. The error terms are normally distributed.

↳ x is non-stochastic.
↳ $u_t \sim \mathcal{N}(0, \sigma^2)$.

SEE THE PROPERTIES OF THE ESTIMATORS.

Goodness of Fit: R^2 and Adjusted R^2

- R^2: Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. R^2 values range from 0 to 1.

a measure of how well one regression model actually fits the data.

R^2: the square of the correlation between y & $\hat{y}$.

Another way to define R^2: TSS = ESS + RSS (Total SS = Expl. SS + RSS)
i.e $\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t \hat{u}_t^2$
$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \in [0,1]$.

- **Adjusted** R^2: Adjusts R^2 for the number of predictors in the model, providing a more accurate measure when comparing models with different numbers of independent variables.

→ A way to address the 2nd limitation.

Problems with R^2 as a Goodness of Fit Measure

- **Limitations**:
  1. R^2 changes if the dependent variable is reparameterized. (bc. it is defined in terms of variation about the mean.)
  2. R^2 never decreases with the addition of more regressors.
  3. High R^2 values are common in time series regressions, which can be misleading.

we can test e.g $H_0: \beta_2 = 0 \& \beta_3 = 0$.

The F-Test for Overall Significance

test statistic $= \frac{RRSS - URSS}{URSS} \times \frac{T-r-1}{r-q} \sim F_{r-q, T-r-1}$

RSS from restricted; RSS of unrestricted; values of the 2 d.f.

T: nb of observat°. q: nb of restrictions. r-q: nb. of regressors in unrestricted reg.

- **Testing Hypotheses**: The F-test compares the fit of a restricted model (with fewer predictors) to an unrestricted model. If the F-statistic is significantly large, we reject the null hypothesis that the restricted model is sufficient.

we can not test hypotheses which are NOT LINEAR OR which are MULTIPLICATIVE. E.g: $H_0: \beta_2 \beta_3 = 1$.

for which the coeff. are freely determined by the data.

Violation of CLRM Assumptions

- **Detection and Consequences**: Violations of assumptions can lead to biased or inefficient estimators. Various tests (like the F-test and Chi-square test) and methods (like transforming variables or adding more data) are used to detect and correct these issues.

Asymptotically, the 2 tests are equivalent.

For small samples F-test is preferable.

Multicollinearity → when explanatory variables are highly correlated w/ each other.

↳ e.g $x_3 = 2x_2$. → ONLY CORRELATION BETWEEN THE X's !

- **Solutions**: To address multicollinearity, one can drop collinear variables, transform them into ratios, or collect more data. Multicollinearity does not affect the goodness of fit but makes the estimation of individual coefficients unreliable.

→ If ignored: high R^2 but high standard errors. → Use PCA, Lasso, Ridge ...

Functional Form Misspecification

- **Ramsey's RESET Test**: This test checks if the functional form of the regression model is correct by adding higher-order terms of the fitted values. If the test indicates misspecification, transforming the data (e.g., using logarithms) might help.

i.e $\hat{y}_t^2, \hat{y}_t^3 \dots$

⚠ we have assumed that the appropriate functional form was LINEAR.

Parameter Stability → we assumed that the β_i are constant for the entire sample period.

- **Testing Stability**: Use parameter stability tests, like the Chow test, to check if regression coefficients are stable over time. This involves splitting the data into sub-periods and comparing the models' residual sum of squares (RSS).

→ or Predictive Failure Test.

Assumption 1: $\mathbb{E}[v_t] = 0$. → OK as long as there is a constant term in the regression.

Assumption 2: Homoskedasticity. → To detect heteroskedasticity: WHITE'S TEST (few assumptions).
2 regressions → classical. → regress the error.

If Ass.2 is not valid: OLS still unbiased but Not BLUE.

How to deal with it?
(III) Transformation of variables into logs.
(II) White's correction.
(I) If we know the form of heteroskad. (e.g $Var(u_t) = \sigma^2 z_t^2$), we can use GLS: e.g $\frac{y_t}{z_t} = \beta_1 \frac{1}{z_t} + \beta_2 \frac{x_{2t}}{z_t} + \beta_3 \frac{x_{3t}}{z_t} + v_t$ where $Var(v_t) = \frac{Var(u_t)}{z_t^2} = \sigma^2$.

Assumption 3: $cov(u_i, u_j) = 0$ for $i \neq j$, i.e., "there is no pattern in the errors".

→ To detect autocorrelation: **Durbin-Watson (DW) test**: it assumes: $u_t = \rho u_{t-1} + v_t$ w/ $v_t \sim N(0, \sigma_v^2)$ i.e 1st order autocorr.

It test $H_0: \rho = 0$. The test stastistic: $DW = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=2}^{T} \hat{u}_t^2} \approx 2(1-\hat{\rho}) \in [0,4]$. ← estimated correlation. ($\hat{\rho} \in [-1,1]$).

If DW is near 2: don't reject H_0.

⚠ Intermediate region where we can neither reject nor not reject H_0.

RULE OF THUMB ↔ DW.

↳ Other test: Breusch-Godfrey Test: r^{th} order autocorr.

- If ass.3. is not valid: OLS still unbiased but NOT BLUE.
- How to deal with it? GLS or Robust or dynamical models (ARMA, VAR).

Assumption 5: normality.

→ Jarque-Bera normality test: a normal distribution is NOT SKEWED (3rd moment.) & has a KURTOSIS = 3 (4th moments.).

What should we do if not normal? Not obvious... use methods w/o this ass.; or use dummy variable (e.g for oct. 1987).

- See the method for building a CLRM model, slides 75-77.
- ARTICLE "Determinants of Sovereign Credit Ratings", Cantor & Packer (1996).

↳ CONCLUSIONS:
- 6 factors play a big role in determining sovereign credit ratings: incomes, GDP growth, inflation, external debt, industrialized or not, & default history.
- The ratings provide more information on yields than all of the macro factors put together.
- We cannot determine well what factors influence how the markets will react to ratings announcements.

COMMENTS ON THE PAPER:
- Only 79 observations for rating announcements.
- Little attempt @ diagnostic checking.
- Where did the factors (explanatory variables) come from?