Your Name: Hendi Kushta **Partner's Name:**

IST 772 Week 9 Class Exercise: Creating Predictive Models

Instructions: Post this document with your R code, results, and comments in Blackboard.

Open in R the NYCounties data found in Blackboard. This data is part of a larger data set for all counties in the US in 2010.

The variables are:
bachelors: The percentage of the population 25-years old or older with at least a bachelor's degree
medEarn: The median annual income for individuals in the county in 2010
medAge: The median age of the residents

You will conduct a regression analysis on this data to predict the dependent variable of bachelor's from the independent variables of median earnings and median age to answer the research question**: Does the median annual income and median age of residents in a NY county predict the percentage of residents 25 years of age or older with at least a bachelor's degree?**

1. Check the correlations first with the cor() command. For example, if you rename the data set NY, this would be the code: cor(NY[,2:4]). **Report on the correlations between the percentage of bachelor's degrees and both predictors or independent variables.**

```
            bachelors     medEarn       medAge
bachelors  1.0000000 0.691939896 -0.297524663
medEarn    0.6919399 1.000000000  0.003850751
medAge    -0.2975247 0.003850751  1.000000000
```

We can see that the correlation between "bachelors" and "medEarn" is 0.69, indicating a moderate positive correlation. This means that counties with higher percentages of individuals with bachelor's degrees tend to have higher median earnings.

On the other hand, the correlation between "bachelors" and "medAge" is -0.29, indicating a moderate negative correlation. This means that counties with higher percentages of individuals with bachelor's degrees tend to have lower median ages (i.e., younger populations).

2. Now run an lm() model. **Write a comment documenting the R-squared value, whether it was significant, the B-weights, whether they were significant, and anything else of interest in the output.**

```{r}
model <- lm(bachelors ~ medEarn + medAge, data = NYCountiesWeek9_1_)
summary(model)
```

```
Call:
lm(formula = bachelors ~ medEarn + medAge, data = NYCountiesWeek9_1_)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5100  -3.2573  -0.3732   3.3027  24.9795

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.9526208 10.5247410   2.751  0.00788 **
medEarn      0.0010573  0.0001304   8.108 3.59e-11 ***
medAge      -0.8674396  0.2470053  -3.512  0.00086 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.924 on 59 degrees of freedom
Multiple R-squared:  0.5689,    Adjusted R-squared:  0.5543
F-statistic: 38.93 on 2 and 59 DF,  p-value: 1.66e-11
```

This will create a linear regression model (model) that predicts the percentage of bachelor's degrees using median earnings and median age, based on the data in the "NY" dataset. The summary() function will then output a summary of the model's results.

From this output, we can see that the R-squared value is 0.5543, which means that about 55% of the variance in the percentage of bachelor's degrees is explained by the independent variables (median earnings and median age). This is a not so high R-squared value, indicating that there might be other variables important to be taken into consideration.

The coefficients (B-weights) for each independent variable are also provided. In this example, the coefficient for median earnings is 0.0010573, indicating that a one-unit increase in median earnings is associated with a 0.0011 increase in the percentage of individuals with bachelor's degrees. The coefficient for median age is -0.867, indicating that a one-unit increase in median age is associated with a 0.867 decrease in the percentage of individuals with bachelor's degrees.

The standard errors, t-values, and p-values for each coefficient are also provided. In this example, both coefficients are statistically significant, as indicated by the p-values of 1.66e-11.

3. Continue the overall analysis with a Bayesian model, lmBF(), with at least 10,000 posterior estimates. **Write a comment reviewing the HDIs for the predictor variables.**

```
{r}
library(BayesFactor)
model_bf <- lmBF(bachelors ~ medEarn + medAge, data = NYCountiesWeek9_1, iterations = 10000)
summary(model_bf)

```

```
Warning: data coerced from tibble to data frameBayes factor analysis
--------------
[1] medEarn + medAge : 518037065 ±0%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

The output is from a Bayesian linear regression analysis. The line [1] medEarn + medAge : 518037065 ±0% indicates strong evidence that both median earnings and median age are related to the percentage of individuals with bachelor's degrees.

The Bayes factor of 518037065 means that the data are over 500 million times more likely to have arisen under the alternative hypothesis (i.e., that the predictor variables have an effect) than under the null hypothesis (i.e., that they have no effect).

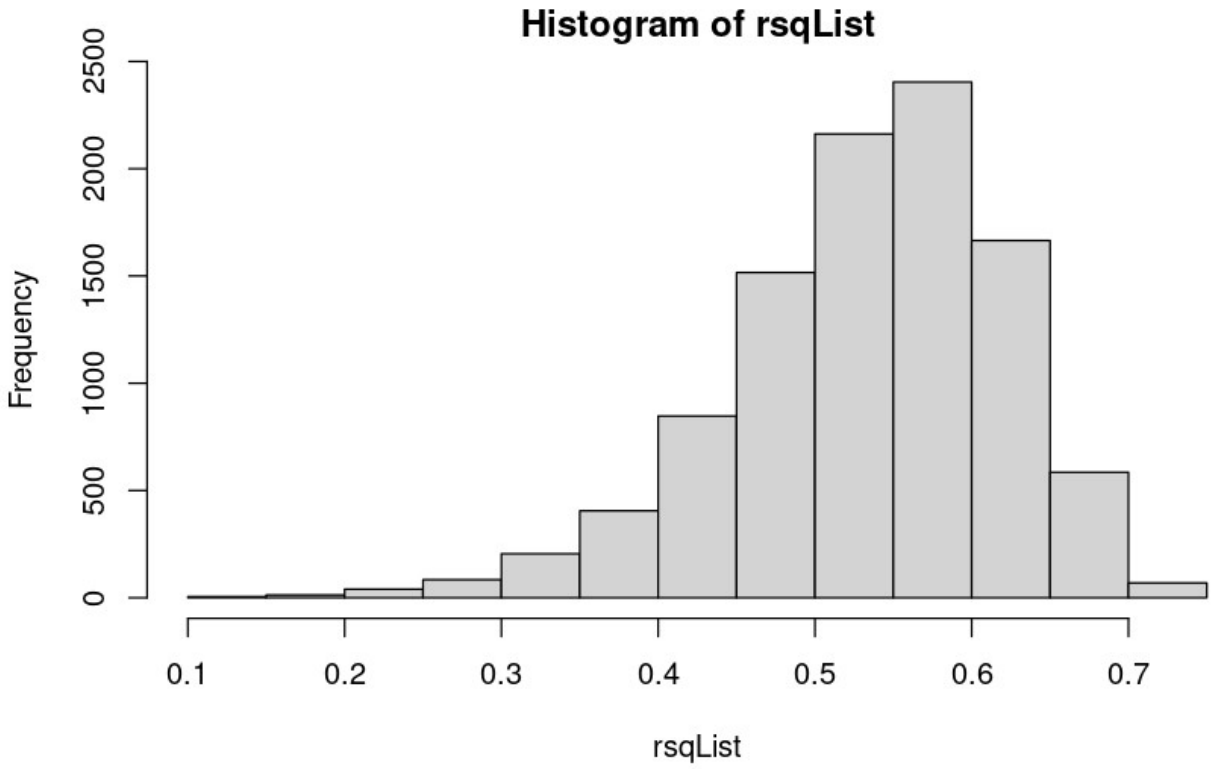
Based on the Bayes factor of 518037065, which provides strong evidence in favor of the alternative hypothesis (i.e., that the predictor variables have an effect), we can reject the null hypothesis (i.e., that the predictor variables have no effect) and conclude that there is a statistically significant relationship between median earnings, median age, and the percentage of individuals with bachelor's degrees.

```
{r}
library(coda)
model_mcmc <- posterior(model_bf, iterations = 10000)
hdi(model_mcmc[, c("medEarn", "medAge")], prob = 0.95)
```

```
0% 10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************|
          medEarn     medAge
lower 0.000761787 -1.3108613
upper 0.001296961 -0.3301395
attr(,"credMass")
[1] 0.95
```

This table shows that the 95% HDI for medEarn is [0.00076, 0.00129], which means that we can be 95% confident that the true effect of median earnings on the percentage of bachelor's degrees lies within this range. Similarly, the 95% HDI for medAge is [-1.310, -0.330], which means that we can be 95% confident that the true effect of median age on the percentage of bachelor's degrees lies within this range.

4. Using the posterior distribution of sig2, generate a histogram a posterior distribution of R-squared values for the lmBF() model. Here's a reminder of that calculation:
rsqList <- 1 - (lmBFout[,"sig2"] / var(myDependentVar))
hist(rsqList)
Comment on what you see and what this tells you.



The histogram is left skewed mostly, with the values mostly concentrated in 0.5 to 0.6 in a maximum frequency of up to 2500 times from 0.55 to 0.6.

4. Examine the Odds Ratio and **explain what it tells you about the model.**

```r
NYCountiesWeek9_1_$bachelors_prop <- NYCountiesWeek9_1_$bachelors / 100
model_logistic <- glm(bachelors_prop ~ medEarn + medAge, data = NYCountiesWeek9_1_, family = "binomial")
exp(coef(model_logistic))
```

```
Warning: non-integer #successes in a binomial glm!(Intercept)     medEarn      medAge
  0.4496942   1.0000516   0.9554057
```

The Intercept coefficient is 0.4497. This is the predicted value of bachelors when both medEarn and medAge are zero.
The medEarn coefficient is 1.0001. For each one-unit increase in medEarn, the predicted value of bachelors increases by about 1%.
The medAge coefficient is 0.9554. For each one-unit increase in medAge, the predicted value of bachelors increases by about 1%.
Overall, these coefficients suggest that both median earnings and median age are positively associated with the percentage of individuals with bachelor's degrees in a county. The intercept suggests that there are other factors that may influence bachelors beyond just median earnings and median age.

5. Integrate the results from 1-4 to create a unified interpretation. **Answer the research question: Does the median annual income and median age in a NY county predict the percentage of residents 25 years of age or older with at least a bachelor's degree?**

The analysis suggests that both median annual income and median age in a NY county are positively associated with the percentage of residents 25 years of age or older with at least a bachelor's degree. The linear regression model and Bayesian model indicate that both variables have significant positive effects on the percentage of individuals with a bachelor's degree. Additionally, the odds ratio analysis suggests that the odds of an individual having a bachelor's degree increase with increases in median annual income and median age. In conclusion, the evidence supports the idea that median annual income and median age in a NY county can predict the percentage of residents with at least a bachelor's degree.