HW1 Part 2

Please Submit Neatly labeled knitted HTML or PDF.

**Q1)**

Use the "airbnb\_data.csv" provided and answer the following questions on Linear Regression:

a): Fit a multiple linear regression model using price as the response variable and all others as predictor variables (Note: remove 'id' columns). Which variables are statistically significant in determining the price?

b): Interpret the coefficients for predictors: room type(Shared Room), bedrooms?

c) Predict the price (nearest dollar) for a listing with the following factors: bedrooms = 1, accommodates = 2, reviews = 70, overall\_satisfaction = 4, and room\_type= 'Private room'.

d) Identify outliers using Cook's distance approach. Remove points having Cook's distance > 1. Rerun the model after removal of these points and print summary.

**Solutions**(All the codes are provided in the .rmd file)

a: Statistically significant variables: overall\_satisfaction, accommodates, bedrooms (can be concluded from p-values in the summary)

b: Interpretations are as follows:

1) Room type(Shared Room): Holding all other variables constant, a listing for a shared room has an estimated price of 76.67 USD less than an entire home/apt.

2) Bedrooms: Holding all other variables constant, the estimated price of a listing increases by 85.64 USD with an incremental bedroom in the property.

c: The estimated price for such a listing is 66.2 dollars.

d: 94 , 95

Rubric for Peer Assessment

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Question No.** | **Total Pts.** | **Ratings** |  |  |
| 1a | 4 | Marks: 4    Fitted the model and identified the significant variables correctly | Marks: 2    Modeled correctly but significant variables identification  incorrect | Marks: 0    Incorrect answer |
| 1b | 4 | Marks: 4    Interpreted the coefficients of both the predictors, room type (Shared room) and Bedroom correctly. | Marks: 2    Interpreted the coefficients of only one of the predictors correctly. | Marks: 0    Incorrect answer. |
| 1c | 4 | Marks: 4    Prepared the prediction data correctly and predicted the correct results. | Marks: 2    Prepared the prediction data correctly but incorrect results | Marks: 0    Incorrect answer. |
| 1d | 4 | Marks: 4    Identified the outliers correctly and removed them. Finally, fitted the new  Model correctly. | Marks: 2    Identified the outliers correctly but did not fit the new model. | Marks: 0    Incorrect answer |

**Q2)**

Use the "direct\_marketing.csv" provided and answer the following questions on Linear Regression:

Create indicator variables for the ‘History’ column. Considering the base case as None (i.e., create *Low*, *Medium* and *High* variables with 1 denoting the positive case and 0 the negative) and few additional variables *LowSalary*, *MediumSalary* and *HighSalary* based on the customer history type i.e., *MediumSalary* = Medium\*Salary etc.

Question a: Fit a multiple linear regression model using *AmountSpent* as the response variable and the indicator variables along with their salary variables as the predictors

*AmountSpent = b0 + b1\*Salary + b2\*Low + b3\*Medium + b4\*High + b5\*LowSalary + b6\*MediumSalary + b7\*HighSalary*

Question b: What is the amount spent by a customer for each historic type provided their salary is $10,000 based on the model constructed in question a?

Use the "airbnb\_data.csv" provided and answer the following questions (part c and partd) on Linear Regression:

Perform Log transformation for the variables *price* and *overall\_satisfaction,* make necessary transformations suggested in the class.

Question c: Fit all four models i.e., linear-linear, linear-log, log-linear and log-log regression models using *price* as the response variable and *overall\_satisfaction* as the predictor.

Question d: Which of the four models has the best R^2? Do you have any comments on the choice of the dependent variable?

**Solutions:**

Question b:

It is $91.87 for a customer with 'High' history

It is $83.74 for a customer with 'Medium' history

It is $29.98 for a customer with 'Low' history

It is $25.60 for a customer with 'No' history

Question d:

The linear -log model has the highest , Note that R^2 values are very small as this is not a very good predictor variable.

Rubric for Peer Assessment

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Question No.** | **Total Pts.** | **Ratings** |  |  |
| 2a | 4 | Marks: 4    Fitted the model and performed the transformations right | Marks: 2    Modeled correctly but didn’t perform the transformations right | Marks: 0    Incorrect answer |
| 2b | 4 | Marks: 4  All correct | Marks: 2  Some correct | Marks: 0    Incorrect answer. |
| 2c | 6 | Marks: 6    All correct | Marks: 4    Log transformations done wrong | Marks: 0    Incorrect answer. |
| 2d | 2 | Marks: 2    Linear-log + some comments | Marks: 1    No comments | Marks: 0    Incorrect answer |

**Q3)**

The attached titanic\_data.csv file was obtained from this source: <http://math.ucdenver.edu/RTutorial/>

It has been cleaned to remove all rows which contain missing values. We will perform a logistic regression on this cleaned dataset.

The dataset contains the following columns:

‘Name’ - Passenger Name - factor

‘PClass’ - Passenger Class (1st, 2nd, 3rd) - factor

‘Age’ - Passenger Age - number

‘Sex’ - Passenger Sex – female, male

‘Survived’ – 1 if passenger survived, 0 if not - number

After converting the survived variable to be a factor with two levels, 0 and 1, perform a logistic regression on the dataset using ‘survived’ as the response and ‘Sex’ as the explanatory variable.

a. Display the model summary.

b. What does the value of the intercept coefficient represent in this model?

c. Determine the probability of survival for females.

d. Determine the probability of survival for males.

**Solutions:**

*#library(tidyverse)*  
 *#titanicDF <- read.csv('http://math.ucdenver.edu/RTutorial/titanic.txt',sep='\t')*  
 *#data <- drop\_na(titanicDF)*  
 *#write.csv(data, 'titanic\_data.csv', row.names = FALSE)*

**a. Display the model summary.**

data <- **read.csv**('titanic\_data.csv')  
 data**$**Survived <- **as.factor**(data**$**Survived)  
 model1 <- **glm**(Survived **~**Sex, family = 'binomial', data = data)  
 **summary**(model1)

##   
## Call:  
 ## glm(formula = Survived ~ Sex, family = "binomial", data = data)  
 ##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6735 -0.6776 -0.6776 0.7524 1.7800   
##   
## Coefficients:  
 ## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.1172 0.1367 8.171 3.05e-16 \*\*\*  
 ## Sexmale -2.4718 0.1783 -13.861 < 2e-16 \*\*\*  
 ## ---  
 ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
 ##   
## (Dispersion parameter for binomial family taken to be 1)  
 ##   
## Null deviance: 1025.57 on 755 degrees of freedom  
 ## Residual deviance: 796.64 on 754 degrees of freedom  
 ## AIC: 800.64  
 ##   
## Number of Fisher Scoring iterations: 4

**b. What does the value of the intercept coefficient represent in this model?**

The intercept value of 1.1172 is the log odds for females, since they are the reference group in the model.

**c. Determine the probability of survival for females.**

For females, p = (Prob of Survival = 1|Male =0) = odds/(1+odds).

Odds of survival for females = exp(1.1172) = 3.056

p = 3.056/(1+3.056) = 0.753

**d. Determine the probability of survival for males.**

Since females are the reference group, we have to add the coefficient for males to the intercept to determine survival odds for males.

Odds of survival for males = exp(1.1172-2.4718) = 0.258

p = 0.258/(1+0.258) = 0.205

HW1 Week3 Peer Assessment Rubric

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Question No.** | **Total Pts.** | **Ratings** |  |  |
| 3a & 3b | 4 | Marks: 4    Fitted the model and interpreted the value of the coefficients correctly | Marks: 2    Modeled correctly but interpretation of the value of the coefficients incorrect | Marks: 0    Incorrect answer |
| 3c & 3d | 4 | Marks: 4    Probability of survival rates for both Females and males correct | Marks: 2    Probability of only one of the survival rates for Females and males correct. | Marks: 0    Incorrect answer. |