Scaler Companion beta

Revision Notes: R Squared and Adjusted R Squared

Introduction

These notes cover the concepts of R Squared and Adjusted R Squared in the context of regression analysis, particularly focusing on linear regression. The lecture also delves into the assumptions of linear regression and provides some practical insights with examples.

1. R Squared (R^2)

- **Definition:** R Squared is a statistical measure that signifies the proportion of variance in the dependent variable that can be explained by the independent variables in the model 【4:3†source】.
- **Formula**: $R^2 = 1 - (RSS/TSS)$, where RSS is the Residual Sum of Squares and TSS is the Total Sum of Squares 【4:3†source】.
- **Interpretation**:
  - R^2 values range from 0 to 1.
  - An R^2 close to 1 indicates that a large proportion of the variance in the dependent variable is explained by the model.
  - R^2 can also be negative, indicating a worse model fit than a horizontal line through the mean of all data points. This implies that the model is not explaining the data better than the mean 【4:14†source】【4:7†source】.

2. Adjusted R Squared

- **Purpose**: Adjusted R Squared adjusts the R^2 value for the number of predictors in the model. It is used when comparing models with different numbers of predictors 【4:0†source】.
- **Behavior**:
  - Adding more variables to a model will always increase R^2, but Adjusted R^2 increases only if the added variables improve the model sufficiently 【4:16†source】.
  - It penalizes the model for including variables that do not improve the predicted value's explanatory power 【4:0†source】.

Scaler Companion beta

3. Assumptions of Linear Regression

The session also introduced the assumptions underlying linear regression. Only one was covered in detail due to a technical issue:

- **Linearity**: The relationship between the dependent variable and each independent variable is linear 【4:5†source】【4:6†source】.

4. Model Evaluation and Comparison

- **Training vs. Testing R Squared**:
  - The training R^2 provides insight into how well the model fits the data it was trained on.
  - The testing R^2 evaluates the model's performance on unseen data, often using a train-test split to achieve this 【4:12†source】【4:15†source】.
- **Impact of Irrelevant Features**: Adding random or irrelevant features can increase R^2 due to overfitting. However, Adjusted R^2 may decrease, indicating that these features do not provide meaningful information 【4:10†source】【4:16†source】.

5. Practical Example with Statsmodels

- **Installation and Setup**: Introduction to a Python library, Statsmodels, used for obtaining statistical details of a regression model 【4:17†source】.
- **Hands-on Scenario**: A practical demonstration involved creating and fitting a regression model using this library and observing changes in R^2 and Adjusted R^2 with the addition of random variables 【4:17†source】.

Examination Analogy

- **Analogy**: A training set is akin to a textbook that you study from, while a testing set is similar to an exam that evaluates how well you have understood the textbook. This highlights the importance of testing R^2 in assessing model performance 【4:15†source】.

Adjusted R^2 help in evaluating regression models, with a particular focus on practical applications and the common pitfalls of model evaluation.