HW1 Part 2

**Question 1)**

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

**Question 1a:**

Remove ‘id’columns (‘room\_id’, ‘survey\_id’, ‘host\_id’) and ‘city’ from your dataset, and fit a multiple linear regression model using price as the response variable and all others as predictor variables. (Note: Do not fit a model using id columns and city as predictors). Which variables are statistically significant at a 95% confidence interval.

**Question 1b:**

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

**Question 1c:**

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'.

**Question 1d:**

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.

**Answer 1)**

**Answer 1a:**

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

*# Load data and print head*

*airbnb\_data <- read.csv("airbnb\_data.csv",header = TRUE)*

*head(airbnb\_data)*

*# create model using 'lm' and print summary*

*model1 <- lm(price ~ room\_type + reviews + overall\_satisfaction + accommodates + bedrooms, data = airbnb\_data)*

*summary(model1)*

**Answer 1b:**

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.

**Answer 1c:**

The estimated price for such a listing is 66.2 dollars.

*# prepare prediction data and use 'predict' to find the value*

*pred\_data = data.frame(bedrooms = 1, accommodates = 2, reviews = 70, overall\_satisfaction = 4, room\_type = 'Private room')*

*predict(model1, pred\_data)*

**Answer 1d:**

Outliers to remove: 94 , 95

*# Use cook's distance to identify outliers*

*cooks <-cooks.distance(model1)*

*which(cooks>1)*

*# remove the outliers*

*airbnb\_data\_2 = airbnb\_data[-c(94,95),]*

*# creating new model and print summary*

*model2 <- lm(price ~ room\_type + reviews + overall\_satisfaction + accommodates + bedrooms, data = airbnb\_data\_2)*

*summary(model2)*

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 |

**Question 2)**

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 2a:**

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 2b:**

What is the amount spent by a customer for each historic type(None, Low, Medium, and High) 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 part d) on Linear Regression. **DO NOT** remove outliers from the dataset:

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

**Question 2c:**

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. (Note: Because overall\_satisfaction contains ‘0’ values, you will need to use log(x+1) transformations instead of log(x) transformations)

**Question 2d:**

Which of the four models has the best R^2? Do you have any comments on the choice of the independent variable?

**Answer 2)**

**Answer 2a:**

*library(tidyverse)*

*library(readr)*

*# Load data and print head*

*direct\_marketing <- read\_csv("direct\_marketing.csv",*

*col\_types = cols(Catalogs = col\_integer(), Children = col\_integer()))*

*head(direct\_marketing)*

*# Create indicator variables*

*direct\_marketing <- direct\_marketing %>%*

*mutate(Low = ifelse(History == "Low",1,0))%>%*

*mutate(Medium = ifelse(History == "Medium",1,0))%>%*

*mutate(High = ifelse(History == "High",1,0))%>%*

*mutate(LowSalary = Low\*Salary)%>%*

*mutate(MediumSalary = Medium\*Salary)%>%*

*mutate(HighSalary = High\*Salary)*

*# create model using 'lm' and print summary*

*model = lm(AmountSpent~Salary + Low + Medium + High + LowSalary + MediumSalary +*

*HighSalary,data = direct\_marketing)*

*summary(model)*

**Answer 2b:**

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

*# prepare prediction data and use 'predict' to find the value*

*pred\_data = data.frame(Salary = 10000,*

*High = 1,*

*Medium = 0,*

*Low = 0,*

*HighSalary = 10000,*

*MediumSalary = 0,*

*LowSalary = 0)%>%*

*add\_row(Salary = 10000,*

*High = 0,*

*Medium = 1,*

*Low = 0,*

*HighSalary = 0,*

*MediumSalary = 10000,*

*LowSalary = 0)%>%*

*add\_row(Salary = 10000,*

*High = 0,*

*Medium = 0,*

*Low = 1,*

*HighSalary = 0,*

*MediumSalary = 0,*

*LowSalary = 10000)%>%*

*add\_row(Salary = 10000,*

*High = 0,*

*Medium = 0,*

*Low = 0,*

*HighSalary = 0,*

*MediumSalary = 0,*

*LowSalary = 0)*

*#make prediction*

*predict(model, pred\_data)*

**Answer 2c:**

*#Read data*

*airbnb\_data <- read\_csv("airbnb\_data.csv")*

*#Perform Transformations*

*airbnb\_data <- airbnb\_data %>%*

*mutate(log\_OverallSatisfaction = log(overall\_satisfaction+1))%>%*

*mutate(log\_Price = log(price))*

*#Linear-Linear Model*

*model2 = lm(price~overall\_satisfaction,data = airbnb\_data)*

*summary(model2)*

*#Linear-Log Model*

*model3 = lm(price~log\_OverallSatisfaction,data = airbnb\_data)*

*summary(model3)*

*#Log-Linear Model*

*model4 = lm(log\_Price~overall\_satisfaction,data = airbnb\_data)*

*summary(model4)*

*#Log-Log Model*

*model5 = lm(log\_Price~log\_OverallSatisfaction,data = airbnb\_data)*

*summary(model5)*

**Answer 2d:**

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 |

**Question 3)**

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.

**Question 3a:**

Display the model summary.

**Question 3b:**

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

**Question 3c:**

Determine the probability of survival for females.

**Question 4d:**

Determine the probability of survival for males.

**Answer 3)**

**Answer 3A:**

*#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)*

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

**Answer 3b.**

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

**Answer 3c.**

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

**Answer 3d.**

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. |