Week 8: Regression 2

Multivariate Regression

Dr Giuseppe Brandi

Northeastern University London

Outline

Introduction to Multivariate Regression

- Regression analysis studies relationships between dependent and independent variables.
- In multivariate regression, there are two or more independent variables.
- Extends simple regression to better capture the effects of multiple factors.
- Common uses: Prediction, explanation, and theory testing.

Why Use Multivariate Regression?

- More realistic models: Real-world outcomes are influenced by multiple factors.
- Better explanatory power: Increases the percentage of variance explained by the model.
- Avoids omitted variable bias: By including multiple predictors, we reduce the likelihood of missing key factors.

The Regression Model

- The general form:

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_k x_k + \epsilon$$

- y: Dependent variable
- $x_1, x_2, \ldots, x_k$: Independent variables
- $w_1, w_2, \ldots, w_k$: Coefficients representing the relationship between each x and y.

Simple vs. Multiple Regression

Simple Regression:

- One dependent variable y predicted from one independent variable x.
- Single regression coefficient.
- R^2: proportion of variation in y predictable from x.

Multiple Regression:

- One dependent variable y predicted from multiple independent variables $x_1, x_2, \ldots, x_k$.
- One regression coefficient for each independent variable.
- R^2: proportion of variation in y predictable by set of independent variables x's.

Interpreting Coefficients

- w_0 is the intercept: The expected value of y when all x's are zero.
- Each w_i represents the change in y for a one-unit change in x_i, holding all other variables constant.
- Example: If $w_1 = 2$, then a one-unit increase in x_1 leads to a 2-unit increase in y, assuming all other x's are constant.

Assumptions of Multivariate Regression

- Independence: The observations are independent of each other.
- Linearity: The relationship between each independent variable and the dependent variable is linear.
- Homoscedasticity: Constant variance of errors.
- No multicollinearity: Independent variables should not be highly correlated.
- Normality of residuals: Residuals (errors) are normally distributed.

Violation of Assumptions

- Independence: Can lead to autocorrelation, often seen in time series data.
- Non-linearity: Consider transformations or polynomial terms.
- Heteroscedasticity: May need to apply weighted least squares.
- Multicollinearity: Addressed using variance inflation factor (VIF) or removing correlated variables.

Fitting the Model: Ordinary Least Squares (OLS)

- OLS is used to estimate the coefficients of the regression model.
- Goal: Minimize the sum of squared errors (SSE) between the observed and predicted values of y.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Evaluating the Model: R^2 and Adjusted R^2

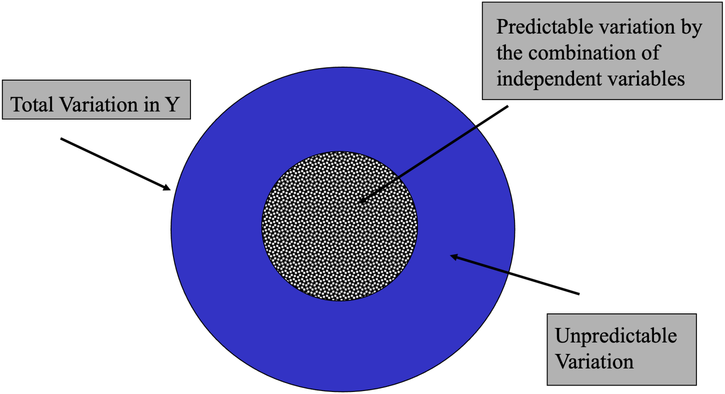
- R^2: The proportion of the variance in the dependent variable that is explained by the independent variables.

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

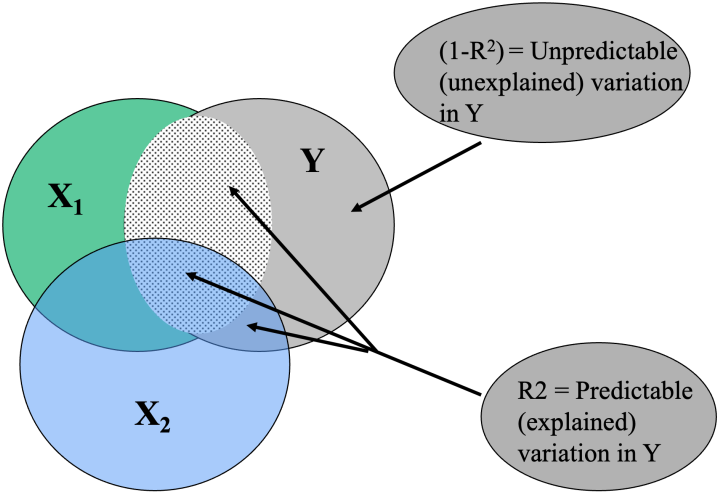
- Adjusted R^2: Adjusts for the number of predictors in the model.
- Formula:

$$R^2_{\text{adj}} = 1 - \left(\frac{SSE/(n-k-1)}{SST/(n-1)} \right)$$

Predictable vs Unpredictable Variation



Predictable vs Unpredictable Variation



Example: Self-Concept and Academic Achievement

Examining the relation between academic achievement (AA), grades, general self-concept (GSC), and academic self-concept (ASC):

- **General self-concept (GSC)**: Perception of self across various areas.
- **Academic self-concept (ASC)**: Perception of self specifically in academic contexts.
- Hypothesis: AA and ASC are more closely related than AA and GSC. Grades may best be predicted by AA and GSC, not ASC.

Example: Academic Achievement Prediction

- Dependent variable: Academic achievement (AA)
- Independent variables: Academic self-concept (ASC), General self-concept (GSC)
- Model:

$$\widehat{AA} = 36.83 + 3.52x_{ASC} - 0.44x_{GSC}$$

- Interpretation:
  - $w_{ASC} = 3.52$: Each unit increase in ASC predicts a 3.52-unit increase in AA.
  - $w_{GSC} = -0.44$: GSC has a negative effect on AA.

Example Model

Predicting AA:

$$\widehat{AA} = \widehat{w_0} + \widehat{w_1} x_{\text{ASC}} + \widehat{w_2} x_{\text{GSC}}$$

Example prediction:

$$\widehat{AA} = 36.83 + (3.52)(6) + (-0.44)(4) = 56.23$$

What is Multicollinearity?

- Occurs when two or more independent variables are highly correlated.
- Leads to unreliable estimates of regression coefficients.
- Makes it difficult to determine the individual effect of each independent variable.

Detecting Multicollinearity

- Variance Inflation Factor (VIF): A measure of how much the variance of a coefficient is inflated due to multicollinearity.

$$VIF = \frac{1}{1 - R^2}$$

- A VIF above 10 indicates significant multicollinearity.
- Pairwise correlation matrix: High correlations between independent variables suggest multicollinearity.

Handling Multicollinearity

- Drop one of the correlated variables.
- Combine variables into a single index (e.g., sum or average of highly correlated variables).
- Use regularization methods like Ridge Regression or LASSO.

Significance Tests

Testing R^2

- Test R^2 through an F-test.
- F-test: Testing the significance of the model as a whole.

Testing w's

- Test each partial regression coefficient w using t-tests.
- t-test: Testing the significance of each individual predictor.

Testing the Overall Model: F-Test

- The F-test evaluates whether the independent variables as a group explain a significant portion of the variation in y.
- Formula:

$$F = \frac{(R^2/k)}{(1 - R^2)/(n - k - 1)}$$

- If the calculated F-statistic is greater than the critical value from the F-distribution table, we reject the null hypothesis.
- Alternatively, if the p-value associated with the F-statistic is lower than a onfidence level α, we reject the null hypothesis.

Testing Individual Coefficients: t-Test

- The t-test is used to test whether each coefficient is significantly different from zero.
- Null hypothesis: $w_i = 0$
- Formula for the t-statistic:

$$t = \frac{w_i}{\mathrm{SE}(w_i)}$$

- If the t-statistic is greater than the critical value, we reject the null hypothesis.
- Alternatively, if the p-value associated with the t-statistic is lower than a confidence level α, we reject the null hypothesis.

Interaction Effects

- Interaction occurs when the effect of one independent variable depends on the level of another independent variable.
- Interaction terms can be added to the model as products of independent variables:

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 (x_1 \cdot x_2) + \epsilon$$

- Example: Does the effect of study time on grades depend on pre-test scores?

Polynomial Regression

- Non-linear relationships between independent and dependent variables can be modeled using polynomial terms:

$$y = w_0 + w_1 x + w_2 x^2 + \cdots + w_k x^k + \epsilon$$

- Example: Modeling the effect of age (quadratic relationship) on income.

Regularization: Ridge and LASSO

- Regularization techniques are used to prevent overfitting by adding penalties for large coefficients.
- Ridge Regression: Adds a penalty proportional to the sum of squared coefficients:

$$E(\mathrm{w}) = \frac{1}{2}\sum_{i=1}^{n}(y-\hat{y})^2 + \frac{\alpha}{2}\sum_{j=1}^{m}{w_j}^2$$

- LASSO (Least Absolute Shrinkage and Selection Operator): Adds a penalty proportional to the sum of absolute values of the coefficients:

$$E(\mathrm{w}) = \frac{1}{2}\sum_{i=1}^{n}(y-\hat{y})^2 + \frac{\alpha}{2}\sum_{j=1}^{m}|w_j|$$

Parameter α controls the **strength of the penalty**.

Model Selection: Stepwise Regression

- Stepwise regression adds or removes predictors based on their statistical significance.
- Can be forward (start with no predictors) or backward (start with all predictors).
- Useful for selecting the best combination of variables in large datasets.

Conclusion

- Multivariate regression allows for the inclusion of multiple predictors, improving the accuracy of models.
- Model assumptions must be checked to ensure validity.
- Tools such as the F-test, t-test, and R^2 help assess the fit and significance of the model.
- Beware of multicollinearity and choose the best model using appropriate selection methods.