Assignment 4 - Multiple regression

INSTRUCTIONAL DETAILS

Multiple regression is a statistical technique used to model the relationship between multiple independent variables and a dependent variable. It is used to predict the value of the dependent variable based on the values of the independent variables.

Here is an example of how to perform multiple linear regression in Python using the scikit-learn library:

```
from sklearn.linear_model import LinearRegression
import numpy as np

# Assume we have a dataset with three independent variables and one dependent variable
X = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9], [10, 11, 12]])
y = np.array([1, 2, 3, 4])

# Create a LinearRegression model and fit it to the data
model = LinearRegression()
model.fit(X, y)

# Use the model to make predictions on new data
new_data = np.array([[13, 14, 15]])
prediction = model.predict(new_data)
print(prediction)
```

This will output the predicted value of the dependent variable for the given independent variables in new_data.

It's also possible to perform multiple regression with non-linear relationships between the independent and dependent variables. In this case, you can use a non-linear model such as a decision tree or a support vector machine.

Here is an example of how to perform multiple non-linear regression using a decision tree:

```
from sklearn.tree import DecisionTreeRegressor
import numpy as np

# Assume we have a dataset with three independent variables and one dependent variable
X = np.array([[1, 2, 3], [4, 5, 6], [7, 8, 9], [10, 11, 12]])
y = np.array([1, 2, 3, 4])

# Create a DecisionTreeRegressor model and fit it to the data
model = DecisionTreeRegressor()
model.fit(X, y)

# Use the model to make predictions on new data
new_data = np.array([[13, 14, 15]])
prediction = model.predict(new_data)
print(prediction)
```

This will output the predicted value of the dependent variable for the given independent variables in new_data.

In multiple regression, the goal is to model the relationship between a response variable (also known as the dependent variable) and one or more predictor variables (also known as the independent variables). The model is typically represented using a linear equation of the form:

$$Y = \beta0 + \beta1X1 + \beta2X2 + ... + \betapXp$$

where Y is the response variable, X1, X2, ..., Xp are the predictor variables, and β0, β1, β2, ..., βp are the coefficients (also known as the weights or parameters) of the model. The coefficients represent the strength and direction of the relationship between each predictor variable and the response variable.

To interpret the output of a multiple regression model, you will typically want to look at the values of the coefficients and the significance of each coefficient. The coefficient values tell you the strength and direction of the relationship between each predictor variable and the response variable. A positive coefficient indicates a positive relationship (i.e., as the value of the predictor variable increases, the value of the response variable also increases), while a negative coefficient indicates a negative relationship (i.e., as the value of the predictor variable increases, the value of the response variable decreases). The magnitude of the coefficient indicates the strength of the relationship; a large coefficient indicates a strong relationship, while a small coefficient indicates a weak relationship.

The significance of each coefficient can be determined using a hypothesis test, which allows you to determine whether the coefficient is significantly different from zero (i.e., whether there is a statistically significant relationship between the predictor variable and the response variable). If the p-value for a coefficient is less than a predetermined significance level (usually 0.05), then the coefficient is considered significant.

It is also important to consider the overall fit of the model, which can be evaluated using measures such as the R-squared value. The R-squared value indicates the percentage of the variance in the response variable that is explained by the model. A high R-squared value indicates a good fit, while a low R-squared value indicates a poor fit.

Other measures that can be used to evaluate the fit of a multiple regression model include the F-statistic and the adjusted R-squared value. The F-statistic is a measure of the overall significance of the model, while the adjusted R-squared value adjusts the R-squared value for the number of predictor variables in the model.

It is also important to consider the assumptions of the multiple regression model, including the linearity of the relationships between the predictor variables and the response variable, the independence of the errors (i.e., the residuals), and the homoscedasticity of the errors (i.e., the constant variance of the errors). Violations of these assumptions can affect the interpretability of the model and the validity of the statistical inferences drawn from the model.

BACKGROUND

Multiple regression is a statistical technique used to analyze the relationship between multiple independent variables and a single dependent variable. It is a useful tool for businesses because it allows them to understand how multiple factors contribute to a particular outcome or response.

For example, a business might use multiple regression to understand how factors such as marketing spend, product quality, and customer satisfaction contribute to overall sales. This can help the business identify which factors have the greatest impact on sales and inform decisions about how to allocate resources in order to maximize sales.

Multiple regression can also be used to predict future outcomes based on the relationships identified through the analysis. For example, a business might use multiple regression to predict future sales based on expected changes in marketing spend, product quality, and customer satisfaction.

Overall, multiple regression is an important tool for businesses because it helps them understand the relationships between different factors and how they contribute to a particular outcome. This can inform decision-making and help businesses achieve their goals.

RESEARCH QUESTION

As a call center sales business your goal is to optimize sales (revenue, $). Using previous months sales as a dependent variable (outcome variable), explore all possible combinations of independent variables to try to achieve the highest coefficient of determination (r^2, highest predictive ability against your dependent). Please make sure you test all of the fundamental assumptions and report any issues.

Data analysis can be a valuable tool in optimizing the operations of a call center. Here are a few ways in which data analysis can be used to improve the performance of a call center:

Analyzing call volume: Data analysis can be used to analyze call volume data to identify trends and patterns in the volume of calls received by the call center. This can help identify times of peak demand and allow the call center to adjust staffing levels accordingly.

Analyzing call data: Data analysis can be used to analyze data on the calls themselves, such as the length of the calls, the reason for the call, and the outcome of the call. This can help identify areas for improvement in the call center's processes and help identify training needs for staff.

Analyzing customer satisfaction: Data analysis can be used to analyze customer satisfaction data, such as survey results or ratings, to identify trends and patterns in customer satisfaction. This can help the call center identify areas for improvement and tailor its services to better meet the needs of its customers.

Analyzing staff performance: Data analysis can be used to analyze data on staff performance, such as call handling time, call resolution rate, and customer satisfaction ratings. This can help the call center identify areas for improvement in staff performance and provide targeted training and support.

REQUIREMENTS FOR SUBMISSION
See “Assignment 1 - Descriptives” for a detailed list of submission requirements.

FORMATTING
See “Assignment 1 - Descriptives” for a detailed list of assignment formatting guidelines. Also, assignment formatting guidelines can be found in the course document cache.

DATASET DETAILS (all data sets can be found here or here)
levers.csv

DATASET FIELDS
emp_id
prev_month_sales
employee_age
sales_training
employment_tenure
conversion
compensation_plan
num_calls