Coding Assignment 3

Team 06

Due: 2022-12-09 23:59

A Florida health insurance company wants to predict annual claims for individual clients. The company pulls a random sample of 50 customers. The owner wishes to charge an actuarially fair premium to ensure a normal rate of return. The owner collects all of their current customer’s health care expenses from the last year and compares them with what is known about each customer’s plan.

The data on the 50 customers in the sample is as follows:

* Charges: Total medical expenses for a particular insurance plan (in dollars)
* Age: Age of the primary beneficiary
* BMI: Primary beneficiary’s body mass index (kg/m2)
* Female: Primary beneficiary’s birth sex (0 = Male, 1 = Female)
* Children: Number of children covered by health insurance plan (includes other dependents as well)
* Smoker: Indicator if primary beneficiary is a smoker (0 = non-smoker, 1 = smoker)
* Cities: Dummy variables for each city with the default being Sanford

Answer the following questions using complete sentences and attach all output, plots, etc. within this report.

insurance <- read.csv("../Data/insurance\_0126\_Group6.csv")

gt(head(insurance))

Table

Description automatically generated

## **Question 1**

Randomly select three observations from the sample and exclude from all modeling (i.e. n=47). Provide the summary statistics (min, max, std, mean, median) of the quantitative variables for the 47 observations.

set.seed(123)

index <- sample(seq\_len(nrow(insurance)), size = 3)

train <- insurance[-index,]

test <- insurance[index,]

summary(train)

## Charges Age BMI Female

## Min. : 1149 Min. :18.00 Min. :17.48 Min. :0.0000

## 1st Qu.: 3442 1st Qu.:26.00 1st Qu.:26.23 1st Qu.:0.0000

## Median : 8060 Median :38.00 Median :30.97 Median :1.0000

## Mean :10241 Mean :39.62 Mean :30.89 Mean :0.5745

## 3rd Qu.:11845 3rd Qu.:53.00 3rd Qu.:35.09 3rd Qu.:1.0000

## Max. :47897 Max. :64.00 Max. :46.53 Max. :1.0000

## Children Smoker WinterSprings WinterPark

## Min. :0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000

## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000

## Median :1.000 Median :0.0000 Median :0.0000 Median :0.0000

## Mean :1.043 Mean :0.1064 Mean :0.2553 Mean :0.2553

## 3rd Qu.:2.000 3rd Qu.:0.0000 3rd Qu.:0.5000 3rd Qu.:0.5000

## Max. :4.000 Max. :1.0000 Max. :1.0000 Max. :1.0000

## Oviedo

## Min. :0.0000

## 1st Qu.:0.0000

## Median :0.0000

## Mean :0.2128

## 3rd Qu.:0.0000

## Max. :1.0000

train %>%

tbl\_summary(statistic = list(all\_continuous() ~ c("{mean} ({sd})",

"{median} ({p25}, {p75})",

"{min}, {max}"),

all\_categorical() ~ "{n} / {N} ({p}%)"),

type = all\_continuous() ~ "continuous2")

Table

Description automatically generated

Table

Description automatically generated

According to the summary statistic output, the dependent variable, Charges is skewed to the right since it mean is higher than the median by about 2 thousand dollars. Since the standard deviation is high, some outliers could exist in the data.

## **Question 2**

Provide the correlation between all quantitative variables

cor(insurance)

## Charges Age BMI Female Children

## Charges 1.00000000 0.63156188 0.11984359 0.12478792 -0.12251184

## Age 0.63156188 1.00000000 0.09410864 0.25563075 0.03021326

## BMI 0.11984359 0.09410864 1.00000000 -0.03888301 0.12632864

## Female 0.12478792 0.25563075 -0.03888301 1.00000000 0.03355780

## Children -0.12251184 0.03021326 0.12632864 0.03355780 1.00000000

## Smoker 0.82667950 0.28018237 -0.02194003 0.09412416 -0.28427400

## WinterSprings 0.18155886 -0.02163867 0.05569997 -0.01911798 0.23096064

## WinterPark -0.01385958 -0.05225310 0.22359816 -0.07445808 -0.07495949

## Oviedo -0.09338461 -0.08435673 -0.05672896 0.03942083 0.00000000

## Smoker WinterSprings WinterPark Oviedo

## Charges 0.82667950 0.18155886 -0.01385958 -0.09338461

## Age 0.28018237 -0.02163867 -0.05225310 -0.08435673

## BMI -0.02194003 0.05569997 0.22359816 -0.05672896

## Female 0.09412416 -0.01911798 -0.07445808 0.03942083

## Children -0.28427400 0.23096064 -0.07495949 0.00000000

## Smoker 1.00000000 0.04318713 0.02365303 0.06400530

## WinterSprings 0.04318713 1.00000000 -0.33309617 -0.29844401

## WinterPark 0.02365303 -0.33309617 1.00000000 -0.31480009

## Oviedo 0.06400530 -0.29844401 -0.31480009 1.00000000

According to the correlation output, “Age” and “Smoker” variables are strongly and directly related to the dependent variable, Charges.

## **Question 3**

Run a regression that includes all independent variables in the data table. Does the model above violate any of the Gauss-Markov assumptions? If so, what are they and what is the solution for correcting?

original\_model <- lm(Charges ~., data = train[,c(1:9)] ) *#pulling only original data column before data transformation*

summary(original\_model)

##

## Call:

## lm(formula = Charges ~ ., data = train[, c(1:9)])

##

## Residuals:

## Min 1Q Median 3Q Max

## -9468.8 -1130.9 -95.9 1084.4 10860.4

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -8148.07 3083.95 -2.642 0.0119 \*

## Age 293.40 41.59 7.055 2.08e-08 \*\*\*

## BMI 138.52 91.89 1.507 0.1400

## Female -681.82 1154.14 -0.591 0.5582

## Children 317.54 478.69 0.663 0.5111

## Smoker 24978.91 2037.60 12.259 8.91e-15 \*\*\*

## WinterSprings 2289.36 1622.62 1.411 0.1664

## WinterPark -1069.77 1604.70 -0.667 0.5090

## Oviedo -1980.10 1608.94 -1.231 0.2260

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 3719 on 38 degrees of freedom

## Multiple R-squared: 0.9026, Adjusted R-squared: 0.8821

## F-statistic: 44.04 on 8 and 38 DF, p-value: < 2.2e-16

Chart, line chart

Description automatically generated

vif(original\_model)

## Age BMI Female Children Smoker

## 1.301281 1.114801 1.106755 1.191726 1.341536

## WinterSprings WinterPark Oviedo

## 1.701470 1.664110 1.473749

## **Question 3 Answers=>**

No significant VIFs are found for independent variables and no concerning multicollinearity is detected.

The “Residual vs Fitted” model tests for the third Gauss-Markov assumption. The model shows that our results are non-linear as the data curves down and then shoots up. If it were properly specified, it would be a perfectly horizontal line with a bit of noise of around it. This is a violation of the third assumption. The solution would be to consider using ratios or percentages rather than raw data.

The “Normal Q-Q” model” checks the sixth Gauss-Markov assumption, which is that we have a normally distributed error term. In this case, there is a violation of the sixth assumption because there are large variables going far outside of our line (top right corner). To correct this violation we can fit the models to several sets of transformed data and then check to see which transformation appears to produce the most normally distributed residuals.

The “Scale-Location” model tests for heteroscedasticity. Our results show that there is a violation of the fourth Gauss-Markov assumption as the data is not in a horizontal line with corresponding noise; rather as the size gets larger, our data is spread is widening creating a “fanning out” effect. To correct this violation, we should look for subgroups in data and analyze separately. We should also use summary data (like the mean value) rather than the raw data.

The “Residuals vs Leverage” model checks for outliers. There are no Gauss-Markov violations shown.

## **Question 4**

Implement the solutions from question 3, such as data transformation, along with any other changes you wish. Use the sample data and run a new regression. How have the fit measures changed? How have the signs and significance of the coefficients changed?

*#to detect non-linear relationships and non-normally distributed variables*

scatterplotMatrix(train[,c(1,2,3,5)])

Diagram

Description automatically generated

Some nonlinearity between Charges and Age have been found.

*#Data Transformation*

*#taking natural logirthm of Charges to bring possible outliers closer*

par(mfrow=c(1,2))

hist(train$Charges) *#before*

train$lnCharges <- log(train$Charges)

hist(train$lnCharges) *#after*

Chart, histogram

Description automatically generated

Question#4 Answer==>

Taking natural logirthm of Charges transfrom the dependent variable to be more normally distributed.

*#Plotting the relationships After Transformation*

scatterplotMatrix(train[,c(10,2,3)]) *# grabbing lnCharges*

Graphical user interface, diagram

Description automatically generated

*#to determine if the relationship between Charges and Age is logarithmic or quadratic.*

train$lnAge <- log(train$Age)

train$AgeSquared <- train$Age^2

scatterplotMatrix(train[,c(10,11,3)]) *# lnChares & lnAge*

Diagram

Description automatically generated

scatterplotMatrix(train[,c(10,12,3)]) *# lnChares & AgeSquared*

Chart, diagram

Description automatically generated

*# New Regression*

*# Models to look at the regression coefficients.*

*# Model 1lnAge: Age with a logaritmic shape*

model\_1lnAge <- lm(lnCharges ~., data = train[,c(10,11,3:9)] )

summary(model\_1lnAge)

##

## Call:

## lm(formula = lnCharges ~ ., data = train[, c(10, 11, 3:9)])

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.38585 -0.12001 -0.00025 0.04639 1.27010

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.980685 0.385580 7.730 2.58e-09 \*\*\*

## lnAge 1.541723 0.103166 14.944 < 2e-16 \*\*\*

## BMI -0.000688 0.006219 -0.111 0.9125

## Female 0.135518 0.078717 1.722 0.0933 .

## Children 0.080173 0.032527 2.465 0.0183 \*

## Smoker 1.254884 0.136858 9.169 3.59e-11 \*\*\*

## WinterSprings 0.202225 0.109201 1.852 0.0718 .

## WinterPark -0.096943 0.108649 -0.892 0.3779

## Oviedo -0.014314 0.108391 -0.132 0.8956

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.2518 on 38 degrees of freedom

## Multiple R-squared: 0.9354, Adjusted R-squared: 0.9218

## F-statistic: 68.76 on 8 and 38 DF, p-value: < 2.2e-16

*# or*

tbl\_regression(model\_1lnAge,

estimate\_fun = ~style\_sigfig(.x, digits = 4)) %>% as\_gt() %>%

gt::tab\_source\_note(gt::md(paste0("Adjusted R-Squared: ",round(summary(model\_1lnAge)$adj.r.squared\* 100,digits = 2),"%")))

| **Characteristic** | **Beta** | **95% CI***1* | **p-value** |
| --- | --- | --- | --- |
| lnAge | 1.542 | 1.333, 1.751 | <0.001 |
| BMI | -0.0007 | -0.0133, 0.0119 | >0.9 |
| Female | 0.1355 | -0.0238, 0.2949 | 0.093 |
| Children | 0.0802 | 0.0143, 0.1460 | 0.018 |
| Smoker | 1.255 | 0.978, 1.532 | <0.001 |
| WinterSprings | 0.2022 | -0.0188, 0.4233 | 0.072 |
| WinterPark | -0.0969 | -0.3169, 0.1230 | 0.4 |
| Oviedo | -0.0143 | -0.2337, 0.2051 | 0.9 |
| Adjusted R-Squared: 92.18% | | | |
| *1* CI = Confidence Interval | | | |

par(mfrow=c(2,2))

plot(model\_1lnAge)

Diagram

Description automatically generated

*# Model 2AgeSq:Quadratic Relationship between Age and Charges*

model\_2AgeSq <- lm(lnCharges ~., data = train[,c(10,3:9,12)] )

summary(model\_2AgeSq)

##

## Call:

## lm(formula = lnCharges ~ ., data = train[, c(10, 3:9, 12)])

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.54586 -0.13583 0.01408 0.14348 1.24806

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 7.5596702 0.2443221 30.941 < 2e-16 \*\*\*

## BMI -0.0012183 0.0077385 -0.157 0.8757

## Female 0.1989881 0.0966186 2.060 0.0463 \*

## Children 0.1027793 0.0401811 2.558 0.0146 \*

## Smoker 1.2074779 0.1727264 6.991 2.53e-08 \*\*\*

## WinterSprings 0.2941239 0.1373456 2.141 0.0387 \*

## WinterPark -0.1163559 0.1348752 -0.863 0.3937

## Oviedo 0.1048258 0.1361433 0.770 0.4461

## AgeSquared 0.0004974 0.0000434 11.460 6.78e-14 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.3128 on 38 degrees of freedom

## Multiple R-squared: 0.9003, Adjusted R-squared: 0.8793

## F-statistic: 42.88 on 8 and 38 DF, p-value: < 2.2e-16

*# or*

tbl\_regression(model\_2AgeSq,

estimate\_fun = ~style\_sigfig(.x, digits = 4)) %>% as\_gt() %>%

gt::tab\_source\_note(gt::md(paste0("Adjusted R-Squared: ",round(summary(model\_2AgeSq)$adj.r.squared\* 100,digits = 2),"%")))

| **Characteristic** | **Beta** | **95% CI***1* | **p-value** |
| --- | --- | --- | --- |
| BMI | -0.0012 | -0.0169, 0.0144 | 0.9 |
| Female | 0.1990 | 0.0034, 0.3946 | 0.046 |
| Children | 0.1028 | 0.0214, 0.1841 | 0.015 |
| Smoker | 1.207 | 0.8578, 1.557 | <0.001 |
| WinterSprings | 0.2941 | 0.0161, 0.5722 | 0.039 |
| WinterPark | -0.1164 | -0.3894, 0.1567 | 0.4 |
| Oviedo | 0.1048 | -0.1708, 0.3804 | 0.4 |
| AgeSquared | 0.0005 | 0.0004, 0.0006 | <0.001 |
| Adjusted R-Squared: 87.93% | | | |
| *1* CI = Confidence Interval | | | |

par(mfrow=c(2,2))

plot(model\_2AgeSq)

Diagram, engineering drawing

Description automatically generated

*# # Model 3lnBMI: BMI with a logaritmic shape*

train$lnBMI <- log(train$BMI)

model\_3lnBMI <- lm(lnCharges ~., data = train[,c(10,11,4:9,13)] )

summary(model\_3lnBMI)

##

## Call:

## lm(formula = lnCharges ~ ., data = train[, c(10, 11, 4:9, 13)])

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.37769 -0.10989 -0.01101 0.04304 1.26948

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 2.92002 0.65974 4.426 7.83e-05 \*\*\*

## lnAge 1.53895 0.10345 14.876 < 2e-16 \*\*\*

## Female 0.13607 0.07869 1.729 0.0919 .

## Children 0.07963 0.03255 2.447 0.0191 \*

## Smoker 1.25585 0.13683 9.178 3.50e-11 \*\*\*

## WinterSprings 0.19922 0.10918 1.825 0.0759 .

## WinterPark -0.10265 0.10830 -0.948 0.3492

## Oviedo -0.01630 0.10855 -0.150 0.8814

## lnBMI 0.01531 0.18218 0.084 0.9335

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.2518 on 38 degrees of freedom

## Multiple R-squared: 0.9354, Adjusted R-squared: 0.9218

## F-statistic: 68.75 on 8 and 38 DF, p-value: < 2.2e-16

*# or*

tbl\_regression(model\_3lnBMI,

estimate\_fun = ~style\_sigfig(.x, digits = 4)) %>% as\_gt() %>%

gt::tab\_source\_note(gt::md(paste0("Adjusted R-Squared: ",round(summary(model\_3lnBMI)$adj.r.squared\* 100,digits = 2),"%")))

| **Characteristic** | **Beta** | **95% CI***1* | **p-value** |
| --- | --- | --- | --- |
| lnAge | 1.539 | 1.330, 1.748 | <0.001 |
| Female | 0.1361 | -0.0232, 0.2954 | 0.092 |
| Children | 0.0796 | 0.0137, 0.1455 | 0.019 |
| Smoker | 1.256 | 0.979, 1.533 | <0.001 |
| WinterSprings | 0.1992 | -0.0218, 0.4203 | 0.076 |
| WinterPark | -0.1027 | -0.3219, 0.1166 | 0.3 |
| Oviedo | -0.0163 | -0.2360, 0.2035 | 0.9 |
| lnBMI | 0.0153 | -0.3535, 0.3841 | >0.9 |
| Adjusted R-Squared: 92.18% | | | |
| *1* CI = Confidence Interval | | | |

par(mfrow=c(2,2))

plot(model\_3lnBMI)

Diagram, engineering drawing

Description automatically generated

## Question 4 Answer=>

After the data transformation on both dependent variable, Charges and 2 independent variables, Age and BMI, regression Model3 shows improved measures. Residual Standard Error has decreased to 0.2518. Adjusted R^2 has increased to 92%. F statistic is also showing a higher number of 68.75 compared to that of Original model. The signs of coefficient didn’t change. Additionally, the independent variable “Children” has become statistically significant for Model3 which was not significant for Original Model.

## **Question 5**

Use the 3 withheld observations and calculate the performance measures for your best two models. Which is the better model? (Remember that “better” depends on whether your outlook is short or long run)

test$lnCharges <- log(test$Charges)

test$lnAge <- log(test$Age)

test$AgeSquared <- test$Age^2

test$lnBMI <- log(test$BMI)

test$original\_model\_pred <- predict(original\_model, newdata = test)

test$model\_1lnAge\_pred <- predict(model\_1lnAge,newdata = test) %>% exp()

test$model\_2AgeSq\_pred <- predict(model\_2AgeSq,newdata = test) %>% exp()

test$model\_3lnBMI\_pred <- predict(model\_3lnBMI,newdata = test) %>% exp()

*# Finding the error*

test$error\_OM <- test$original\_model\_pred - test$Charges

test$error\_1lnAge<- test$model\_1lnAge\_pred - test$Charges

test$error\_2SqAge<- test$model\_2AgeSq\_pred - test$Charges

test$error\_3lnBMI<- test$model\_3lnBMI\_pred - test$Charges

*# Bias*

*# Original Model*

mean(test$error\_OM)

## [1] 4879.738

*# Model 1lnAge*

mean(test$error\_1lnAge)

## [1] 2656.491

*# Model 2SqAge*

mean(test$error\_2SqAge)

## [1] 7168.83

*#Model 3lnBMI*

mean(test$error\_3lnBMI)

## [1] 2627.923

BIAS [1] 4879.738 [1] 2656.491 [1] 7168.83 [1] 2627.923 (best)

*# MAE*

mae <- **function**(error\_vector){

error\_vector %>%

abs() %>%

mean()}

*# Original Model*

mae(test$error\_OM)

## [1] 5721.517

*# Model 1lnAge*

mae(test$error\_1lnAge)

## [1] 9328.276

*# Model 2SqAge*

mae(test$error\_2Sq)

## [1] 12033.55

*#Model 3lnBMI*

mae(test$error\_3lnBMI)

## [1] 9300.512

MAE [1] 5721.517 [1] 9328.276 [1] 12033.55 [1] 9300.512 better

*### RMSE*

rmse <- **function**(error\_vector){

error\_vector^2 %>%

mean() %>%

sqrt()}

*# Original Model*

rmse(test$error\_OM)

## [1] 6627.591

*# Model 1lnAge*

rmse(test$error\_1lnAge)

## [1] 11853.78

*# Model 2SqAge*

rmse(test$error\_2SqAge)

## [1] 17117.81

*#Model 3lnBMI*

rmse(test$error\_3lnBMI)

## [1] 11808.4

RMSE [1] 6627.591 [1] 11853.78 better [1] 17117.81 [1] 11808.4

*### MAPE*

mape <- **function**(error\_vector, actual\_vector){

(error\_vector/actual\_vector) %>%

abs() %>%

mean()

}

*# Original Model*

mape(test$error\_OM, test$Charges)

## [1] 0.3584357

*# Model 1lnAge*

mape(test$error\_1lnAge, test$Charges)

## [1] 0.406728

*# Model 2SqAge*

mape(test$error\_2SqAge, test$Charges)

## [1] 0.4912408

*# Model 3lnBMI*

mape(test$error\_3lnBMI, test$Charges)

## [1] 0.4066504

# Question 5 Answer

Summary of Performance Metrics BIAS [1] 4879.738 [1] 2656.491 lower bias better for long run [1] 7168.83 [1] 2627.923 (best) over prediction MAE [1] 5721.517 [1] 9328.276 [1] 12033.55 [1] 9300.512 better RMSE [1] 6627.591 [1] 11853.78 [1] 17117.81 [1] 11808.4 better MAPE [1] 0.3584357 [1] 0.406728 [1] 0.4912408 [1] 0.4066504 better

is our interest long run or short run? short run which model is the best? =>Model3lnBMI

The original model is bad model with high BIAS value and High RMSE value and. Comparing the proposed models with transformed data, Model3lnBMI has better BIAS and RMSE value.Model3lnBMI has higher MAE value than that of the Original Model due to possible outliers, however, after normalizing the errors, the MAPE value of Model3lnBMI became almost the same as the original MAPE. With Lowest RMSE among the models and lowest BIAS value, Model3lnBMI is the best model for the short run.

## **Question 6**

Provide interpretations of the coefficients, do the signs make sense? Perform marginal change analysis (thing 2) on the independent variables.

*### Marginal Change Analysis*

model\_3lnBMI$coefficients["lnAge"]

## lnAge

## 1.53895

model\_3lnBMI$coefficients["Female"]

## Female

## 0.1360697

model\_3lnBMI$coefficients["Children"]

## Children

## 0.07963402

model\_3lnBMI$coefficients["Smoker"]

## Smoker

## 1.255853

model\_3lnBMI$coefficients["WinterSprings"]

## WinterSprings

## 0.1992206

# Question 6 Answer==>

Marginal Change Analysis lnAge 1.53895 Female 0.1360697 Children 0.07963402 Smoker 1.255853 WinterSprings 0.1992206

According to the Model3 coefficients, Age is the strongest determinant for the premium charges and smoker found to be the second in order. Coefficient sings also show the true relationship between each independent variable and the charges. lnAge, Female, Children and Smoker and Winter Springs all tested statistically significant. WinterPark, Oviedo, and LnMBI did not test statistically significant, so we are unable to change perform marginal change analysis for these variables. Our coefficient for lnAge represents that a 1% change will affect charges by 1.5%. Our coefficient for Female represents that a 1% change will affect charges by 1.5%. Our coefficient for Children represents that a 1% change will affect charges by .08%. Our coefficient for Winter Springs represents that a 1% change will affect charges by .2%.The coefficients are consistent with what we expected regarding their sign. For example, we assumed that an increase in age, an increase in BMI, and being a smoke would increase charges. Also, women tend to incur greater health costs than men, particularly during the reproductive years, so the positive sign is also consistent with what we expect.

## **Question 7**

An eager insurance representative comes back with five potential clients. Using the better of the two models selected above, provide the prediction intervals for the five potential clients using the information provided by the insurance rep.

| **Customer** | **Age** | **BMI** | **Female** | **Children** | **Smoker** | **City** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 60 | 22 | 1 | 0 | 0 | Oviedo |
| 2 | 40 | 30 | 0 | 1 | 0 | Sanford |
| 3 | 25 | 25 | 0 | 0 | 1 | Winter Park |
| 4 | 33 | 35 | 1 | 2 | 0 | Winter Springs |
| 5 | 45 | 27 | 1 | 3 | 0 | Oviedo |

PredictionCUS1 <- data.frame(Age = 60,

lnAge=4.094345 ,

BMI = 22,

lnBMI=3.091042,

Female = 1,

Children=0,

Smoker=0,

city=0,

WinterSprings=0,

WinterPark=0,

Oviedo=1)

predict(model\_3lnBMI, newdata =PredictionCUS1,interval = "confidence",

level = .5)

## fit lwr upr

## 1 9.388107 9.304008 9.472206

C1<-predict(model\_3lnBMI, newdata =PredictionCUS1,interval = "confidence",

level = .5)

*# calculate the inverse log using exp( ) function.*

antilog\_C1 = exp (C1)

print(antilog\_C1)

## fit lwr upr

## 1 11945.47 10981.95 12993.52

PredictionCUS2 <- data.frame(Age = 40,

lnAge=3.688879 ,

BMI = 30,

lnBMI=3.401197,

Female = 0,

Children=1,

Smoker=0,

city=1,

WinterSprings=0,

WinterPark=0,

Oviedo=0)

predict(model\_3lnBMI, newdata =PredictionCUS2,interval = "confidence",

level = .5)

## fit lwr upr

## 1 8.728727 8.669521 8.787933

C2<-predict(model\_3lnBMI, newdata =PredictionCUS2,interval = "confidence",

level = .5)

antilog\_C2 = exp (C2)

print(antilog\_C2)

## fit lwr upr

## 1 6177.858 5822.709 6554.668

PredictionCUS3 <- data.frame(Age = 25,

lnAge=3.218876 ,

BMI = 25,

lnBMI=3.218876,

Female = 0,

Children=0,

Smoker=1,

city=0,

WinterSprings=0,

WinterPark=1,

Oviedo=0)

predict(model\_3lnBMI, newdata =PredictionCUS3,interval = "confidence",

level = .5)

## fit lwr upr

## 1 9.076193 8.974388 9.177998

C3<-predict(model\_3lnBMI, newdata =PredictionCUS3,interval = "confidence",

level = .5)

antilog\_C3 = exp (C3)

print(antilog\_C3)

## fit lwr upr

## 1 8744.609 7898.18 9681.747

PredictionCUS4 <- data.frame(Age = 33,

lnAge=3.496508 ,

BMI = 35,

lnBMI=3.555348,

Female = 1,

Children=2,

Smoker=0,

city=0,

WinterSprings=1,

WinterPark=0,

Oviedo=0)

predict(model\_3lnBMI, newdata =PredictionCUS4,interval = "confidence",

level = .5)

## fit lwr upr

## 1 8.849962 8.790622 8.909302

C4<-predict(model\_3lnBMI, newdata =PredictionCUS4,interval = "confidence",

level = .5)

antilog\_C4 = exp (C4)

print(antilog\_C4)

## fit lwr upr

## 1 6974.123 6572.32 7400.491

PredictionCUS5 <- data.frame(Age = 45,

lnAge=3.806662 ,

BMI = 27,

lnBMI=3.295837,

Female = 1,

Children=3,

Smoker=0,

city=0,

WinterSprings=0,

WinterPark=0,

Oviedo=1)

predict(model\_3lnBMI, newdata =PredictionCUS5,interval = "confidence",

level = .5)

## fit lwr upr

## 1 9.187415 9.115552 9.259277

C5<-predict(model\_3lnBMI, newdata =PredictionCUS5,interval = "confidence",

level = .5)

antilog\_C5 = exp (C5)

print(antilog\_C5)

## fit lwr upr

## 1 9773.352 9095.658 10501.54

# Question 7 Answer=>

Prediction Intervals Customer1: fit lwr upr 11945.47 10981.95 12993.52 Customer2: fit lwr upr 6177.858 5822.709 6554.668 Customer3: fit lwr upr 8744.609 7898.18 9681.747 Customer4: fit lwr upr 6974.123 6572.32 7400.491 Customer5: fit lwr upr 9773.352 9095.658 10501.54

## **Question 8**

The owner notices that some of the predictions are wider than others, explain why.

Using individual observances instead of the mean increases uncertainty and results in a wider range of possible values. The more spread the data is, the larger the variance is in relation to the mean. This can cause a prediction interval with a wider range of values. Extrapolation error could also cause the wider prediction.

## **Question 9**

Are there any prediction problems that occur with the five potential clients? If so, explain.

Model3lnBMI’s high MAPE value can signal some potential problems with prediction accuracy for the potential clients. However, Model3’s high MAPE value can be due to the frequent changes in premium charges or the” zero” values in the data set. Since there is no industry standard for a good MAPE value, the model’s high MAPE can be ignored. For instance, customer1 is predicted to be charged the most for his/age despite his/her low BMI while customer2 is being charged the least for being a non-smoker in his middle age with higher BMI. Since the charges reflect the predictors well, there is no concerning prediction problems.