"MULTIPLE REGRESSION AND ISSUES IN REGRESSION ANALYSIS"

MSR = Mean Regression Sum of Squares
MSE = Mean Squared Error
RSS = Regression Sum of Squares
SSE = Sum of Squared Errors/Residuals
α = Level of Significance

F_C = Critical F taken from F Distribute Table
H_0 = Null Hypothesis
$H_\propto$ = Alternative Hypothesis
X = Independent Variable
Y = Dependent Variable
F = F Statistic (calculated)

1. INTRODUCTION

- Multiple linear regression models are more sophisticated.
- They incorporate more than one independent variable.

2. MULTIPLE LINEAR REGRESSIONS

- Allows determining effects of more than one independent variable on a particular dependent variable
- $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \cdots b_nX_n + E_i$
- Tells the impact on Y by changing X_1 by 1 unit keeping other independent variables same.
- Individual slope coefficients (e.g. b_1) in multiple regressions known as partial regression/slope coefficients.

2.1 Assumption of the Multiple Linear Regression Model

- Relationship b/w Y and $X_1, X_2, X_3, \ldots X_n$ is linear.
- Independent variables are not random and no exact linear relationship exists b/w 2 or more independent variables.
- Expected value of error terms is 0.
- Variance of error term is same for all observations.
- Error term is uncorrelated across observations.
- Error term is normally distributed.

2.2 Predicting the Dependent Variable in a Multiple Regression Model

- Obtain estimates of regression parameters.
  - $estimates = \hat{b_0}, \hat{b_1}, \hat{b_0}, \ldots \hat{b_n}$
  - $regression\ parameters = b_0, b_1, b_1, \ldots b_k$
- Determine assumed values of $\hat{X}_{1i}, \hat{X}_{2i} \ldots \hat{X}_k$
- Compute predicted value of $\hat{Y}$ using $\hat{Y}_i = \hat{b}_0 + \hat{b}_1\hat{X}_{1i} + \hat{b}_2\hat{X}_{2i} + \cdots + \hat{b}_k\hat{X}_{ki}$
- To predict dependent variable:
  - Be confident that assumptions of the regression are met.
  - Predictions regarding X must be within reliable range of data used to estimate the model.

2.3 Testing Whether All Population Regression Coefficients Equals Zero

- $H_0 \Rightarrow$ All slope coefficients are simultaneously = 0, none of the X variable helps explain Y.
- To test H_0 F-test is used.
- T-test cannot be used.

$$F = \frac{MSR}{MSE} = \frac{RSS/k}{SSE/(n-(k+1))}$$

2.3 Testing Whether All Population Regression Coefficients Equals Zero

- Where

$$RSS = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$$

$$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

  n = no. of observation
  k = no. of slope coefficients
- Decision rule ⇒ reject H_0 if F > F_C (for given α).
- It is a one-tailed test.
- d_f numerator =k
- d_f denominator =n-(k+1).
- For k and n the test statistic representing H_0, all slope coefficients are equal to 0, is $F_{k,n-(k+1)}$
- In F-distribution table f_0 $F_{k,n-(k+1)}$ where K represents column and n-(k+1) represents row.
- Significance of F in ANOVA table represents 'p value'.
- ↑ F-statistic ↓ chances of Type I error.

2.4 Adjusted R^2

- R^2 ↑ with addition of independent variables (X) in regression
- $Adjusted\ R^2\ \left(\overline{R}^2\right) = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2)$.
- When k ≥ 1 ⇒ $R^2 > \overline{R}^2$
- $\overline{R}^2$ can be –ve but R^2 is always +ve.
- If $\overline{R}^2$ is used for comparing regression models.
  - Sample size must be the same
  - Dependent variable is defined in the same way.
- ↑$\overline{R}^2$ Does not necessarily indicate regression is well specified.

3. USING DUMMY VARIABLES IN REGRESSION

- Dummy variable ⇒ takes 1 if particular condition is true & 0 when it is false.
- Diligence is required in choosing no. of dummy variables.
- Usually n-1 dummy variables are used where n= no. of categories.

4. VIOLATIONS OF REGRESSION ASSUMPTIONS

4.1 Heteroskedasticity

- Variance of errors differs across observations ⇒ heteroskedastic
- Variance of errors is similar across observations ⇒ homoskedastic
- Usually no systematic relationship exists b/w X & regression residuals.
- If systematic relationship is present ⇒ heteroskedasticity can exist.

4.1.1 The Consequence of Heteroskedasticity

- It can lead to mistake in inference.
  - Does not affect consistency.
- F-test becomes unreliable.
- Due to biased estimators of standard errors, t-test also becomes unreliable.
- Most likely result of heteroskedasticity is that the:
  - estimated standard errors will be underestimated.
  - t-statistic will be inflated.
- Ignoring heteroskedasticity leads to significant relationship that does not exist actually.
  - It becomes more serious while developing investment strategy using regression analysis.
- Unconditional heteroskedasticity ⇒ when heteroskedasticity of error variance is not correlated with independent variables in the multiple regression.
  - Create major problems for statistical inference.
- Conditional heteroskedasticity ⇒ when heteroskedasticity of error variance is correlated with the independent variables.
  - It causes most problems.
  - Can be tested & corrected easily through many statistically software packages.

4.1.2 Testing for Heteroskedasticity

- Breush-Pagan test is widely used.
- Regression squared residuals of regression on independent variables.
  - Independent variables explain much of the variation of errors ⇒ conditional heteroskedasticity exists.
- H_0 = no conditional heteroskedasticity exists.
- H_α = conditional heteroskedasticity exist
- Under Breush-pagan test statistic = nR^2
  - R^2: from regression of squared residuals on X
- Critical value ⇒ calculated χ^2 distribution.
  - d_f = no. of independent variables
- Reject H_0 if test-static > critical value.

4.1.3 Correcting for Heteroskedasticity

Robust Standard Errors

- Corrects standard error of estimated coefficients.
- Also known as heteroskedasticity consistent standards errors or white-corrected standards errors.

Generalized Least Squares

- Modify original equation.
- Requires economic expertise to implement correctly on financial data.

4.2 Serial Correlation

- Regression errors correlated across observations.
- Usually arises in time-series regression.

4.2.1 The Consequences of Serial Correlation

- Incorrect estimate of regression coefficient standard errors
- Parameter estimates become inconsistent & invalid when Y is lagged onto X under serial correlation.
- Positive serial correlation ⇒ positive (negative) errors ↑ chance of positive (negative) errors
- Negative serial correlation ⇒ positive (negative) errors ↑ chance of negative (positive) errors
- It leads to wrong inferences
- If positive serial correlation:
  - Standard errors underestimated
  - T-statistic & F-statistics inflated
  - Type-I error ↑
- If negative serial correlation
  - Standard errors overestimated
  - T-statistics & F-statistics understated
  - Type-II error ↑

4.2.2 Testing for Serial Correlation

- Variety of tests, most common → Durbin-Watson test
- $DW = \frac{\sum_{t=2}^{T}(\hat{E}_t - \hat{E}_{t-1})^2}{\sum_{t=1}^{T}\hat{E}_t^2}$

  Where $\hat{E}_t$= regression residual for period t.
- For large sample size Durbin-Watson statistic (d) is approximately

  →DW ≈ 2(1-r)

  →where r = sample correlation b/w regression residuals of t and t-1
- Values of DW can range from 0 to 4.
- DW = 2 ⇒ r=0 ⇒ no serial correlation.
- DW = 0 ⇒ r=1 ⇒ perfectly positively serially correlated.
- DW = 4 ⇒ r = -1 ⇒ perfectly negatively serially correlated.
- For positive serial correlation:
  - H_0 ⇒ No positive serial correlation
  - H_α ⇒ Positive serial correlation
  - $DW < dl$ ⇒ reject H_0
  - $DW > du$ ⇒ do not reject H_0
  - dl ≤ $DW \leq du$ ⇒ inconclusive.

4.2.2 Testing for Serial Correlation

- For negative serial correlation:
  - $H_0 \Rightarrow$ No negative serial correlation.
  - $H_\alpha \Rightarrow$ Negative serial correlation.
  - $DW > 4 - dl \Rightarrow$ Reject H_0.
  - $DW < 4 - du \Rightarrow$ do not reject H_0
  - $4 - du \leq DW \leq 4 - dl \Rightarrow$ inconclusive.

4.2.3 Correcting for Serial Correlation

- Adjust the coefficient standard errors.
  → Recommended method
- Hansen's method ⇒ most prevalent one.

- Modify regression equation.
- Extreme care is required.
- May lead to inconsistent parameters estimates.

4.3 Multicollinearity

- Occurs when two or more independent variables (X) are highly correlated with each other.
- Regression can be estimated but result becomes problematic.
- Serious practical concern due to commonly found approximate linear relation among financial variables.

4.3.1 The Consequences of Multicollinearity

- Difficulty in detecting significant relationships.
- Estimates become extremely imprecise & unreliable though consistency is unaffected.
- F-statistic is unaffected.
- Standard errors of regression can ↑.
  - Causing insignificant t-tests
  - Wide confidence interval
  - Type II error↑

4.3.2 Detecting Multicollinearity

- Multicollinearity is a matter of degree rather than the presence / absence.
- ↑ Pair wise correlation does not necessarily indicate presence of Multicollinearity
- ↓ Pair wise correlation does not necessarily indicate absence of Multicollinearity
- With 2 independent variables ⇒ correlation is a useful indicator.
- ↑ R^2 significant, F-statistic significant, insignificant t-statistic on slope coefficients ⇒ classic symptom of Multicollinearity

4.3.3 Correcting Multicollinearity

- Exclude one or more regression variables.
- In many cases, experimentation is done to determine variable causing Multicollinearity

5. MODEL SPECIFICATION AND ERRORS IN SPECIFICATION

- Model specification ⇒ set of variables included in regression.
- Incorrect specification leads to biased & inconsistent parameters

5.1 Principles of Model Specification

- Model grounded on economic reasoning.
- Functional form of variables compatible with nature of variables
- Parsimonious ⇒ each included variable should play an essential role
- Model is examined for the violation of regression assumptions.
- Model is tested for the validity & usefulness of the out of sample data.

5.2 Misspecified Functional Form

- One or more variables are omitted. If omitted variable is correlated with remaining variable, error term will also be correlated with the latter and the:
  - result can be biased & inconsistent.
  - estimated standard errors of the coefficients will be inconsistent.
- One or more variables may require transformation.
- Pooling of data from different samples that should not be pooled.
  - Can lead to spurious results.

5.3 Times-Series Misspecification (Independent Variables Correlated with Errors)

- Including lagged variables (dependent) as independent with serial correlation.
- Including a function of the dependent variable as an independent variable.
- Independent variables measured with error

5.4 Other Types of Time-Series Misspecification

- Nonstationarity: variable properties, e.g. mean, are not constant through time.
- In practice nonstationarity is a serious problem.

6. MODELS WITH QUALITATIVE DEPENDENT VARIABLES

- Qualitative dependent variables ⇒ dummy variables used as dependent instead of independent.
- Probit model⇒ based on normal distribution estimates the probability:
  - of discrete outcome, given values of independent variables used to explain that outcome.
  - that Y=1, implying a condition is met.
- Logit model:
  - Identical to Probit model.
  - Based on logistic distribution.
- Both Logit and Probit models must be estimated using maximum likelihood methods.
- Discriminate analysis ⇒ can be used to create an overall score that is used for classification.
- Qualitative dependent variable models can be used for portfolio management and business management.