DSC 423: Data Analysis and Regression Assignment 06 Midterm

Saransh Thakur

Id: 2020070

**Problem 1**

F-test:

Null hypothesis: B1 = B2 = B3 = 0 (All the betas are equal to zero)

Alternative hypothesis: At least one Bi =/= 0

It means at least one beta should not be equal to zero.

T-test:

Null hypothesis: B1 = 0

Alternative hypothesis: B1=/=0

If we reject the null hypothesis and accept the alternative hypothesis that means the independent variable has an impact on the dependent variable.

If we fail to reject the null hypothesis, we are saying that we don’t know if an independent variable should be used to predict the dependent variable.

F-test can be used to determine the usefulness of the regression model whereas the T-test is used to evaluate the utility of a particular independent variable in a regression model. Both f-test and t-test should be under the significant value of 0.05.

**Problem 2**

Residual is the difference between an observed value of the response variable and the value predicted by the regression line.

The assumption about residual: -

1. **Normality of residual: -** It tells if a dataset is well modeled by a normal distribution. It means that residual should be normally distributed, then the regression line is a good fit.
2. **Unbiased: -** It is assumed that the residuals have an average value of zero means the error terms are normally distributed around zero.
3. **Homoscedastic: -** Means “same stretch”, the spread of the residuals should be the same.
4. **Linearity: -** The relationship between the independent variable and dependent variable should be linear.

These assumptions are made to check whether our model is a good fit or not.

To verify normality, we use a Q-Q plot if the observation lies on the straight line, then it means our data is normally disturbed.

We take the average value of points on a thin vertical strip to verify unbiased. If it is zero, then residuals are unbiased.

We use a residual plot to verify homoscedastic if the residual shows the same stretch as the value increase, it means it’s homoscedastic.

To check linearity, we draw a scatter plot between the independent variable and dependent variable, if the plot seems to be an alien in a straight line, then our assumption is true.

**Problem 3**

Once we have our model, we can judge the quality of our model using f-test, t-test, R squared, adjusted R square, the sum of square error, mean square error, and Root mean square error.

The f-test and p-test should be under the significance value of 0.05.

R square & adjusted R squared should be high; it measures the variability in the dependent variable which is explained by the independent variable.

The sum of square error, MSE (mean square error), and RMSE (root mean square error) should be low.

The lower the errors, the better a model fits a dataset.

**Problem 4**

1. The **first-order model** to predict y given x1 and x2 are

Y = B0 + B1x1 + B2x2 where

B0 is y intercept.

X1, x2 are qualitative variables that are not a function of other independent variables which means that they all are different, and Bi represents the slope of the line relating y to xi when all the other x’s are held fixed.

1. The **interaction model** to predict y given x1 and x2 are

Y = B0 + B1x1 + B2x2 + B3x1x2 where

B0 is y intercept.

(B1+B3x2) represent the change in y for every 1 unit increase in x1 holding x2 fixed.

(B2+B3x1) represent the change in y for every 1 unit increase in x2 holding x1 fixed.

We have a general rule for interaction terms, if our interaction term satisfies the t-test then we always include the children term, no matter what their t-test.

1. The **second-order model** to predict y are

Y = B0 + B1x1 + B2x2 + B3x1\*x2 + B4 x1^2 + B5 x2^2 where

B0 is y-intercept.

B4 and B5 talk about the rate of curvature.

If B4 and B5 are +ve the curve will be upward, if it’s negatived the curve is in a downward direction.

A first-order model is used when we have a basic model with several independent variables.

In the case of a twist in the regression plane, the interaction term may be used, whereas, in the case of a curve in the regression plane, the second-order model can be used.

**Problem 5**

Beta - 0 = -1338.95134

Beta - 1 = 12.74057

Beta - 2 = 85.95298

Regression line is y = -1338.95134 + 12.74057 age + 85.95298 numbids

This line minimizes the sum of square errors, and it best fits the data.

For, Beta 1 and Beta 2 we reject the null hypothesis and accept the alternative hypothesis which means that B1 and B2 have an impact on Y, so we will keep the prediction for Beta 1 and Beta2.

SSE (Sum of Square error) – 516727

R^2(R-squared) is 0.8923 which means 89.23% of the variability in Y is explained by our model.

MSE (Mean square error) is 17818

MSE (Mean Squared Error) represents the difference between the original and predicted values which are extracted by squaring the average difference over the data set. Lower the MSE, the better a model fits a dataset.

MSE = Σ (ŷi – Yi)2 / n

where:

* Σ is a symbol that means “sum”
* ŷi is the predicted value for the ith observation
* Yi is the observed value for the ith observation
* n is the sample size

RMSE (Root mean square error) is 133.48467

RMSE (Root Mean Squared Error) tells us the square root of the average squared difference between predicted and actual values in a dataset.

The lower the RMSE, the better a model fits a dataset.

RMSE = √Σ (ŷi – Yi)2 / n

where:

* Σ is a symbol that means “sum”
* ŷi is the predicted value for the ith observation
* Yi is the observed value for the ith observation
* n is the sample size

**Problem 6**

Validating a model means that we want our model to work well on future or unseen points. We can validate our model using two K-fold cross-validation methods and leave one out cross-validation.

**K-fold cross Validation: -**

It has a single parameter k, which tells the number of groups into which the data sample is split. The rule of thumb is to take k=10. We randomly divide the data into k equal folds to select a final model using cross-validation. Now each fold is treated as a validation set and we train our model on that. We do that for each fold for k iteration and at last, we take the average for that. It gives us unbiased results.

**Leave one out cross-validation: -**

The number of folds equals the number of instances in the data set, making it a particular example of cross-validation. The algorithm is implemented once for each instance, with all the other cases acting as a training set and the selected instance as a single-item test set.

Problem 7

When comparing models with different numbers of variables, Adj-R2 is the best model. When the number of variables is increased, R2 increases as well. R2 will grow even if a pointless variable is added to the model.

While using Adj-R2, we should constantly compare models with different numbers of independent variables. Only if the new variable improves the model, the Adj-R2 is increased.

So, adjusted R2 is better than R2.

Problem 8

A parsimonious model is a model that achieves the desired level of goodness of fit using as few explanatory variables as possible. It tells that the simplest explanation is most likely the right one.

A model with few parameters but a high level of goodness of fit should be preferred over a model with a plethora of parameters and only a slightly higher level of goodness of fit.

Models with fewer parameters are easier to understand and explain and they perform better on the new data.

Model number one:

House price = 8,830 + 81\* (sq. ft.)

**R2 adjusted: 0.7734**

Model number two:

House price = 8,921 + 77\*(sq. ft.) + 7\*(sq. ft.)2 – 9\*(age) + 600\*(rooms) + 38\*(baths)

**R2 adjusted: 0.7823**

The first model contains only one explanatory variable and has an adjusted R2 of.7734, whereas the second model contains five explanatory variables and has a slightly higher adjusted R2.

We would prefer to use the first model based on the principle of parsimony because each model has roughly the same ability to explain the variation in house prices, but the first model is much easier to understand and explain.

Problem 9

Categorical variables (qualitative variables) are variables that divide observations into groups. They have different values known as levels. We can’t use the qualitative variable directly into equations, so to solve this problem we convert qualitative variables into a binary variable. So, we use a dummy variable to convert qualitative variables into binary.

E(Y) = B0 + B1x

Where x = {1 if level A

{0 if level B

B0 is UB (mean of base level)

B1 = Ua – Ub

The number of the dummy variable for a qualitative variable is always 1 less than the number of levels.

If one dummy variable of qualitative variable has a significant p-value we should keep the other dummy variables of the qualitative variable even if they do not have a significant p-value.

Problem 10

**Forward stepwise**

Forward stepwise selection is a variable selection method which:

1. Begins with a model that contains no variables (called the *Null Model*).
2. Then starts adding the most significant variables one after the other.
3. Until a pre-specified stopping rule is reached or until all the variables under consideration are included in the model.

Advantages: -

1. Can work well on a very large dataset, even larger than the sample size.
2. It does not have to consider the full model.

Disadvantages: -

1. Given collinearity, none of them may be retained in the model.
2. Do not consider all possible combinations.
3. When the sample size is small in comparison to the number of variables, the variable selection becomes highly unstable.

**Backward stepwise selection**: - It is a variable selection method which:

1. **Begins**with a model that contains all variables under consideration (called the Full Model)
2. **Then** starts removing the least significant variables one after the other.
3. **Until**a pre-specified stopping, the rule is reached or until no variable is left in the model

Advantages: -

1. The full model has the advantage of considering the effects of all variables at the same time.
2. In the situation of collinearity (where variables in a model are correlated with one another), backward stepwise may be the best way to maintain them all in the model.

Disadvantages: -

1. If the number of candidate variables exceeds the sample size, this is not a good option.
2. Does not consider every possible combination of potential predictors.
3. When the sample size is small in comparison to the number of variables, the variable selection will be highly unstable.

**All-possible regression: -**

It is the combination of both forward and backward stepwise collections.

Advantages: -

1) It fits all possible models based on the independent variables.

2) The number of models that fit quickly multiplies.

3) It shows the best-fitting models in various sizes.

Disadvantages: -

* 1. It is slow, and the time cost may be excessive.