REGRESSION MODELS REPORT

1. Linear Regression in Real Life

Article written by Carolina Bento, Real world problems solved with Math.
The article explains how linear regression can be used to estimate real-life costs, using the example of planning a road trip from San Francisco to Las Vegas. The key idea is to predict how much money will be spent on gas for the 1200-mile journey based on data collected about the car's fuel efficiency over time.

To solve this, first analyze the relationship between miles driven and gas expenditure using past data. By plotting this data, it becomes clear that there is a linear connection between these two variables. A linear regression model helps predict future costs based on this past data.

In this model, dependent variables are the costs of gas, and independent variables are the miles driven. The model is described as a simple linear equation:

y = Beta0 + Beta1x.

Here, Beta0 is the intercept (the starting point when no miles have been driven), and Beta1 is the coefficient that determines how the dependent variable changes as the independent variable changes.

The model's parameters (Beta0, Beta1) are estimated using Ordinary Least Squares (OLS), which minimizes the error between the predicted and actual data points. After the model is built, it can be used to estimate how much money should be budgeted for gas for the road trip.

Finally, the article concludes by emphasizing that linear regression is a useful tool for making predictions when there's a linear relationship between the variables involved.

2. Popular Applications of Linear Regression for Businesses

This article resumes linear regression in business from Jigsaw academy, the online school on analytics.

Linear Regression: A Powerful Statistical Technique

Linear regression is a statistical method that demonstrates how variations in independent variables influence changes in a dependent variable. The dependent variable (predictor or factor of interest) could be metrics like sales, pricing, performance, or risk. The independent variables (explanatory variables) help explain the factors that impact the dependent variable.

Applications:

- Provides insights into *consumer behavior*.
- Helps businesses understand *factors influencing profitability*.
- Assists in evaluating trends and *forecasting sales*.
- Analyzes the *effectiveness of marketing strategies*.
- Assesses financial and *insurance service risks*.

Features:

- *Limited applications*: Works only when the dependent variable is continuous.
- *Flexibility*: Non-linear relationships can be transformed to fit the linear regression model.

3. Good and Bad Regression Analysis

This article was written by Benjamin Obi Tayo Ph.D. Published in Towards AI, Feb 1, 2019. He shows that a good regression analysis requires careful attention to hyperparameters and model understanding to avoid poor results.

As mentioned, regression models are widely used in machine learning to predict continuous target variables. This article discusses good and bad practices in building regression models, using a simple linear regression model to predict housing prices from the Housing dataset.

Linear Regression with Gradient Descent: A simple linear regression model is implemented with gradient descent to predict an outcome variable (y) based on a predictor (X), demonstrated through Python code.

Application to Housing Data: The Housing dataset is preprocessed by selecting features and standardizing them. The model uses the strongest correlation between the number of rooms (RM) and median house value (MEDV) to predict house prices.

Hyperparameter Tuning: Different learning rate values (eta) are tested, with the best performance at eta = 0.0001, which produces the highest R-square value.

General Remarks and Conclusion: The article stresses the importance of selecting the right hyperparameters for regression models. Poor choices can lead to suboptimal models. Understanding model details and hyperparameter tuning is essential for achieving good performance. Using the model without understanding its intricacies can result in inaccurate predictions.

4. How to Choose the Best Regression Model

This article discusses how to Choose the Best Regression Model from Minitab Blog Editor on 2/28/2019. The author shows that selecting the best regression model involves combining statistical methods with domain knowledge. Theory, model simplicity, and residual analysis are crucial to achieving accurate, generalizable results.

Choosing the correct regression model can be challenging, especially when working with sample data. This post offers practical advice and statistical methods for selecting the best regression model, focusing on avoiding bias and ensuring precision.
Here are some of them:

Balance in Model Specification:

- **Too few predictors** can result in **biased** estimates (underspecified model).
- **Too many predictors** can lead to **less precise** estimates (overspecified model).
- The ideal model includes the **right number of predictors**, providing **unbiased** and **precise** estimates.

Statistical Methods for Model Selection:

- **Adjusted R-squared** and **Predicted R-squared** help identify models with **higher precision** and **avoid overfitting**.
- **P-values** indicate statistically **significant predictors**, guiding the r**emoval of non-significant terms** in the model.
- **Stepwise regression** and **Best subsets regression** assist in **identifying useful predictors**. Minitab also provides Mallows’ Cp to manage precision vs. bias.

Real-World Complications:

- **omitted variable bias**: Variables not included in the analysis may bias results.
- **Sampling issues**, like unusual data or data collection errors, can distort results.
- **Multicollinearity** complicates model interpretation by **reducing the significance of predictors**.
- **Data mining** can result in false findings of significance due to random correlations.
- **Model differences**: Different statistical methods might suggest varying models, and stepwise methods don't always identify the best model.

Recommendations for Finding the Best Model:

- **Theory**: Start with theoretical knowledge about predictors and their expected relationships. Adjust models based on theory, even if some predictors have insignificant p-values.
- **Complexity**: Simpler models generally offer more precise predictions. Avoid overly complex models that are tailored to the dataset and lack generalizability.
- **Residual Plots:** Check residual plots to identify model inadequacies. Patterns in residuals suggest the need for model adjustments, such as addressing curvature in underspecified models.