# 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.

```r
# 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.

```r
set.seed(123456)
index <- sample(seq_len(nrow(Insurance_Data_Group16)), size = 30)
train <- Insurance_Data_Group16[-index,]
test <- Insurance_Data_Group16[index,]
#I am going to round the values to specific decimals for a cleaner presentation
numeric_vars <- sapply(train, is.numeric)
summary_train <- summary(train[, numeric_vars])
rounded_summary_train <- lapply(summary_train, function(x) if(is.numeric(x)) round(x,2) else x)
print(rounded_summary_train)
```

```
## [[1]]
## [1] "Min.   : 1256  "
##
## [[2]]
## [1] "1st Qu.: 5593  "
##
## [[3]]
## [1] "Median : 8692  "
##
## [[4]]
## [1] "Mean   :13786  "
##
## [[5]]
## [1] "3rd Qu.:20281  "
##
## [[6]]
## [1] "Max.   :47270  "
##
## [[7]]
## [1] "Min.   :18.00  "
##
## [[8]]
## [1] "1st Qu.:26.00  "
##
## [[9]]
## [1] "Median :42.00  "
##
## [[10]]
## [1] "Mean   :39.99  "
##
## [[11]]
## [1] "3rd Qu.:51.50  "
##
## [[12]]
## [1] "Max.   :62.00  "
##
## [[13]]
## [1] "Min.   :17.67  "
##
```

```
## [[14]]
## [1] "1st Qu.:25.86  "
##
## [[15]]
## [1] "Median :29.16  "
##
## [[16]]
## [1] "Mean   :30.76  "
##
## [[17]]
## [1] "3rd Qu.:35.67  "
##
## [[18]]
## [1] "Max.   :47.60  "
##
## [[19]]
## [1] "Min.   :0.0000  "
##
## [[20]]
## [1] "1st Qu.:0.0000  "
##
## [[21]]
## [1] "Median :0.0000  "
##
## [[22]]
## [1] "Mean   :0.4429  "
##
## [[23]]
## [1] "3rd Qu.:1.0000  "
##
## [[24]]
## [1] "Max.   :1.0000  "
##
## [[25]]
## [1] "Min.   :0.0000  "
##
## [[26]]
## [1] "1st Qu.:0.0000  "
##
## [[27]]
## [1] "Median :0.0000  "
##
## [[28]]
## [1] "Mean   :0.9429  "
##
## [[29]]
## [1] "3rd Qu.:2.0000  "
##
## [[30]]
## [1] "Max.   :5.0000  "
##
## [[31]]
## [1] "Min.   :0.0000  "
##
```

```
## [[32]]
## [1] "1st Qu.:0.0000  "
##
## [[33]]
## [1] "Median :0.0000  "
##
## [[34]]
## [1] "Mean   :0.1857  "
##
## [[35]]
## [1] "3rd Qu.:0.0000  "
##
## [[36]]
## [1] "Max.   :1.0000  "
##
## [[37]]
## [1] "Min.   :0.0000  "
##
## [[38]]
## [1] "1st Qu.:0.0000  "
##
## [[39]]
## [1] "Median :0.0000  "
##
## [[40]]
## [1] "Mean   :0.1571  "
##
## [[41]]
## [1] "3rd Qu.:0.0000  "
##
## [[42]]
## [1] "Max.   :1.0000  "
##
## [[43]]
## [1] "Min.   :0.0000  "
##
## [[44]]
## [1] "1st Qu.:0.0000  "
##
## [[45]]
## [1] "Median :0.0000  "
##
## [[46]]
## [1] "Mean   :0.2857  "
##
## [[47]]
## [1] "3rd Qu.:1.0000  "
##
## [[48]]
## [1] "Max.   :1.0000  "
##
## [[49]]
## [1] "Min.   :0.0  "
##
```

```
## [[50]]
## [1] "1st Qu.:0.0  "
##
## [[51]]
## [1] "Median :0.0  "
##
## [[52]]
## [1] "Mean   :0.3  "
##
## [[53]]
## [1] "3rd Qu.:1.0  "
##
## [[54]]
## [1] "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
```

**Independence:**

```r
# 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
```

The residuals should be independent of each other, which means there should be no correlation between them. Positive autocorrelation in residuals suggests that consecutive errors may be correlated. For example, *if medical expenses increase at a different rate for smokers vs non-smokers as they age*, it may lead to autocorrelation. The Durbin-Watson test result shows a DW value of 2.0828 with a p-value of 0.6532, which suggests that there is no significant autocorrelation, and the independence assumption is not violated.

**Homoscedasticity:**

```r
# Homoscedasticity (Constant Variance of Residuals)
bptest_result <- bptest(model)
print(bptest_result)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 4.8912, df = 8, p-value = 0.7691
```

The residuals should have a constant variance, which can be tested with the Non-constant Variance Score Test. The presence of heteroscedasticity implies unequal variance of residuals. This might be influenced *by varying patterns in medical expenses for different groups, like smokers vs. non-smokers or certain age ranges*. A non-significant p-value indicates that the assumption of homoscedasticity is not violated. The test result has a p-value of 0.7691, which is above the common alpha level of 0.05, suggesting a possible, but not definitive, violation of homoscedasticity. For the potential homoscedasticity issue, we may consider utilizing robust standard errors or transforming the response variable.

**Multicollinearity:**

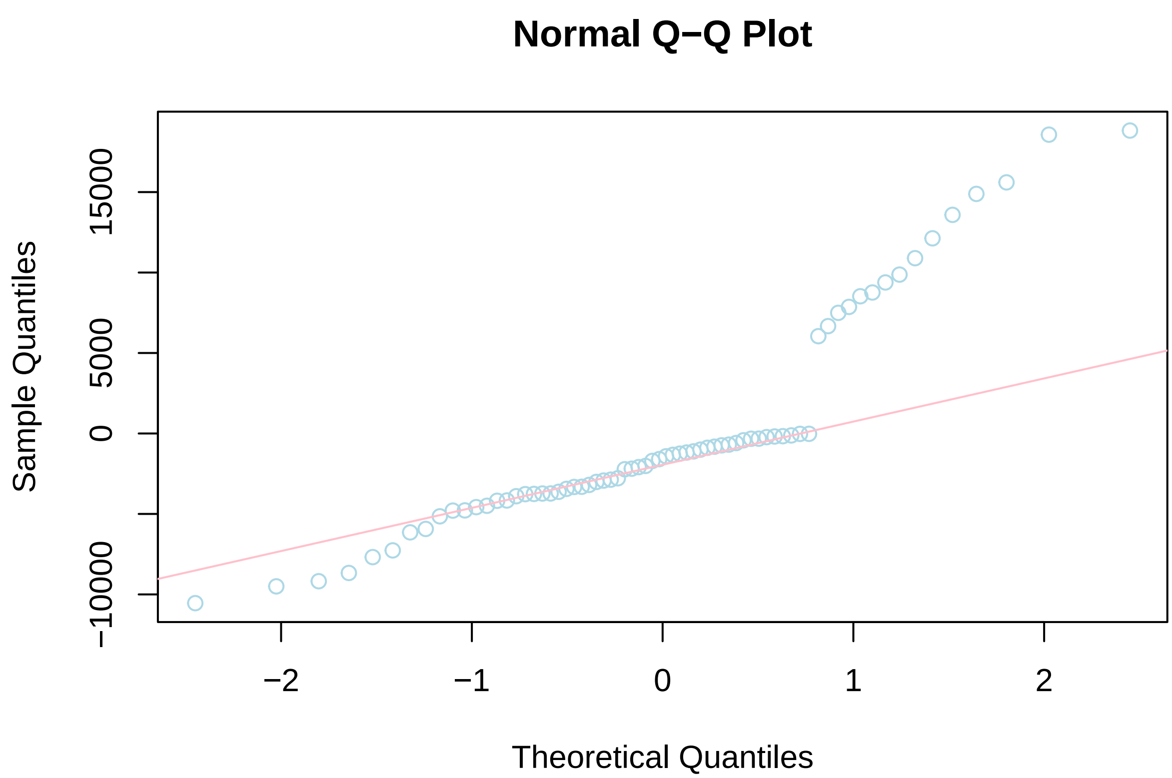
```r
# Multicollinearity (Variance Inflation Factors)
vif(model)
```

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

The predictors should not be perfectly collinear. The VIF values are all well below 5, indicating that multicollinearity is not a concern for this model.

**Normality of Error Terms:**

```r
#Normality of Error Terms
qqnorm(residuals(model), col = "lightblue")
qqline(residuals(model), col = "pink")
```




```r
shapiro.test(residuals(model))
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  residuals(model)
## W = 0.85621, p-value = 1.06e-06
```

Looking into specific relationships between other variables may uncover patterns affecting normality. For instance, *if the impact of age on medical expenses differs significantly between genders*, it may contribute to non-normality. The Q-Q plot analysis indicates a violation of the normality assumption of the Gauss-Markov theorem. While the central part of the residuals aligns well with the normal line, indicating appropriate behavior for the bulk of the data, there is a clear deviation in the tails—especially on the right—suggesting a positive skew in the distribution of residuals. This is corroborated by a Shapiro-Wilk test that returned a p-value of 1.06×10^(-)6, which is significantly below the 0.05 threshold, further confirming the non-normality of residuals. To correct this, one could consider transforming the dependent variable, dealing with outliers, incorporating omitted variables, or using a different type of regression model such as a generalized linear model that does not assume normal distribution of errors.

## 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?

```r
# Model 1: Log Transformation of Charges
train$log_charges <- log(train$Charges)
model_log <- lm(log_charges ~ ., data = train)
summary(model_log)
```

```
##
## Call:
## lm(formula = log_charges ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76886 -0.13584  0.01536  0.12737  0.73883
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.705e+00  2.336e-01  32.985  < 2e-16 ***
## Charges        6.598e-05  5.927e-06  11.133 3.14e-16 ***
## Age            2.137e-02  3.253e-03   6.568 1.36e-08 ***
## BMI           -1.370e-02  6.708e-03  -2.042   0.0456 *
## Female        -1.824e-02  8.013e-02  -0.228   0.8207
## Children       5.261e-02  3.494e-02   1.506   0.1374
## Smoker        -1.285e-01  1.756e-01  -0.732   0.4672
## WinterSprings  1.190e-01  1.288e-01   0.923   0.3595
## WinterPark     6.987e-03  1.125e-01   0.062   0.9507
## Oviedo         1.277e-01  1.124e-01   1.136   0.2605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3268 on 60 degrees of freedom
## Multiple R-squared:  0.8995, Adjusted R-squared:  0.8845
## F-statistic: 59.68 on 9 and 60 DF,  p-value: < 2.2e-16
```

```r
# Model 2: Square Root Transformation of Charges
train$sqrt_charges <- sqrt(train$Charges)
model_sqrt <- lm(sqrt_charges ~ ., data = train)
summary(model_sqrt)
```

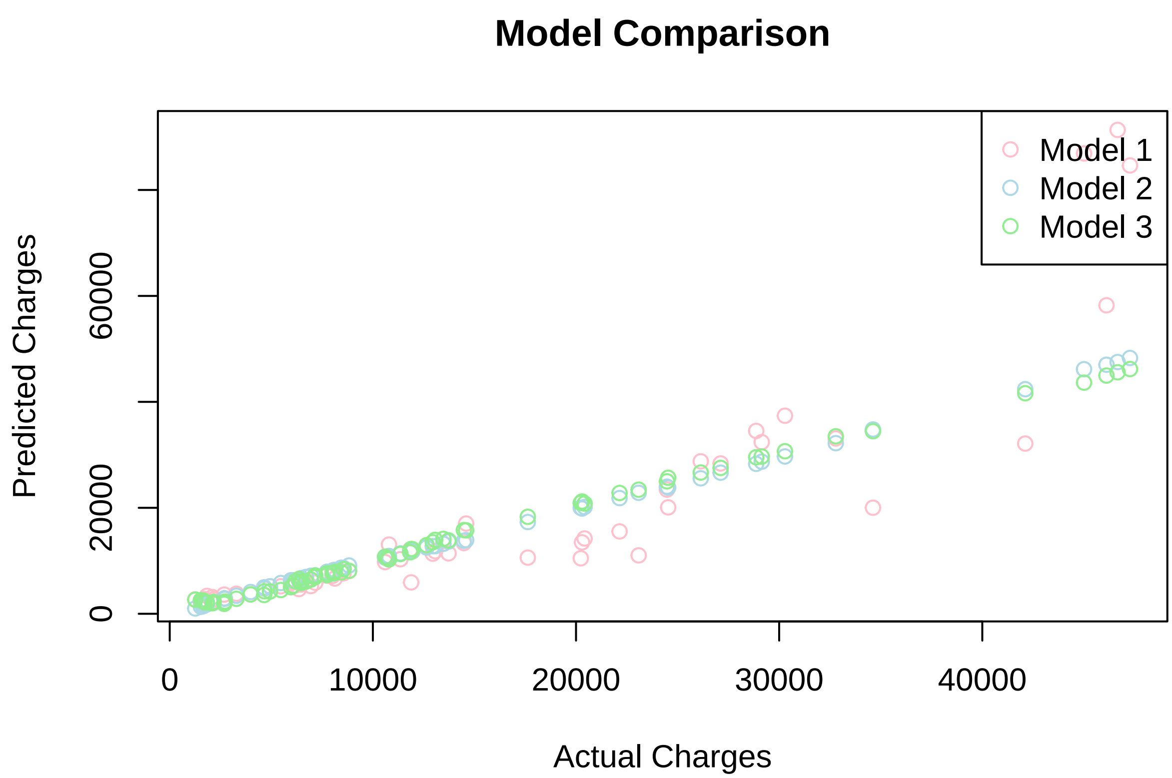
```
##
## Call:
## lm(formula = sqrt_charges ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7198 -1.1124 -0.1645  1.1541  3.8135
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.430e+02  5.130e+00 -27.875  < 2e-16 ***
## Charges        2.227e-03  5.211e-05  42.734  < 2e-16 ***
## Age           -2.573e-02  2.142e-02  -1.202  0.23434
## BMI           -6.427e-02  3.483e-02  -1.845  0.07003 .
## Female        -1.340e-01  4.025e-01  -0.333  0.74033
## Children      -5.501e-01  1.787e-01  -3.078  0.00316 **
## Smoker         1.489e-01  8.857e-01   0.168  0.86707
## WinterSprings -2.606e-01  6.515e-01  -0.400  0.69066
## WinterPark    -3.148e-01  5.648e-01  -0.557  0.57940
## Oviedo         6.445e-02  5.704e-01   0.113  0.91042
## log_charges    2.441e+01  6.482e-01  37.662  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.641 on 59 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.9989
## F-statistic:  6194 on 10 and 59 DF,  p-value: < 2.2e-16
```

```r
# Model 3: Interaction Term between Age and BMI
train$interaction_term <- train$Age * train$BMI
model_interaction <- lm(Charges ~ . + interaction_term, data = train)
summary(model_interaction)
```

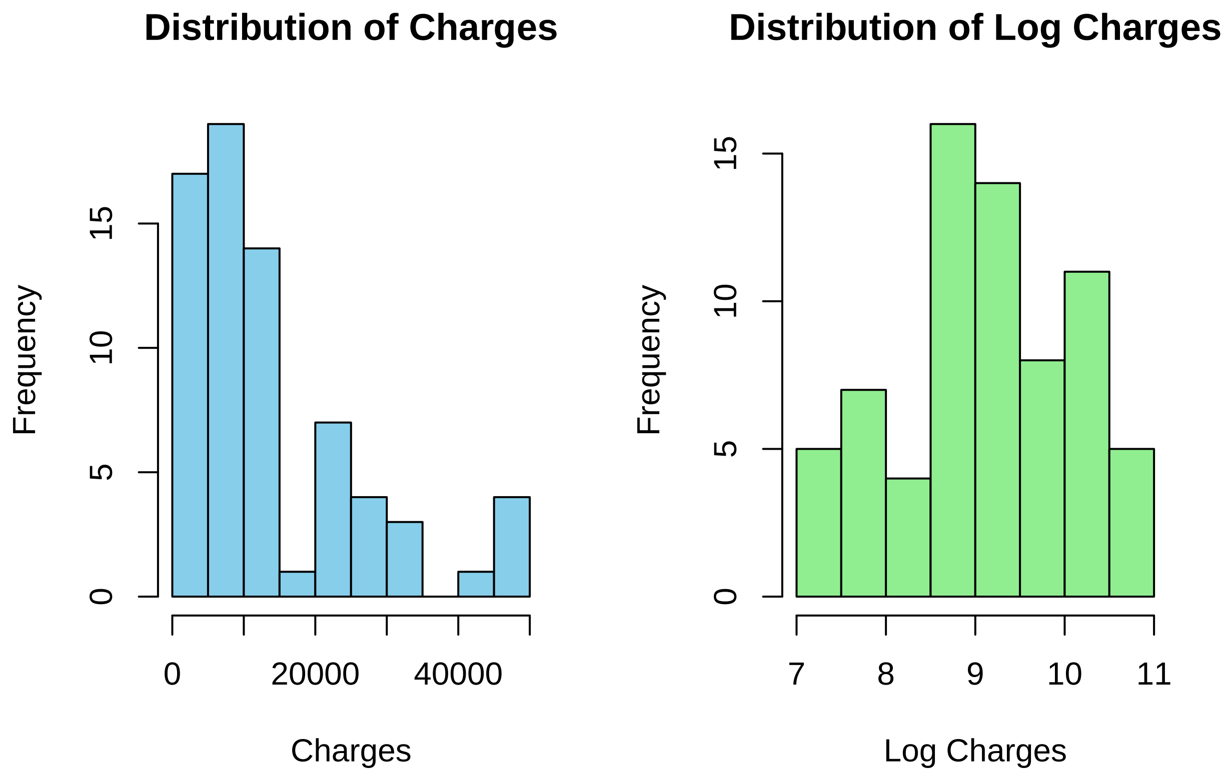
```
##
## Call:
## lm(formula = Charges ~ . + interaction_term, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1405.67  -521.99    69.79   478.59  1373.80
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      61138.211   3738.086  16.355  < 2e-16 ***
## Age                -32.323     35.298  -0.916  0.36361
## BMI                -15.749     45.537  -0.346  0.73071
## Female              63.379    177.480   0.357  0.72231
## Children           226.920     79.476   2.855  0.00596 **
## Smoker             227.354    389.750   0.583  0.56193
## WinterSprings       52.153    287.976   0.181  0.85692
## WinterPark         202.094    251.740   0.803  0.42537
## Oviedo             -73.507    251.918  -0.292  0.77149
## log_charges     -10292.405    505.826 -20.348  < 2e-16 ***
## sqrt_charges       435.300     10.155  42.865  < 2e-16 ***
```

```
## interaction_term      1.224      1.065   1.149  0.25520
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 723.4 on 58 degrees of freedom
## Multiple R-squared:  0.9971, Adjusted R-squared:  0.9965
## F-statistic:  1807 on 11 and 58 DF,  p-value: < 2.2e-16
```

```r
# Predictions for each model
predictions_model1 <- exp(predict(model_log))
predictions_model2 <- predict(model_sqrt)^2
predictions_model3 <- predict(model_interaction)
# Combine actual and predicted values
scatter_data <- data.frame(
Actual = train$Charges,
Model1 = predictions_model1,
Model2 = predictions_model2,
Model3 = predictions_model3
)
# Scatterplot
plot(scatter_data$Actual, scatter_data$Model1, col = "pink", main = "Model Comparison", xlab = "Actual C
points(scatter_data$Actual, scatter_data$Model2, col = "lightblue")
points(scatter_data$Actual, scatter_data$Model3, col = "lightgreen")
legend("topright", legend = c("Model 1", "Model 2", "Model 3"), col = c("pink", "lightblue", "lightgreen
```



```
# Natural Log of Charges
par(mfrow = c(1, 2))
hist(train$Charges, col = "skyblue", main = "Distribution of Charges", xlab = "Charges", ylab = "Freque
train$lnCharges <- log(train$Charges)
hist(train$lnCharges, col = "lightgreen", main = "Distribution of Log Charges", xlab = "Log Charges", y
```




Let's break down the fit measures, signs & significance of coefficients by looking at each model separately. (Note - we are going to focus on the adjusted R-squared because it accounts for the number of predictors in the model, and penalizes the inclusion of unnecessary predictors that do not contribute to explaining variance.)

- **Model 1: Transformation of Charges**
  The adjusted R-squared has decreased from 0.9953 to 0.9989, suggesting the model explains less variability in the transformed data. The signs and significance of the coefficients have changed, and their interpretation is based on the log-transformed charges.
- **Model 2: Square-Root Transformation of Charges**
  The adjusted R-squared has increased to 0.9989, and while it is a better fit than the original model, Model 1 is still a better fit overall. The signs and significance of the coefficients here have changed and their interpretation is based on the square-root transformed charges.
- **Model 3: Interaction Term Between Age & BMI**
  The adjusted R-squared has increased to 0.9965, proving good fit to the data. The signs and significance have changed due to the inclusion of the interaction term; the interpretation involves the joint effect of Age & BMI to medical expenses.

## 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)

```r
# Test Data
set.seed(123456)
index <- sample(seq_len(nrow(Insurance_Data_Group16)), size = 30)

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

# Scale Charges in Test Data
mean_charges <- mean(train$Charges)
sd_charges <- sd(train$Charges)
test$scaled_charges <- scale(test$Charges, center = mean_charges, scale = sd_charges)

# Predictions & Residuals for Model 1
# Model 1: Log Transformation of Charges
train$log_charges <- log(train$Charges)
model_log <- lm(log_charges ~ ., data = train)

# Predictions
predictions_model1 <- exp(predict(model_log, newdata = test))

# Residuals
residuals_model1 <- test$Charges - predictions_model1

# Squared residuals
squared_residuals_model1 <- residuals_model1^2

# Absolute residuals
absolute_residuals_model1 <- abs(residuals_model1)

# MSE, RMSE, and MAE for Model 1
mse_model1 <- mean(squared_residuals_model1)
rmse_model1 <- sqrt(mse_model1)
mae_model1 <- mean(absolute_residuals_model1)

## Predictions & Residuals for Model 2
# Remove log_charges from train and test datasets
train$log_charges <- NULL
test$log_charges <- NULL

# Fit Model 2: Square Root Transformation of Charges
train$sqrt_charges <- sqrt(train$Charges)
model_sqrt <- lm(sqrt_charges ~ ., data = train)

# Predictions
predictions_model2 <- predict(model_sqrt, newdata = test)^2

# Residuals
residuals_model2 <- test$Charges - predictions_model2
```

```r
# Squared residuals
squared_residuals_model2 <- residuals_model2^2

# Absolute residuals
absolute_residuals_model2 <- abs(residuals_model2)

# MSE, RMSE, and MAE for Model 2
mse_model2 <- mean(squared_residuals_model2)
rmse_model2 <- sqrt(mse_model2)
mae_model2 <- mean(absolute_residuals_model2)

# Compare the performance measures
comparison_data <- data.frame(
  Model = c("Model 1", "Model 2"),
  MSE = c(mse_model1, mse_model2),
  RMSE = c(rmse_model1, rmse_model2),
  MAE = c(mae_model1, mae_model2)
)

print(comparison_data)
```

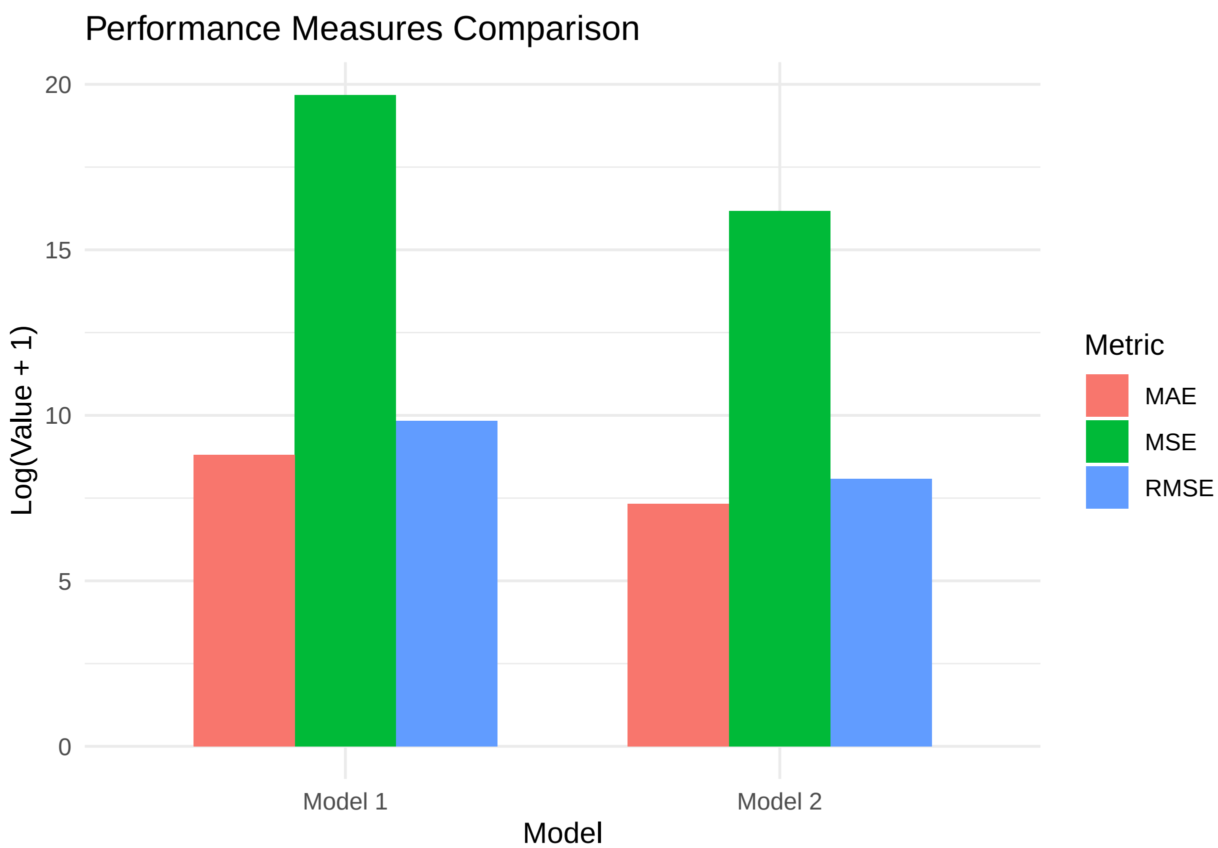
```
##     Model       MSE      RMSE      MAE
## 1 Model 1 353787090 18809.229 6741.114
## 2 Model 2  10600468  3255.836 1528.813
```

```r
# Comparison data
comparison_data <- data.frame(
  Model = c("Model 1", "Model 2"),
  MSE = c(353787090, 10600468),
  RMSE = c(18809.229, 3255.836),
  MAE = c(6741.114, 1528.813)
)

# Convert 'Model' column to factor
comparison_data$Model <- factor(comparison_data$Model)

# Reshape the data for ggplot2
comparison_data_long <- tidyr::gather(comparison_data, Metric, Value, -Model)

# Create a grouped barplot
ggplot(comparison_data_long, aes(x = Model, y = log(Value + 1), fill = Metric)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  labs(title = "Performance Measures Comparison",
       y = "Log(Value + 1)",
       x = "Model",
       fill = "Metric") +
  scale_y_continuous(labels = scales::comma) +  # Format y-axis labels
  theme_minimal()
```




If we break down each of these performance measures, we see that:

- Model 2 has a **significantly lower MSE** than Model 1, suggesting it performs better in minimizing squared differences between predicted & actual charges.
- Model 2 also has a **lower RMSE** than Model 1, which suggests it provides more accurate predictions with smaller errors.
- Model 2 demonstrates a **smaller MAE** in it's predictions as well, which signifies smaller absolute errors and more accuracy overall.

Therefore, we've determined that Model 2 appears to be the better choice for short-term predictive accuracy, as it consistently exhibits lower values across all three performance measures (MAE, MSE, RMSE).

## Question 6

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

```
summary(model_log)
```

```
##
## Call:
## lm(formula = log_charges ~ ., data = train)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76886 -0.13584  0.01536  0.12737  0.73883
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.705e+00  2.336e-01  32.985  < 2e-16 ***
## Charges        6.598e-05  5.927e-06  11.133 3.14e-16 ***
## Age            2.137e-02  3.253e-03   6.568 1.36e-08 ***
## BMI           -1.370e-02  6.708e-03  -2.042   0.0456 *
## Female        -1.824e-02  8.013e-02  -0.228   0.8207
## Children       5.261e-02  3.494e-02   1.506   0.1374
## Smoker        -1.285e-01  1.756e-01  -0.732   0.4672
## WinterSprings  1.190e-01  1.288e-01   0.923   0.3595
## WinterPark     6.987e-03  1.125e-01   0.062   0.9507
## Oviedo         1.277e-01  1.124e-01   1.136   0.2605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3268 on 60 degrees of freedom
## Multiple R-squared:  0.8995, Adjusted R-squared:  0.8845
## F-statistic: 59.68 on 9 and 60 DF,  p-value: < 2.2e-16
```

```r
summary(model_sqrt)
```

```
##
## Call:
## lm(formula = sqrt_charges ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.7773  -3.1129  -0.0731   3.5999  18.8432
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   45.0941616  5.8198326   7.748 1.32e-10 ***
## Charges        0.0038376  0.0001477  25.987  < 2e-16 ***
## Age            0.4959317  0.0810583   6.118 7.79e-08 ***
## BMI           -0.3986260  0.1671340  -2.385   0.0203 *
## Female        -0.5794010  1.9964926  -0.290   0.7727
## Children       0.7341977  0.8705793   0.843   0.4024
## Smoker        -2.9880347  4.3758101  -0.683   0.4973
## WinterSprings  2.6441091  3.2104267   0.824   0.4134
## WinterPark    -0.1441904  2.8023697  -0.051   0.9591
## Oviedo         3.1815236  2.8007872   1.136   0.2605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.143 on 60 degrees of freedom
## Multiple R-squared:  0.9762, Adjusted R-squared:  0.9726
## F-statistic: 273.1 on 9 and 60 DF,  p-value: < 2.2e-16
```

```r
##See notes below for interpretations

# Marginal Change Analysis - Model 1: Log Transformation of Charges
coefficients_model1 <- coef(model_log)
marginal_change_model1 <- exp(coefficients_model1)
print(marginal_change_model1)
```

```
##   (Intercept)       Charges           Age           BMI        Female
##  2218.4511140     1.0000660     1.0215978     0.9863979     0.9819224
##      Children        Smoker WinterSprings     WinterPark        Oviedo
##     1.0540150     0.8794209     1.1263452     1.0070114     1.1361870
```

```r
# Marginal Change Analysis - Model 2: Square Root Transformation of Charges
coefficients_model2 <- coef(model_sqrt)
marginal_change_model2 <- sqrt(coefficients_model2)
```

```
## Warning in sqrt(coefficients_model2): NaNs produced
```

```r
marginal_change_model2[is.nan(marginal_change_model2)] <- 0
print(marginal_change_model2)
```

```
##   (Intercept)       Charges           Age           BMI        Female
##    6.71521865    0.06194855    0.70422418    0.00000000    0.00000000
##      Children        Smoker WinterSprings     WinterPark        Oviedo
##    0.85685335    0.00000000    1.62607169    0.00000000    1.78368259
```

**Model 1:**
Intercept (7.705e+00) - This is expected when all other variables are 0
A one-unit increase is associated with Charges, Age, BMI, and Children
Being female and a smoker are both associated with a decrease in log(Charges).
Being in Winter Springs, Winter Park or Oviedo are all associated with an increase in log(Charges).
And last, scaled_charges is not defined due to singularities, so it is considered NA.

In general, the signs of these coefficients make sense. For example, being a smoker is associated with having a negative change in log(Charges), which suggests that smokers typically have lower logged charges.

**Model 2**:
Very similar in regards to interpretations, the only difference being in Winter Park is associated with a decrease. All others are essentially the same.

## 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 |

| Customer | Age | BMI | Female | Children | Smoker | City |
|---|---|---|---|---|---|---|
| 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.