Coding Assignment 3

Team 16

Due: 2023-12-09 23:59

Contents

A Florida health insurance company wants to predict annual claims for individual clients. The company pulls a random sample of 100 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 100 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.

```
# Bring in the dataset here.
Insurance_Data_Group16 <- read_csv("~/GitHub/ECO6416_Group16/Data/Insurance_Data_Group16.csv",
                                   show_col_types = FALSE)
```

Question 1

Randomly select 30 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 70 observations.

```
set.seed(123456)

index <- sample(seq_len(nrow(Insurance_Data_Group16)), size = 30)

train <- Insurance_Data_Group16[-index,]
test <- Insurance_Data_Group16[index,]

summary(train)
```

```
##     Charges           Age             BMI            Female
##  Min.   : 1256   Min.   :18.00   Min.   :17.67   Min.   :0.0000
##  1st Qu.: 5593   1st Qu.:26.00   1st Qu.:25.86   1st Qu.:0.0000
##  Median : 8692   Median :42.00   Median :29.16   Median :0.0000
##  Mean   :13786   Mean   :39.99   Mean   :30.76   Mean   :0.4429
##  3rd Qu.:20281   3rd Qu.:51.50   3rd Qu.:35.67   3rd Qu.:1.0000
##  Max.   :47270   Max.   :62.00   Max.   :47.60   Max.   :1.0000
##     Children          Smoker       WinterSprings      WinterPark
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.9429   Mean   :0.1857   Mean   :0.1571   Mean   :0.2857
##  3rd Qu.:2.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :5.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      Oviedo
##  Min.   :0.0
##  1st Qu.:0.0
##  Median :0.0
##  Mean   :0.3
##  3rd Qu.:1.0
##  Max.   :1.0
```

Question 2

Provide the correlation between all quantitative variables

```
quantitative_vars <- train[, c("Charges", "Age", "BMI", "Children")]

correlation_matrix <- cor(quantitative_vars, use = "complete.obs")
correlation_matrix
```

```
##              Charges         Age         BMI    Children
## Charges   1.0000000  0.36696097  0.26854917  0.17297484
## Age       0.3669610  1.00000000  0.17056536 -0.04066037
## BMI       0.2685492  0.17056536  1.00000000 -0.06794895
## Children 0.1729748 -0.04066037 -0.06794895  1.00000000
```

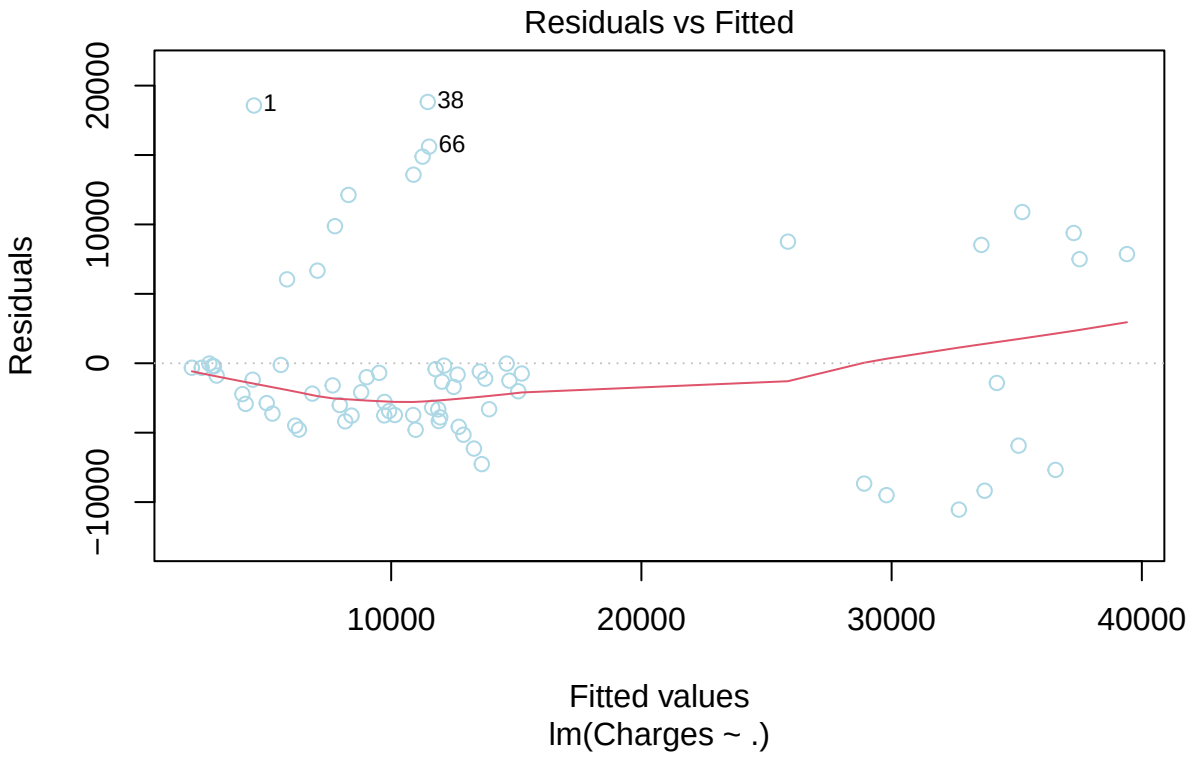
Question 3

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?

```
model <- lm(Charges ~ ., data = train)
model_summary <- summary(model)
print(model_summary)
```

```
##
## Call:
## lm(formula = Charges ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10543.3  -3751.2  -1505.9   -133.6  18822.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5343.87    4999.26  -1.069 0.289312
## Age             232.47      63.66   3.652 0.000543 ***
## BMI             122.65     144.05   0.851 0.397858
## Female         -534.28    1729.63  -0.309 0.758450
## Children       1035.76     743.06   1.394 0.168400
## Smoker        23206.03    2359.09   9.837 3.33e-14 ***
## WinterSprings  -302.37    2783.21  -0.109 0.913843
## WinterPark     2255.73    2412.46   0.935 0.353459
## Oviedo         1381.77    2421.86   0.571 0.570408
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7060 on 61 degrees of freedom
## Multiple R-squared:  0.7085, Adjusted R-squared:  0.6702
## F-statistic: 18.53 on 8 and 61 DF,  p-value: 9.974e-14
```

```
# Linearity and Homoscedasticity (Residuals vs Fitted Plot)
plot(model, which = 1, col = "lightblue")
```



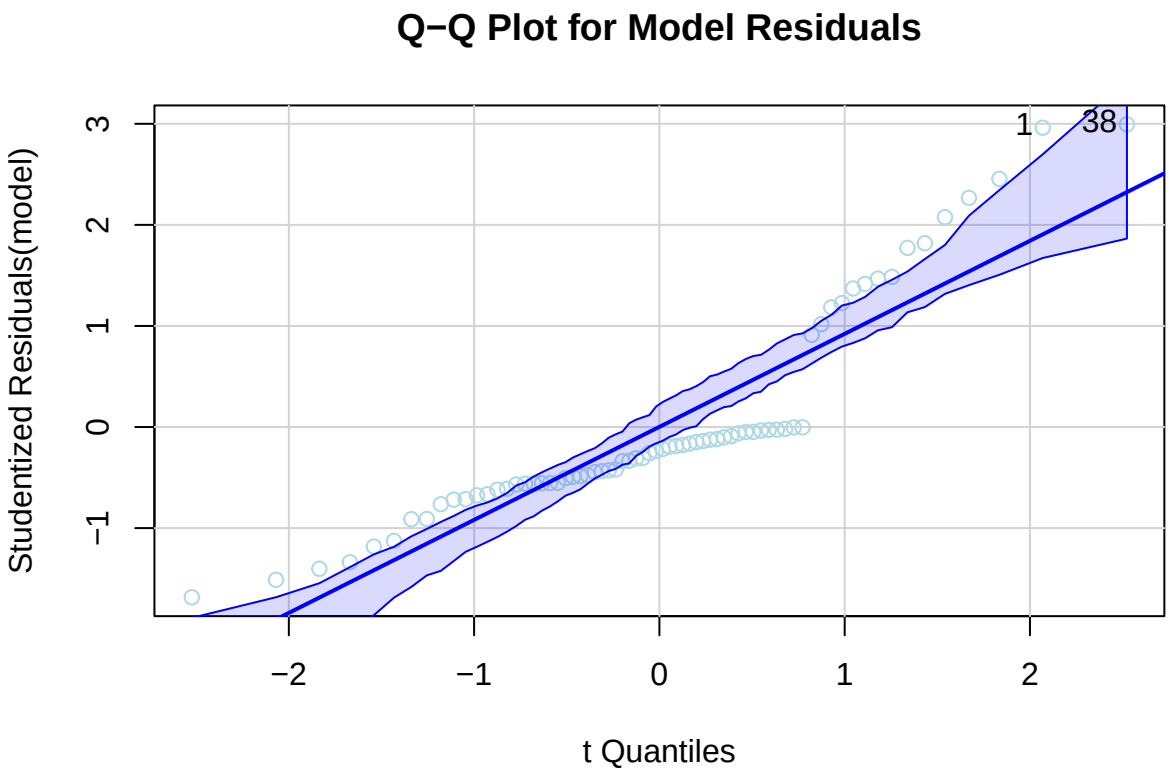

```
# Independence (Durbin-Watson Test)
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 2.0828, p-value = 0.6532
## alternative hypothesis: true autocorrelation is greater than 0
```

```
# Multicollinearity (Variance Inflation Factors)
vif_values <- vif(model)
print(vif_values)
```

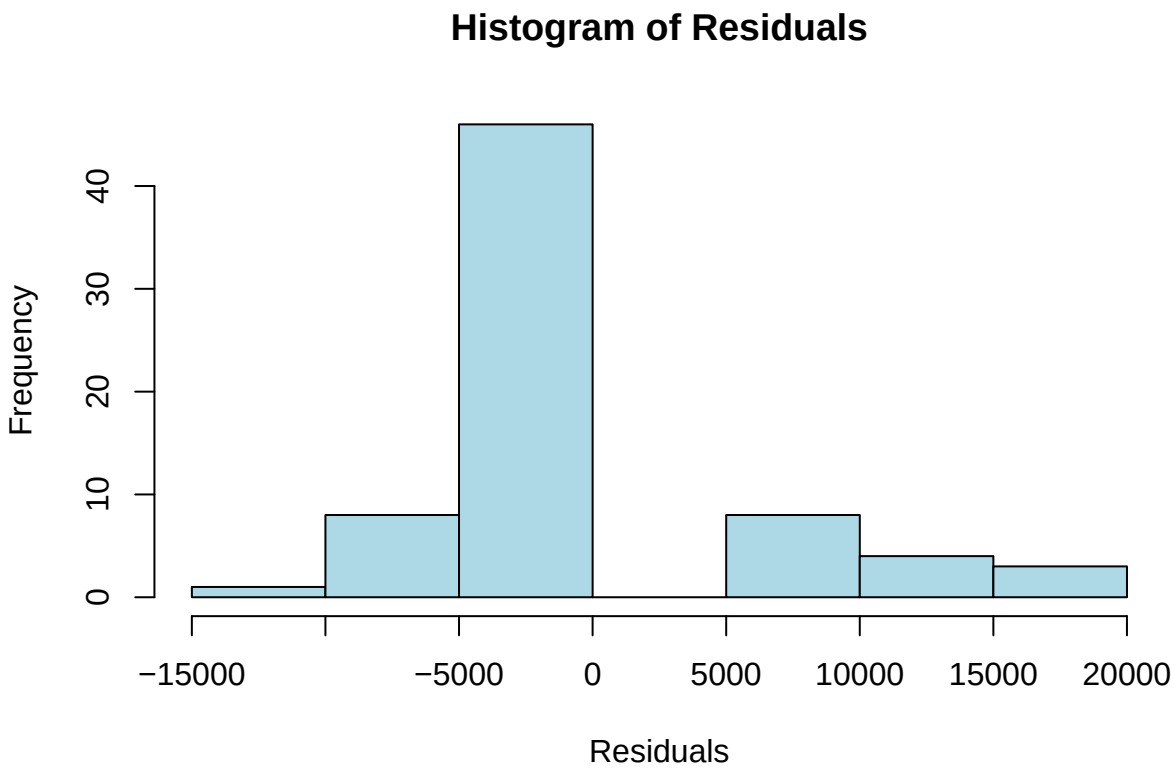
```
##           Age           BMI        Female      Children        Smoker
##      1.112658      1.185349      1.036640      1.038754      1.181960
## WinterSprings    WinterPark        Oviedo
##      1.440891      1.668075      1.729853
```

```
# Normal Distribution of Errors (Q-Q Plot)
qqPlot(model, main = "Q-Q Plot for Model Residuals", col = "lightblue")
```




```
## [1]  1 38
```

```
# Histogram (Residuals)
hist(resid(model), main = "Histogram of Residuals", xlab = "Residuals", col = "lightblue")
```



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?

```
#
```

Question 5

Use the 30 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)

```
#
```

Question 6

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

```
#
```

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

```
#
```

Question 8

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

Question 9

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