DSC 423: Data Analytics and Regression

Assignment 06

Name: Adarsh Shankar

Student Id: 2117611

Honor Statement: “I have completed this work independently. The solutions given are entirely my own work.”

1. (5 points) What is the null and alternative Hypothesis of the F-test? …of a t-test? Explain how each one can be used in the analysis of your regression model.

Answer:

F-test: The f-test determines whether or not each independent variable fits our model. The assumption that all independent variables are equal to zero is the f-null test's hypothesis. A different possibility is that there is more than one independent variable that is not zero. The independent variables in our model are evaluated using the f-test to determine their significance. We can utilise the independent variables in our model if they are statistically significant; otherwise, we cannot.

T-test: The t-test determines the significance of an independent variable. According to the null hypothesis, the independent variable is not significantly impacted by the variable in question. The other possibility is that the variable significantly influences the dependent variable. To ascertain, the t-tests are utilised.

2. (5 points) What are the four assumptions about residuals in the regression model? Why are these assumptions made? How can you verify your assumptions? How can you correct your model if the assumptions are not verified?

Answer:

The four residuals’ assumptions of the regression model are:

a) Independence: The residuals are independent of one another because there is no systematic pattern or connection between them.

b) Normality: The residuals are roughly normally distributed and have a bell-shaped distribution.

c) Equal Variance: If the residuals have an equal variance, the spread of the residuals will be the same for all values of the independent variables.

d) Zero Mean: Because the mean of the residuals is essentially zero, the average deviation from the values predicted by the model is zero.

These Presumptions are made since the validity of the regression model and the modelbased hypothesis testing depends on them. To evaluate these hypotheses, a number of diagnostic charts can be used, such as the residual vs fitted value plot, the normal probability plot, and the residual versus predictor plot. If the assumptions are not supported, you can try to alter the data, add or delete variables, or apply another kind of regression model that might be better suitable for the data.

3. (5 points) How can you judge the quality of a model? What metrics can you use to compare models?

Answer:

R-squared values typically vary from 0% to 100% and have a range from 0 to 1. An Rsquared of 100% indicates that changes in the independent variable fully explain all changes in a security (or another dependent variable).

The effectiveness of the model may be assessed using the R2 value, which shows how dependent the dependent variable is on the independent variables. We can determine whether the independent variables pass the t-test after evaluating whether the model passes the f-test.

When contrasting two models, we look at the R2 value. The best model is considered to have the most value.

4. (5 points) Given a model that predicts y given x1 and x2 write the a) first order model, b) interaction model and c) complete second order model. Which is better, under which circumstances?

Answer:

First order model: β0+ β1x1+ β2x2
Interaction model: β0+ β1x1+ β2x2+ β3x1x2
Complete second order model: β0+ β1x1+ β2x2+ β3x1x2+ β4x1^2+ β5x2^2+ β6x1^2x2^2

If the scatterplot of the model is linear, the first-order model ought to be applied. If the model is not linear, the second-order model and interaction model are utilised, and the model with the best prediction is chosen.

5. (5 points) In the model below, what is Beta-0, Beta-1, Beta-2? What is the regression line? Why was this line chosen? What is the SSE? Can you be certain that x1 and x2 should be in the model? What is R2? What does that mean? What is MSE? What does that mean? RMSE? What does that mean?

Answer:

β0 = -1338.95134
β1 = 12.74057 β2 =
85.95298 the
regression line is
Y = -1338.95134 + 12.74057 * AGE + 85.95298 * NUMBIDS + e

This line minimises the square of the errors' sum. The squares of errors add up to 516727. The variables should be used in the model because the p values are very low and X1 and X2 pass the t-test. According to the R2, 89.23% of the variance in the dependent variable is accounted for. The square of the difference between the estimated value and the actual value provides the result of 17818 for the mean square error (MSE). The root mean square error (RMSE), also known as the standard deviation of the residuals, has a value of 133.48467.

6. (5 points) How can you validate your model? Give two distinctly different methods?

Answer:

There are several methods for validating models including

- train/test splitting
- k-fold cross validation
- leave-one-out cross validation
- leave-one-group cross validation
- time-series cross validation
- Nested Cross-Validation
- Wilcoxon signed-rank test
- 5x2CV combined F test
- 5x2CV paired t test
- McNemar's test are all examples of cross-validation.

Two techniques are:

a)K-fold cross-validation: In this model, the data is divided into k folds, with one-fold acting as the test set and the remaining k-1 folds serving as the training sets. This technique is known as K-fold cross-validation. After that, the procedure is repeated k times. The average of all k folds yields the final model.

b) Train-test split: Our data is divided into train and test data sets using a train-test split model. The model is trained on the train data set and evaluated on the test data set.

7. (5 points) Explain as if to a nonprofessional why adjusted-R2 might be better than R2.

Answer:

Adjusted R2 value is superior to R2 because its value only rises when significant variables are added, whereas the R2 value rises whenever a variable is added. Thus, whenever evaluating a model, we should always use the R2 value.

8. (5 points) Define "parsimonious." Explain its relevance to building regression models.

Answer:

Parsimonious refers to a model that uses the minimum number of variables required to explain the variation in the response variable. In order to prevent overfitting, which happens when a model is too complex and captures too much noise in the data, parsimony is important when creating regression models. Overfitting can produce inaccurate conclusions about the relationships between variables and poor predictions of new data. We can lower the possibility of overfitting and increase the model's generalizability and interpretability by using a parsimonious model.

9. (5 points) Explain how to incorporate categorical features into your model? Be specific.

Answer:

Categorical variables are variables that have fixed values. These variables are included to the model as dummy variables, each of which has a value for the categorical variable. These dummy variables are then used to build the model. For each categorical variable with n values, n-1 dummy variables were created.

10. (5 points) Compare and contrast the benefits and drawbacks of forward stepwise regression, backward stepwise regression, and all-possible regression.

Answer:

Forward stepwise regression: This method of forward stepwise regression begins with an empty model and gradually adds meaningful variables.

Benefits:

- The full model need not be taken into account.
- The model is generated quickly.
- It is simple to manage vast amounts of data.

Drawbacks:

- None of the combinations are tested
- the model's parameters are biased
- it fails badly with small models.

Backward stepwise Regression: Regression using the backward stepwise method entails starting with a complete model and gradually eliminating each non-significant variable to produce a model.

Benefits:

- Every element is considered.
- The model is generated quickly.

Drawbacks:

- The parameters of the model are skewed • it struggles to handle big amounts of data.

All possible regression: Utilizing every possible combination of the independent variables, the model is constructed using this method.

Benefits:

- The whole spectrum of models is made
- and all combinations are present as advantages

Drawbacks:

- Model construction is expensive
- Model development takes time.