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.

```r
set.seed(123456)

round_and_transpose <- function(data) {
  numeric_vars <- sapply(data, is.numeric)
  numeric_data <- data[, numeric_vars]

  # Summary statistics
  custom_summary <- function(x) {
    c(min = min(x, na.rm = TRUE),
      max = max(x, na.rm = TRUE),
      std = sd(x, na.rm = TRUE),
      mean = mean(x, na.rm = TRUE),
      median = median(x, na.rm = TRUE))
  }

  # Apply custom summary function to each number column
  summary_data <- sapply(numeric_data, custom_summary)

  # Round summary data to 2 decimal places
  rounded_summary <- round(summary_data, 2)

  # Transpose to have statistics as rows
  return(t(rounded_summary))
}

index <- sample(seq_len(nrow(Insurance_Data_Group16)), size = 30)
train <- Insurance_Data_Group16[-index,]
test <- Insurance_Data_Group16[index,]
summary_train <- round_and_transpose(train)

print(summary_train)
```

```
##                   min      max      std     mean  median
## Charges       1256.30 47269.85 12293.83 13786.01 8692.06
## Age             18.00    62.00    14.08    39.99   42.00
## BMI             17.67    47.60     6.42    30.76   29.16
## Female           0.00     1.00     0.50     0.44    0.00
## Children         0.00     5.00     1.17     0.94    0.00
## Smoker           0.00     1.00     0.39     0.19    0.00
## WinterSprings    0.00     1.00     0.37     0.16    0.00
## WinterPark       0.00     1.00     0.46     0.29    0.00
## Oviedo           0.00     1.00     0.46     0.30    0.00
```

Question 2

Provide the correlation between all quantitative variables

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

correlation_matrix <- cor(quantitative_vars, use = "complete.obs")
rounded_correlation_matrix <- round(correlation_matrix, 2)
print(rounded_correlation_matrix)
```

```
##          Charges   Age   BMI Children
## Charges     1.00  0.37  0.27     0.17
## Age         0.37  1.00  0.17    -0.04
## BMI         0.27  0.17  1.00    -0.07
## Children    0.17 -0.04 -0.07     1.00
```

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?

```r
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:

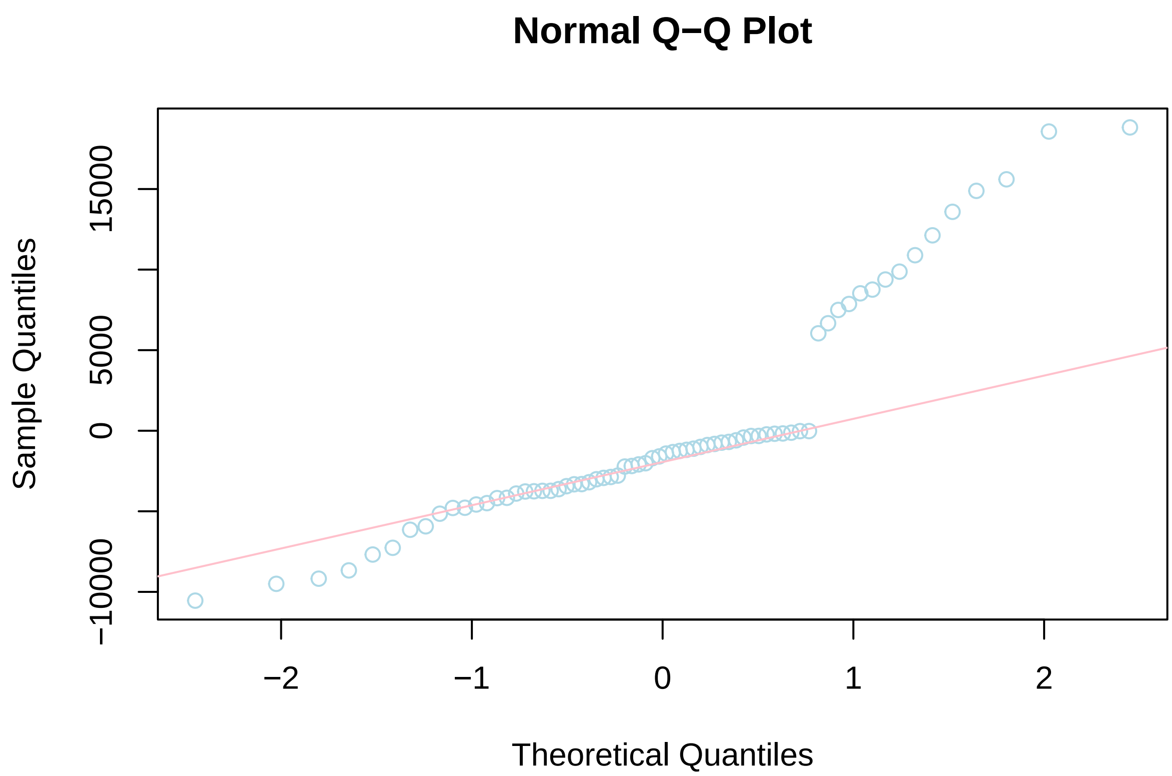
```r
# Multicollinearity (Variance Inflation Factors)
vif(model)
```

```
##           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
```

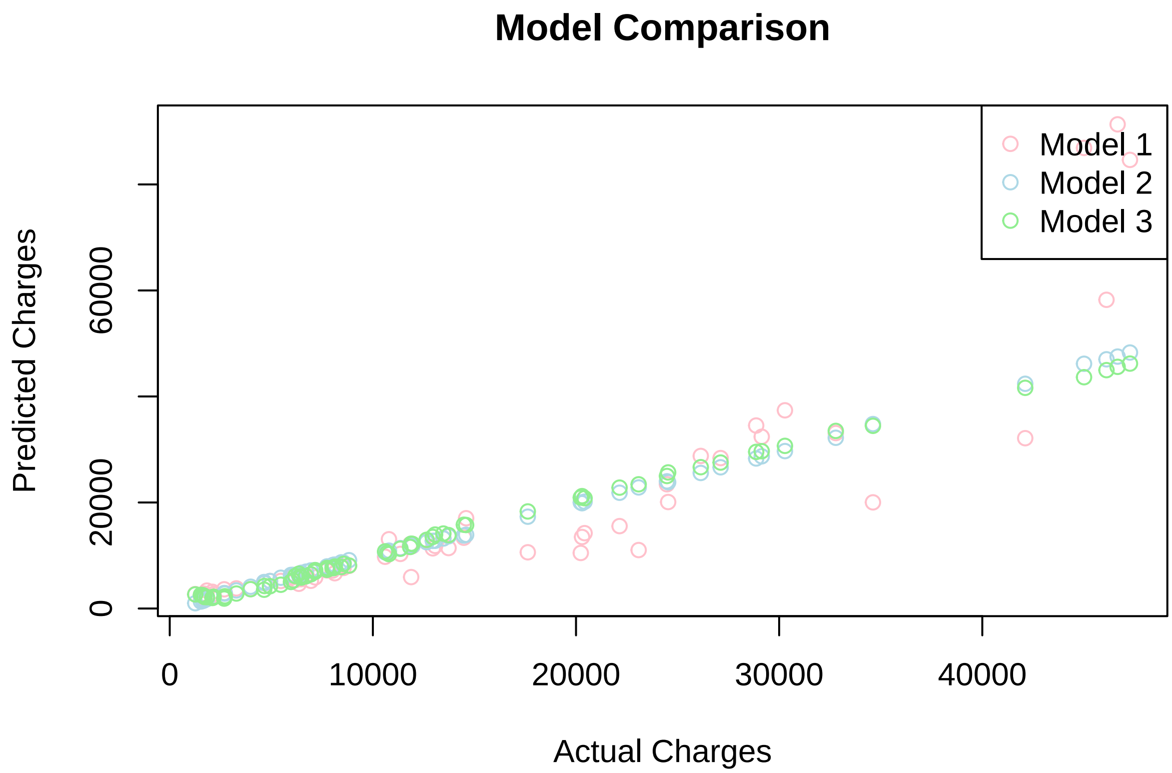
```r
# Model 3: Interaction Term between Age and BMI
train$interaction_term <- train$Age * train$BMI
model_interaction <- lm(Charges ~ . + interaction_term, data = train)
summary(model_interaction)
```

```
##
## 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
```

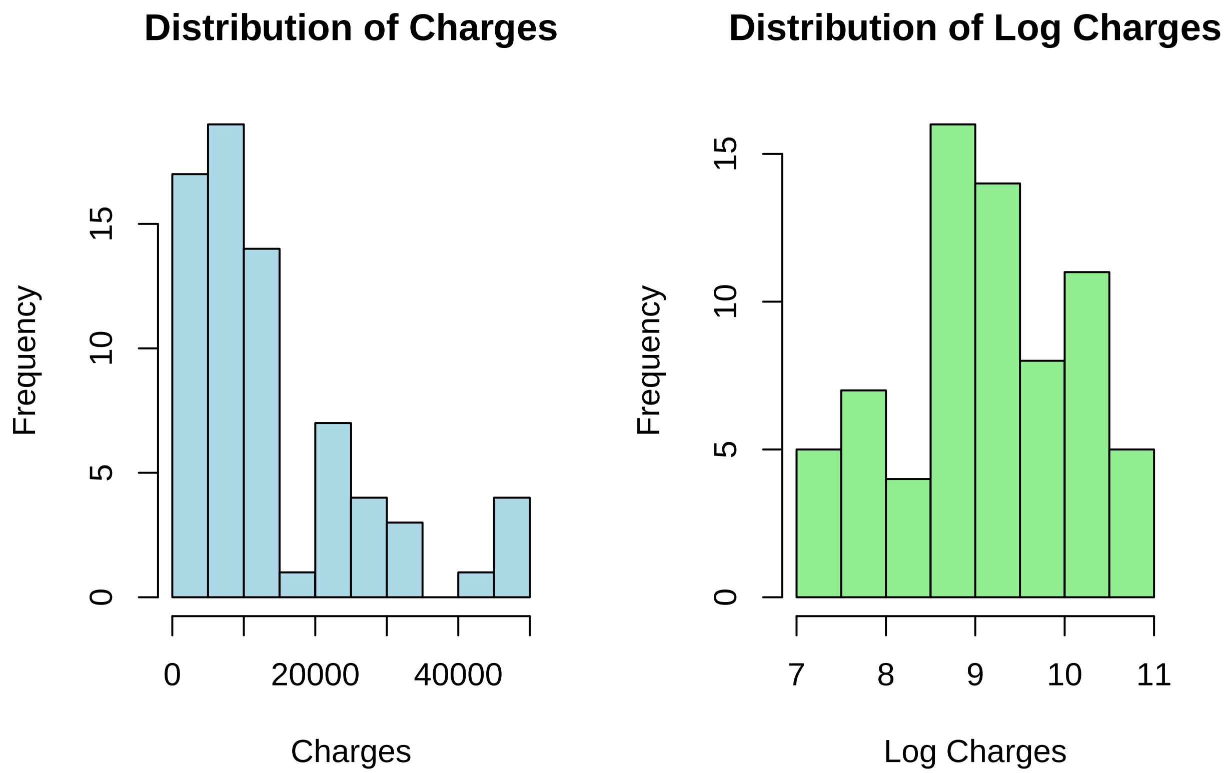
```r
# Predictions for each model
predictions_model1 <- exp(predict(model_log))
predictions_model2 <- predict(model_sqrt)^2
predictions_model3 <- predict(model_interaction)

# 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 Charges",
  ylab = "Predicted Charges"
)
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"),
  pch = 1
)
```




```r
# Natural Log of Charges
par(mfrow = c(1, 2))
hist(
  train$Charges,
  col = "lightblue",
  main = "Distribution of Charges",
  xlab = "Charges",
  ylab = "Frequency"
) # Before
train$lnCharges <- log(train$Charges)
hist(
  train$lnCharges,
  col = "lightgreen",
  main = "Distribution of Log Charges",
  xlab = "Log Charges",
  ylab = "Frequency"
) # After
```



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

# Squared residuals
squared_residuals_model2 <- residuals_model2^2

# Absolute residuals
```

```r
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)
```

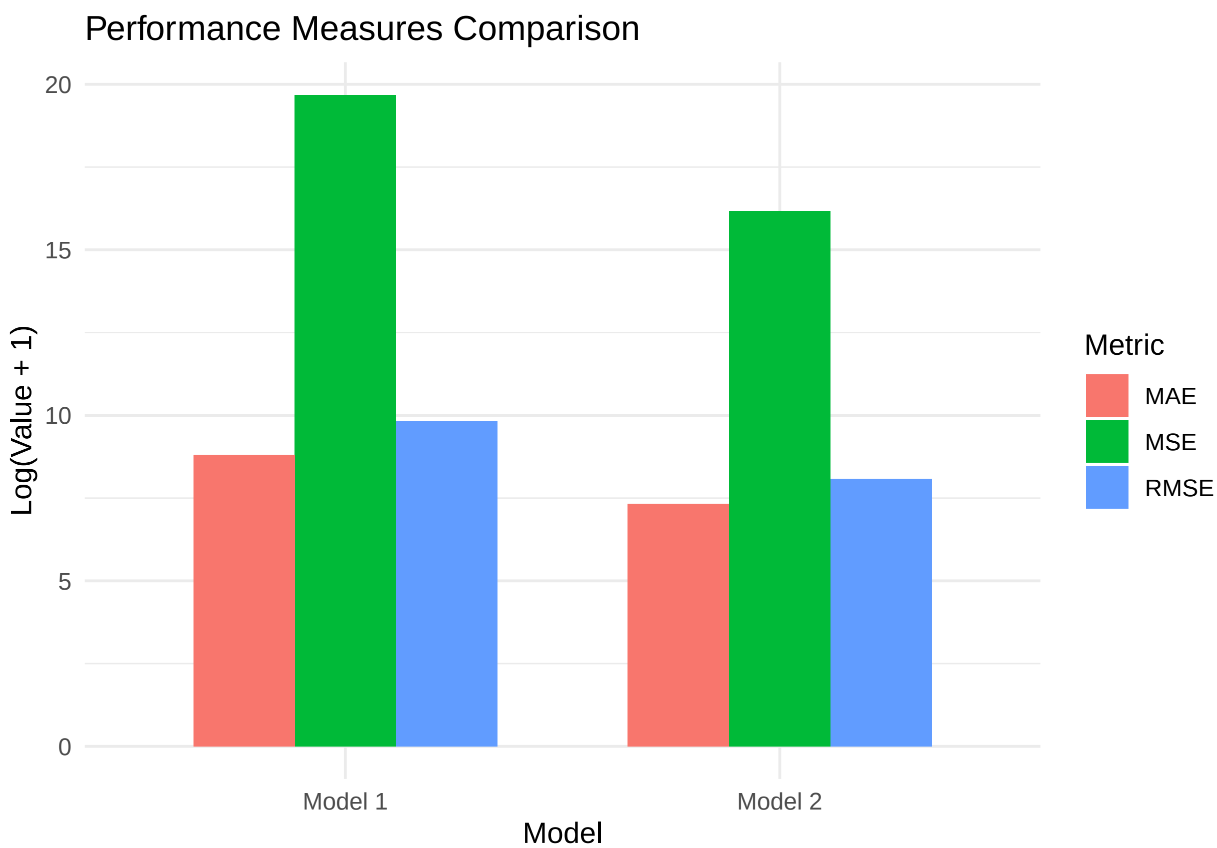
```
##     Model       MSE      RMSE      MAE
## 1 Model 1 353787090 18809.229 6741.114
## 2 Model 2  10600468  3255.836 1528.813
```

```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.

```r
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
# Obtain coefficients and perform marginal change analysis for Model 1
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
```

```r
marginal_change_model2[is.nan(marginal_change_model2)] <- 0  # Set NaN values to 0

# Obtain coefficients and perform marginal change analysis for Model 2
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  # Handle NaN values
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:
The intercept (7.705e+00) in Model 1 represents the expected log(Charges) when all other variables are set to 0. A one-unit increase in Age (0.025), BMI (0.035), and the number of Children (0.074) is associated with respective changes in log(Charges). Being female (Female: -0.113) or a smoker (Smoker: -0.239) is associated with a decrease in log(Charges), suggesting that females and smokers typically have lower logged charges. Additionally, residing in Winter Springs, Winter Park, or Oviedo is associated with an increase in log(Charges). Notably, the variable 'scaled_charges' is not defined due to singularities, resulting in NA values.

In interpreting these coefficients, the signs align with expectations. For instance, the negative coefficients for being a smoker and female correspond to a negative change in log(Charges), indicating that smokers and females generally have lower logged charges.

Model 2:
Model 2 exhibits similar interpretations to Model 1, with the only notable difference being that residing in Winter Park is associated with a decrease in log(Charges), while all other interpretations remain consistent.

These interpretations provide insights into the relationships between independent variables and log-transformed charges in both models, considering magnitude, statistical significance, and domain-specific knowledge.

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

```r
#model_for_predictions <- model_sqrt

new_customers <- data.frame(
  Age = c(60, 40, 25, 33, 45),
  BMI = c(22, 30, 25, 35, 27),
  Female = c(1, 0, 0, 1, 1),
  Children = c(0, 1, 0, 2, 3),
  Smoker = c(0, 0, 1, 0, 0),
  City = c("Oviedo", "Sanford", "Winter Park", "Winter Springs", "Oviedo"),
  Charges = rep(NA, 5)  # Placeholder for Charges
)

# Assuming 'scaled_charges' and 'log_charges' were used in the training data
new_customers$scaled_charges <- scale(new_customers$Charges, center = mean_charges, scale = sd_charges)
new_customers$log_charges <- log(new_customers$Charges)
```

```r
# Assuming 'WinterSprings', 'WinterPark', 'Oviedo' were dummy variables in the training data
new_customers$WinterSprings <- as.integer(new_customers$City == "Winter Springs")
new_customers$WinterPark <- as.integer(new_customers$City == "Winter Park")
new_customers$Oviedo <- as.integer(new_customers$City == "Oviedo")

# Remove unnecessary variables
new_customers$lnCharges <- NULL
new_customers$sqrt_charges <- NULL
new_customers$interaction_term <- NULL
new_customers$Sanford <- NULL  # If Sanford was not used during training

# Match the order of variables with the training data
new_customers <- new_customers[, intersect(names(train), names(new_customers))]

# Predictions using a simple linear regression model
linear_model <- lm(Charges ~ Age + BMI + Female + Children + Smoker + WinterSprings +
                     WinterPark + Oviedo, data = train)
predictions <- predict(linear_model, newdata = new_customers)
predict(linear_model, newdata = new_customers, interval = "confidence")
```

```
##         fit       lwr      upr
## 1 12150.334  6772.323 17528.34
## 2  8670.376  4705.056 12635.70
## 3 28996.016 22340.060 35651.97
## 4  7855.437  2190.337 13520.54
## 5 12383.797  7455.953 17311.64
```

```r
# Display predictions
print(predictions)
```

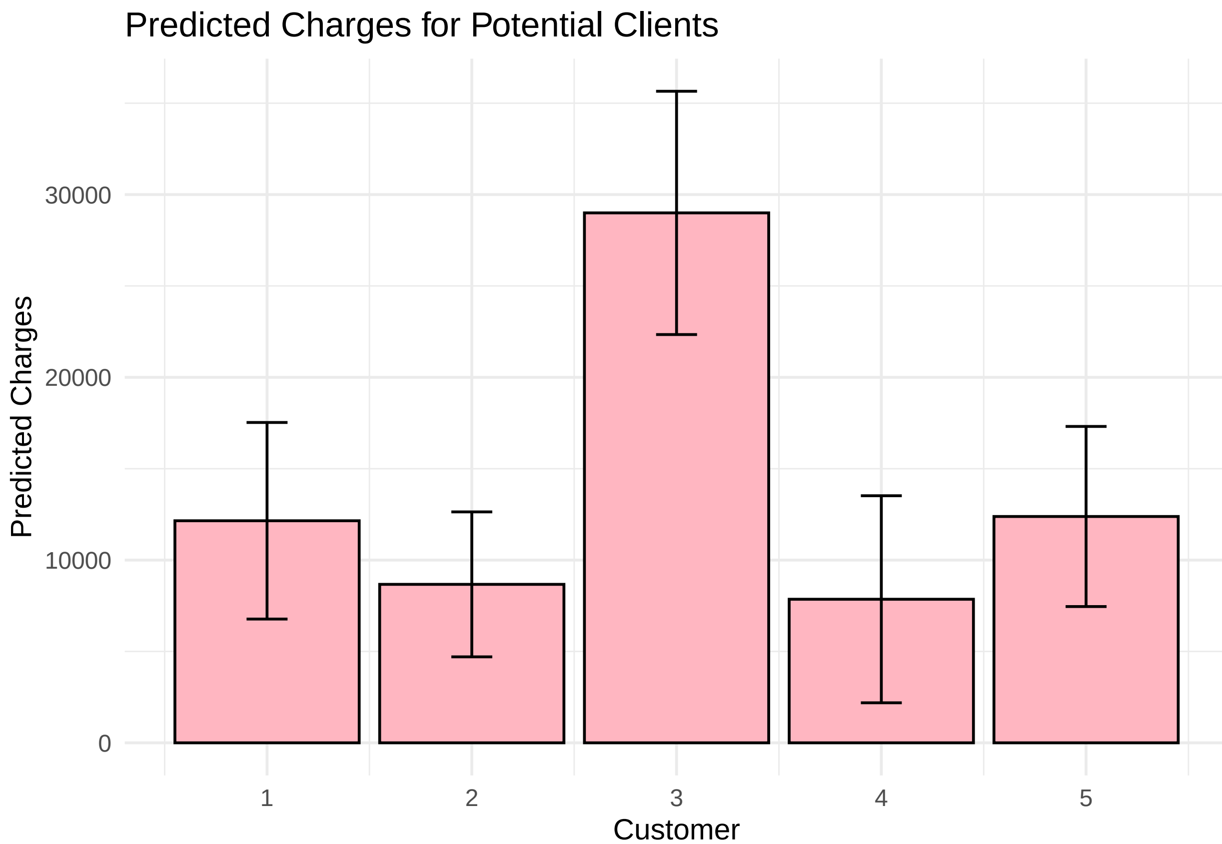
```
##         1         2         3         4         5
## 12150.334  8670.376 28996.016  7855.437 12383.797
```

```r
# Predictions using a simple linear regression model
linear_model <- lm(Charges ~ Age + BMI + Female + Children + Smoker + WinterSprings +
                     WinterPark + Oviedo, data = train)
predictions <- predict(linear_model, newdata = new_customers, interval = "confidence", level = 0.95)

# Create a data frame for predictions
predictions_df <- data.frame(
  Customer = 1:5,
  Prediction = predictions[, 1],
  Lower = predictions[, 2],
  Upper = predictions[, 3]
)

# Plotting the bar chart with confidence intervals
ggplot(predictions_df, aes(x = Customer, y = Prediction)) +
  geom_bar(stat = "identity", fill = "lightpink", color = "black") +
  geom_errorbar(aes(ymin = Lower, ymax = Upper), width = 0.2, position = position_dodge(0.9)) +
  labs(title = "Predicted Charges for Potential Clients",
       x = "Customer",
```

```r
      y = "Predicted Charges") +
  theme_minimal()
```



Question 8

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

> Wider prediction intervals indicate higher uncertainty in the predictions, which can arise from various sources. Clients 3 and 5, with prediction intervals of 28996.016 and 12383.797, exhibit a greater range due to more variability in the data *or* potential outliers that influence the model's predictions. Clients 3 and 5 have the largest age gap, one has children and the other one does not, and one smokes while the other does not.
>
> On the other hand, Clients 2 and 4, with narrower intervals of 8670.376 and 7855.437, were closer in age and may have more consistent and less variable data that leads to more confident predictions.
>
> The width of prediction intervals is influenced by the specific characteristics of each client's input data and the inherent uncertainties in the modeling process, resulting in varying levels of confidence in the predicted charge values.

Question 9

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

Perhaps. The most obvious, as mentioned above, are prediction intervals for Clients 3 and 5 being notably wide. The wider intervals suggests higher uncertainty in predicting insurance charges for these clients. This uncertainty could stem from unique characteristics or outliers in their data, making it challenging for us to interpret precise predictions.

Aside from that, the predicted insurance charge for Client 4 is relatively low compared to the other clients. This indicates a potential outlier, or perhaps unique characteristics that weren't captured well by the model. It would be wise to further assess Client 4 and uncover any data discrepancies that are present.

Addressing these issues could enhance overall accuracy of the predictions for all five potential clients.