How Smoking Habit Affects the Household Expenses (Statistical Analysis using R)

Basic Statistics

```
> summary(MultiRegDataset)
      age            sex                 bmi           children        smoker             region
 Min.   :18.00   Length:1338        Min.   :16.00   Min.   :0.000   Length:1338        Length:1338
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000   Class :character   Class :character
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character   Mode  :character
 Mean   :39.21                      Mean   :30.67   Mean   :1.095
 3rd Qu.:51.00                      3rd Qu.:34.70   3rd Qu.:2.000
 Max.   :64.00                      Max.   :53.10   Max.   :5.000
    expenses
 Min.   : 1122
 1st Qu.: 4740
 Median : 9382
 Mean   :13270
 3rd Qu.:16640
 Max.   :63770
```

Age

Min: The youngest age reported within the observations is 18
1st Quartile: At least 25% of them were within 27 years of age
Median: 50% of them were aged 39 or lesser
Mean: Average age of observations is 39.21 years
3rd Quartile: 75% of them were aged 51 or lesser
Max: The eldest age reported within the observations is 64
Standard Deviation: The high standard deviation of 14.04996 show the data scattered from the average 39.21

Sex - This feature is categorical and includes values ‘Male’ and ‘Female’

Smoker - This feature is categorical and contains values ‘yes’ & ‘no’

Region - This feature is categorical and contains values ‘southwest’, ‘southeast’, ‘northeast’ & ‘northwest’

BMI

Min: The youngest age reported within the observations is 18
1st Quartile: At least 25% of them were within 27 years of age
Median: 50% of them were aged 39 or lesser
Mean: Average BMI of these observations is 30.67
3rd Quartile: 75% of them were aged 51 or lesser
Max: The eldest age reported within the observations is 64
Standard Deviation: The high standard deviation of 6.098382 show the data dispersed from the average 39.21

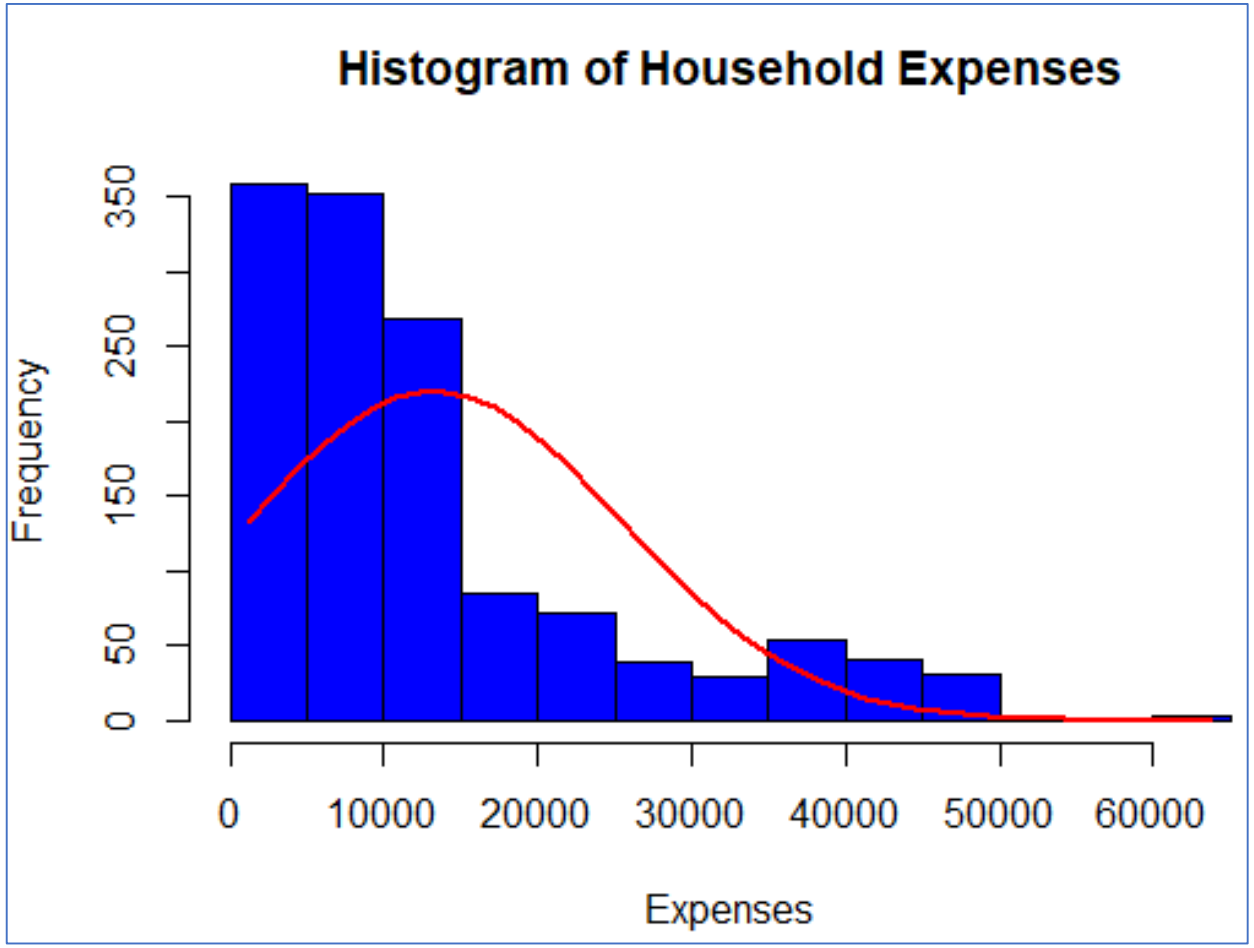
Children

Min: The minimum number of children reported within the observations is 0 (no children)
1st Quartile: At least 25% of them had no children
Median: 50% of them at least had 1 or more children
Mean: Average number of children is 1.095
3rd Quartile: 75% of them had 2 children or lesser
Max: The maximum number of children reported within the observations is 5
Standard Deviation: The standard deviation of 1.205493 show the data dispersed from the average 1.095

Expenses

Min: The minimum household expenses reported within the observations is 1122
1st Quartile: 25% of them had expenses 4740 or lesser
Median: 50% of them have had household expenses 9382 or higher
Mean: Average household expense within the observations is 13270
3rd Quartile: 75% of them had household expenses 16640 or lesser
Max: The highest household expenses reported within the observations is 63770
Standard Deviation: The standard deviation of 12110.01 show the data dispersed from the average 13270

Histogram for Household Expenses



This is a right-skewed (positive-skew) distribution which has a long right tail

T-Test

Null and Alternative Hypothesis for T-Test

Null Hypothesis → Ho: The mean of expenses is 10000

Ho: $\mu = 10000$

Alternative Hypothesis → Ha: The mean of expenses is not 10000

Ha: $\mu \neq 10000$

- Significance level will be considered as 0.05
- 1 sample Two-Tailed Test is selected because 'Ha' is not in the form of "greater than" or "less than"

Conducting T-Test

```
> #Create the mean, standard deviation, and standard error
> mean.x <- mean(x)
> sd.x <- sd(x)
> SE.x <- sd(x) / sqrt(length(x))
>
> #Show the mean, standard deviation and standard error
> mean.x
[1] 13270.42
> sd.x
[1] 12110.01
> SE.x
[1] 331.0675
>
> #State the Ho value and calculate the z-score
> Ho <- 10000
> z <- (mean.x - Ho) / SE.x
> z
[1] 9.878417
>
> #Two-tail Test
> 2*pnorm(abs(z),lower.tail=FALSE)
[1] 5.164217e-23
```

We have received a p-value of 5.164217e-23

It is much lower than significance level 0.05, therefore we reject the null hypothesis that the mean of the expenses is equal to 10000 and fail to reject the alternative hypothesis which states the mean of the expenses is not equal to 10000

Conclusion

We reject the null hypothesis that the mean of expenses is equal to 10000 with a 5% level of confidence

Simple Linear Regression Model

Summary

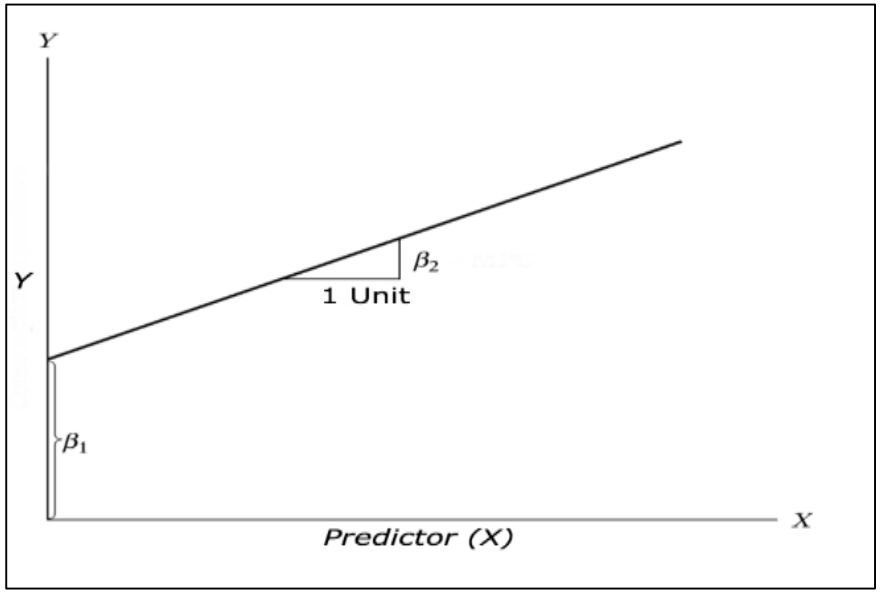
- Linear regression method is used to predict the value of a dependent variable (Y) based on one or more input independent Variables (X).
- The purpose is to establish a linear relationship between the predictor variable(independent) and the response variable (dependent), so that, we could use this formula to estimate the value of the response Y, when only the predictors (X) values are known
- The formula can be generalized as follows: $Y = \beta1 + \beta2X + \varepsilon$

  β1 -> Intercept

  β2 -> Slope. Collectively , they are called regression coefficients

  ε -> Error Term

  Y -> The part of Y the regression model is unable to explain



- Depending on all these above, we will be conducting a linear Analysis for the provided data set. The independent variable will be ‘smoker’ and the dependent variable will be ‘expenses’
- Our model will give the relationship between these two variables and whether the model is statistically significant or not

Null and Alternative Hypothesis

Null Hypothesis → Ho: β=0, co-efficient β of the predictor is zero and not statistically significant (A relationship between 'smoker'[independent variable] and 'expenses' [dependent variable] does not exist).

Alternative Hypothesis → Ha: β ≠0, co-efficient β of the predictor is not equal to zero and is statistically significant (A relationship between 'smoker' [independent variable] and 'expenses' [dependent variable] does exist).

Performing the Test

Step 1: Load Required Libraries for the Analysis and Review Data

```
#Load Libraries
library(dplyr)
library(ggplot2)
```

Step 2: View and Review Dataset

```
#View & Review Dataset
View(MultiRegDataset)
str(MultiRegDataset)
```

Output:

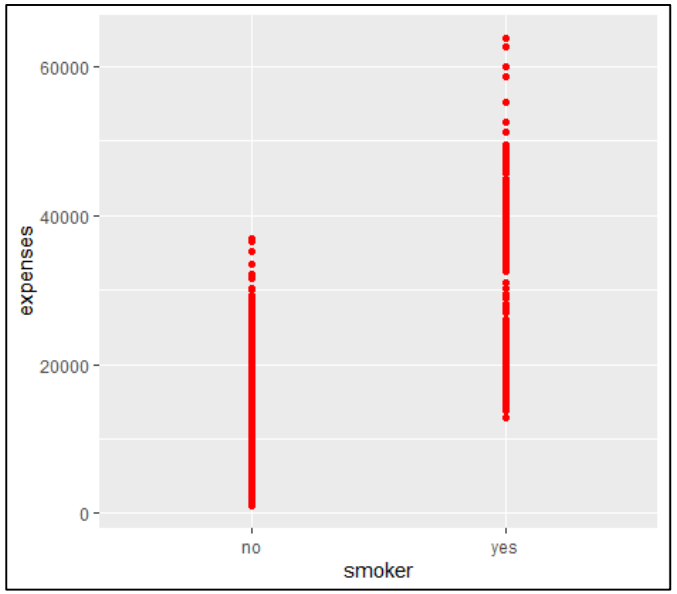
```
> #View & Review Dataset
> View(MultiRegDataset)
> str(MultiRegDataset)
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : chr  "female" "male" "male" "male" ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : chr  "yes" "no" "no" "no" ...
 $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ expenses: num  16885 1726 4449 21984 3867 ...
> g
```

When we look at the data types of the variables, smoker is a categorical data which contains ‘yes’ and ‘no’. ‘expenses’ column contains numerical data. Having the independent variable as categorical data can be problematic and this will be considered during the analysis

Step 3: Examining the correlation

```
#Linear Regression Model
#smoker vs. expenses
ggplot(MultiRegDataset,aes(x = smoker, y = expenses)) + geom_point(colour = "red")  +
  geom_smooth(method = "lm", fill = NA)
```

Output:



Step 4: Building the model

```
#Build Linear Model
simple.fit<-lm(expenses~smoker, data=MultiRegDataset)
LinearModel<-simple.fit

#Summary of Key Statistics of the Model
summary(LinearModel)
```

Output:

```
Call:
lm(formula = expenses ~ smoker, data = MultiRegDataset)

Residuals:
   Min     1Q Median     3Q    Max
-19221  -5042   -919   3705  31720

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8434.3      229.0   36.83   <2e-16 ***
smokeryes    23616.0      506.1   46.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7470 on 1336 degrees of freedom
Multiple R-squared:  0.6198,    Adjusted R-squared:  0.6195
F-statistic:  2178 on 1 and 1336 DF,  p-value: < 2.2e-16
```

Interpretation

Equation for Regression Model

$$Y = \beta 1 + \beta 2X + \varepsilon$$

$$Y(\text{expenses}) = 8434.3 + (23616) * X(\text{smoker}) + \varepsilon$$

- ✓ If the person is a smoker (X=1) our expenses is translate into 32050.3, If the person does not smoke (X=0) our expenses is translate into 8434.3. Since our independent variable is categorical, we will have only two outputs predicted for Y
- ✓ Residual Standard Error - 7470 on 1336 degrees of freedom: This is the standard deviation of the residuals. Smaller is better
- ✓ Multiple / Adjusted R-Square: For one variable, the distinction doesn't really matter. R-squared shows the amount of variance explained by the model. Adjusted R-Square value (0.6195) takes into account the number of variables and is most useful for multiple-regression. Multiple R-squared value 0.6198 (61.98%) indicates that the model fails to explain the variability of the response data around its mean
- ✓ F-Statistic (2178 on 1 and 1336 DF): The F-test checks if at least one variable's weight is significantly different than zero. This is a global test to help asses a model. If the p-value is not significant (greater than 0.05) than your model is essentially not doing anything
- ✓ p-value is 2.2e-16 which is less than the significance value 0.05
- ✓ Max value of 31720 indicates that there is a point which lies far above the predicted regression line and min value of - 19221 indicates that there is a point lies far below the predicted regression line

Evaluation

- ✓ We reject the null hypothesis with a 5% level of confidence which states that the relationship between ‘smoker’(independent variable) and ‘expenses’ (dependent variable) does not exist and we fail reject the alternative hypothesis
- ✓ In conclusion, there is correlation between ‘smoker’(independent variable) and ‘expenses’ (dependent variable) . Our model is statistically significant

Multiple Linear Regression Model

- It is also known simply as multiple regression
- It is a statistical technique that uses several independent variables to predict the outcome of a dependent variable
- The formula can be generalized as follows:

$$y_i = B_0 + B_1x_{i1} + B_2x_{i2} + \ldots + B_px_{ip} + E \text{ where } i = 1,2, \ldots, n$$

β0 -> Intercept

β1 -> Slope. Collectively , they are called regression coefficients

ε -> Error Term

Y -> value of dependent variable

Null and Alternative Hypothesis for the Test

Null Hypothesis → Ho: age = sex = bmi = children = smoke = region = 0 **Alternative Hypothesis** → Ha: at least one βi ≠ 0 (for i = 1, 2, 3)

Significance level will be considered as 0.05

Building the Model

```
#Building Multiple Linear Regression Model

#Create Model1
model1 <- lm(expenses~., data = MultiRegDataset)

#Summary of Model1
print(model1)
summary(model1)
```

Output

```
Call:
lm(formula = expenses ~ ., data = MultiRegDataset)

Coefficients:
    (Intercept)              age          sexmale              bmi         children        smokeryes
      -11941.6            256.8           -131.4            339.3            475.7          23847.5
regionnorthwest  regionsoutheast  regionsouthwest
         -352.8          -1035.6           -959.3

> summary(model1)

Call:
lm(formula = expenses ~ ., data = MultiRegDataset)

Residuals:
     Min       1Q   Median       3Q      Max
-11302.7  -2850.9   -979.6   1383.9  29981.7

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11941.6      987.8 -12.089  < 2e-16 ***
age                256.8       11.9  21.586  < 2e-16 ***
sexmale           -131.3      332.9  -0.395 0.693255
bmi                339.3       28.6  11.864  < 2e-16 ***
children           475.7      137.8   3.452 0.000574 ***
smokeryes        23847.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -352.8      476.3  -0.741 0.458976
regionsoutheast  -1035.6      478.7  -2.163 0.030685 *
regionsouthwest   -959.3      477.9  -2.007 0.044921 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Interpretation

Max value of 29981.7 indicates that there is a point which lies far above the predicted regression line and min value of -11302.7indicates that there is a point lies far below the predicted regression line

Equation for Regression Model

$$y_i = B_0 + B_1x_{i1} + B_2x_{i2} + \ldots + B_px_{ip} + E \text{ where } i = 1,2, \ldots, n$$

Y(expenses) = (-11941.6) + 256.8 *X(age) + (-131.3)*(sexmale) + 339.3*(bmi) + 475.7*(children) + 23847.5*(smokeryes) + (-352.8)*(regionnorthwest) + (-1035.6)*(regionsoutheast) + (-959.3)*regionsouthwest+ ε

- ✓ Residual Standard Error - 6062 on 1329 degrees of freedom: This is the standard deviation of the residuals. Smaller is better
- ✓ Multiple / Adjusted R-Square: For one variable, the distinction doesn't really matter. R-squared shows the amount of variance explained by the model. Adjusted R-Square value (0.7494) takes into account the number of variables and is most useful for multiple-regression. Multiple R-squared value 0.7509 (75.09%) indicates that the model fails to explain the variability of the response data around its mean
- ✓ F-Statistic (500.9 on 8 and 1329 DF): The F-test checks if at least one variable's weight is significantly different than zero. This is a global test to help asses a model. If the p-value is not significant (greater than 0.05) than your model is essentially not doing anything
- ✓ p-value is 2.2e-16 which is less than the significance value 0.05

Evaluation

- Since p-value is lesser than the significance level, we reject the null hypothesis with a 5% level of confidence and we fail to reject the alternative hypothesis
- Features **sexmale** and **regionnorthwest** are not statistically significant because its p-values (0.693255 & 0.458976) are greater than the significance level of 0.05
- The remaining features look statistically significant
- It is standard practice to use the coefficient p-values to decide whether to include variables in the final model
- Having these features in the model can affect model's precision
- For the results above, we could consider removing sexmale and regionnorthwest

Conclusion

- When we compare the adjusted R2 values of both LR and MLR
- MLR has given better results
  - LR – 0.6195
  - MLR – 0.7494

- Therefore, based on these evidences, I would suggest to select MLR for Mr.Hughes's dataset