Inference on the Linear Model; Interpreting the Model

Overview

Testing the Multiple Regression Model

Testing for the Model Utility - ANOVA

1. Check Assumptions: Linearity? Independence? nearly normal? Equal Variance? Normal Population?
2. State Hypotheses:
   $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (model is useless)
   H_A : At least one $\beta_i \neq 0$ (model is of some use)
3. Compute Test statistics: $F = \frac{SSReg/k}{SSResid/(n-k-1)}$
4. Compute p-value. Make a decision based on a pre-specified α. [Note: if p-value$< \alpha$ reject H_0, else do not reject.]
5. Conclude

A Confidence Interval for the Regression Slopes

$$b_i \pm t * SE(\beta_i)$$

where

- t is the critical value of the t-distribution with df $= n - p - 1$

Assumptions

- Linearity Assumption
- Independence Assumption: Look for randomization in the sample or the experiment (the Randomization Condition). Also check the residual plot for lack of patterns or clumping.
- Equal Variance Assumption: the variability of y should be about the same for all values of x. Make sure the spread around the line is nearly consistent.
- Normal Population Assumption: Check if the residuals satisfy the Nearly Normal Condition. Also, still check the Outliers.

t-Test for Individual Coefficients

1. Check Assumptions: Linearity? Independence? nearly normal? Equal Variance? Normal Population?
2. State Hypotheses:
   $H_0 : \beta_i = 0$ (this variable is not needed in model, once others are there already)
   $H_A : \beta_i \neq 0$ (this variable is of some use)
3. Compute Test statistics: $t = \frac{b_i}{SE(b_i)}$
4. Compute p-value. Make a decision based on a pre-specified α. [Note: if p-value$< \alpha$ reject H_0, else do not reject.]
5. Conclude

Final Note

Developing multiple regression models is an inexact science. Some tools that we use to help decide whether one model is better than another include:

- Checking Assumptions: Look at Residual Plots
- Looking at R^2 and adjusted R^2 values
- Overall F-test for model utility
- Individual t-tests for model coefficients
- Parsimony: Simpler models are generally preferred to more complicated ones.

Interpreting the Model Output

Interpreting Model Coefficients

Partial effect plots, also known as partial regression plots or added variable plots, show the relationship between predictor and response variables while adjusting for interference from other predictor variables.

- Partial effect plots are useful for understanding the individual contributions
- They can help with detecting influential points and outliers
- They help with interpreting the slope coefficients.

Interpretations

Interpreting Regression Results

- Regression Coefficients: Partial effect plots
- R^2 and Adjusted R^2
- Residual standard error

Confidence and Prediction Intervals

- The predicted value is our best guess, but now that we have standard errors, we can construct confidence intervals for predictions.
- There are two different questions we can address with confidence intervals in this setting:
  1. For a given x value, we can give a **confidence interval for the predicted mean value of y.**
  2. For a given x value, we can give a **prediction interval for a single value of y.**

Confidence and Prediction Intervals

- Confidence interval of prediction:

$$\hat{y}_i \pm \sqrt{MSE(x_i^t(\mathbf{X}^T\mathbf{X})^{-1})x_i)}$$

- Prediction interval:

$$\hat{y}_i \pm \sqrt{MSE(1 + x_i^t(\mathbf{X}^T\mathbf{X})^{-1})x_i)}$$

Model Complexity and Underfitting/Overfitting

Model Complexity and Underfitting/Overfitting

- Model complexity refers to the sophistication or flexibility of a model in capturing relationships between input features and the target variable. In simpler terms, it reflects how intricate the model is in representing the underlying patterns in the data.
- Underfitting/Overfitting: model complexity is a crucial consideration in machine learning because overly simple models may underfit the data, failing to capture important patterns, while overly complex models may overfit the data, capturing noise instead of true underlying relationships. Achieving the right balance of complexity is essential for building models that generalize well to unseen data.
- A solution is validating the model on Train-Test Split. Other techniques includes Bootstrapping, Cross-Validation, and Regularization.