SRILASYA GARIGIPATY

DATA 1204 FINAL PROJECT

STUDENT #100822953

Description of Research Requirements

Background :

Mr. John Hughes has been collecting data on the effect of personal attributes on household expenses. He has put together a dataset (**MultiRegDataset.csv**) with contains **1338 observations (rows) and 7 features (columns).** The details of the features are as follows:

Independent (Input) variables:

- Age
- Sex
- BMI
- Children
- Smoker
- Region

Dependent (Output) variable:

- Expenses

The Ask from Mr. John Hughes who would like to understand:

- The effect of **smoking** on expenses by creating a linear regression model
- The effect of **all input variables** on expenses by creating a multivariate regression model

Basic Statistics of the Data

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
age	1	1338	39.21	14.05	39.00	39.01	17.79	18.00	64.00	46.00	0.06	-1.25	0.38
sex*	2	1338	1.51	0.50	2.00	1.51	0.00	1.00	2.00	1.00	-0.02	-2.00	0.01
bmi	3	1338	30.67	6.10	30.40	30.50	6.23	16.00	53.10	37.10	0.28	-0.06	0.17
children	4	1338	1.09	1.21	1.00	0.94	1.48	0.00	5.00	5.00	0.94	0.19	0.03
smoker*	5	1338	1.80	0.40	2.00	1.87	0.00	1.00	2.00	1.00	-1.46	0.14	0.01
region*	6	1338	2.52	1.10	3.00	2.52	1.48	1.00	4.00	3.00	-0.04	-1.33	0.03
expenses	7	1338	13270.42	12110.01	9382.03	11076.02	7440.81	1121.87	63770.43	62648.56	1.51	1.59	331.07

The above chart displays the basic statistics of the MultiRegDataset. It is noted with a star that the variables “sex”,”smoker”,and “region” are all categorical variables.

Basic Statistics of the Data

- Correlation ranges from -1 to 1.A negative correlation indicates that as one variable goes up, the other goes down. A correlation of 0 means that two variables are not related at all. A correlation of 1 is perfect correlation, and means that as the first variable changes, the second changes in the same direction, though not necessarily by the same amount.

- From the correlation function on R code, it was noted that the independent numeric variable “age”,”bmi”,”children” had correlations with the dependent variable “expense” at values of 0.29,0.19, and 0.067, respectively . The highest numeric independent variable that is correlated with the dependent variable “expense” is “age” at 0.29. Change in this variables will have a small effect on the “expense” of the household.

- The mean is defined as the mathematical average of the data. The mean, maximum values, and minimum values insight will tell how much the variability is affecting the independent variables and what effect this might have on the dependent variable.The mean of “age” is 39 years and the mean of “expense” is 13270.42 dollars.

- The maximum value of “age” is 64 years, and the minimum value is 18 years. The maximum value of “expense” is 63770.43 dollars, and the minimum is 1121.87 dollars. This indicates that there is a high variability of “expense” and “age” values within the dataset. The mean, maximum values, and minimum values insight will tell how much the variability is affecting the expense of the household, and what kinds of differences are there between the expense of household and the wide range of “age” values. The higher the “age” value, the more likely household “expense” will rise.

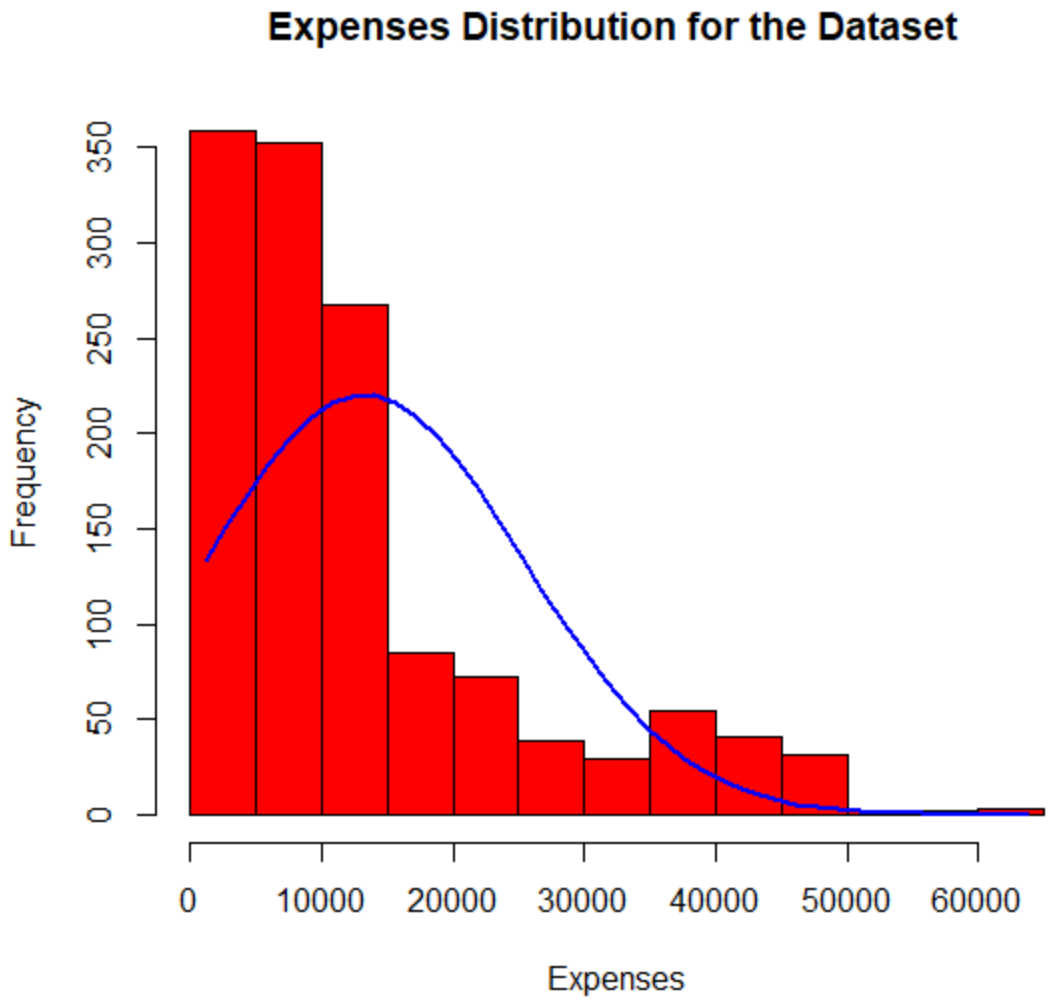
- The variability in “age” may have little effect on the c”expense” as they have only 0.2 correlation value. The higher the “age” value, the more likely expense will be higher, but it is not guaranteed.

- Standard Deviation describes the spread of the observation from the mean. A high standard deviation indicates that the values are spread over a wide range. A low standard deviation indicates that the values are closer to the mean(average) value.

- The standard deviation for “age” variable is 14 years which is far from the mean value of 39 years indicating a large variation in the variable and further exploring shows this variable has a lot of high and low values in different ranges.

- The standard deviation for “expense” variable is 12110 dollars which is not very far from the mean value of 13270 dollars indicating a small variation in the variable and further exploring shows this variable has values within close distance of each other.

Histogram of Dependent Variable “Expense”



- It can be seen from the histogram that it is right skewed.
- The majority of household expense are between the 0 to 15000 dollar range.
- The frequency of higher expenses decreases as the graph moves towards right.
- The expenses stretch out past the 60000 dollar expense.
- This histogram is positively skewed, with skewness coefficient of 1.5, which indicates most of the values in the “expenses” category are elss than mean value.

T-test Hypotheses

- H_0: $\mu s = 10000$(null hypothesis)
- Ha: $\mu s \neq 10000$ (alternative hypothesis)

- The **null hypothesis** indicates that the mean for household expense is equal to 10000 dollars.
- The **alternative hypothesis** indicates that the mean for household expense is not equal to 10000 dollars.

T-test Results

```
        One Sample t-test

data:  MultiRegDataset$expenses
t = 9.8784, df = 1337, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 10000
95 percent confidence interval:
 12620.95 13919.89
sample estimates:
mean of x
 13270.42
```

After conducting the One-Sample T-test, the following insights were determined:

- A **t-value of 9.8784** was found. The greater the magnitude of the t-value, the greater there is evidence to reject the null hypothesis that the mean average household expense was 10000. The larger the absolute value of the t-value, the smaller the p-value, and the greater the evidence against the null hypothesis.
- A **p -value of 2.2*10^-16** was found. The p-value or **probability value** is the probability that of the null hypothesis that the mean average household income being 10000 dollars is true.
- The critical value (confidence interval) used for analysis is 0.05. This means that 5% of the time the null hypothesis that the household expense is 10000 dollars can be rejected, and 95% of the time it cannot be rejected.

T-test Conclusion

```
        One Sample t-test

data:  MultiRegDataset$expenses
t = 9.8784, df = 1337, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 10000
95 percent confidence interval:
 12620.95 13919.89
sample estimates:
mean of x
 13270.42
```

- **p_value > α (Critical value)**: Fail to reject the null hypothesis of the statistical test.
- **p_value ≤ α (Critical value)**: Reject the null hypothesis of the statistical test.

- In this case, the **p-value is 2.2*10^-16** which is less than critical value of 0.05. The null hypothesis Ho that mean household expense is 10000 should be rejected, and accept the alternative hypothesis Ha that mean average household expense is not 10000. This is proven by the T-test performed, which indicated the mean of the expense is 13270 dollars and not equal to null hypothesis that mean of expense is10000 dollars.

Simple Linear Regression Model

Step 1: Load the Data

This model will attempt to fit a simple linear regression model using *smoker* as the explanatory variable and expenses as the response variable.

Step 2: Visualize the Data

Before a simple linear regression model is fit on the data, the data should be visualized to get a clear understanding of it. The purpose of visualizing the data is to check that the relationship between the two variables smoker and expense is roughly linear or not.

Step 3: Perform Simple Linear Regression

Proceed to fit linear regression model using smoker as explanatory variable and expense as response variable.

Step 4:Interpret and Analyze the results.

Based on model summary, get insights about the variables and their relationship.

Interpretation and Evaluation of Simple Linear Regression Model

Results:

```
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)  32050.2      451.3   71.02   <2e-16 ***
smoker2     -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
```

- **Pr(>|t|):** This is the p-value associated with the model coefficients. Since the p-value for smoker (**2*10^-16**) is significantly less than .001 significance level, we can say that there is a statistically significant association between *smoker* and *expenses.*
- **Multiple R-squared:** This number tells us the percentage of the variation in expenses that can be explained by number of smokers studied. In general, the larger the R-squared value of a regression model the better the explanatory variables are able to predict the value of the response variable. In this case, **62%** of the variation in expense can be explained by whether a person smokes or not.
- **Residual standard error:** This is the average distance that the observed values fall from the regression line. The lower this value, the more closely a regression line is able to match the observed data. In this case, the average observed expense falls **7470** points away from the value predicted by the regression line.
- **F-statistic & p-value:** The F-statistic (**2178**) and the corresponding p-value (**2.2*10^-16**) tell us the overall significance of the regression model, i.e. whether explanatory variables in the model are useful for explaining the variation in the response variable. Since the p-value in this example is less than .001 significance level, our model is statistically significant, and *smoker* is deemed to be useful for explaining the variation in *expenses*.

Multi-Linear Regression Model

Step 1: Load the Data

This model will attempt to fit multi -linear regression model using all independent variables as the explanatory variable and expense as the response variable.

Step 2: Visualize the Data

Before a multi-linear regression model is fit on the data, the data should be visualized to get a clear understanding of it. The purpose of visualizing the data is to check that the relationship between the independent variables and expense is roughly linear or not.

Step 3: Create Relationship Model & get the Coefficients

Proceed to fit multi-linear regression model using all independent variables as explanatory variable and expense as response variable.

Step 4: Apply Equation for predicting New Values & Get insights

Based on model summary, get insights about the variables and their relationship and create an equation based on coefficient values of independent variables to predict dependent variable.

Interpretation and Evaluation of Multi-Linear Regression Model

Results:

```
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)      11905.9     1034.0  11.515  < 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 ***
smoker2          -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
```

- **Adjusted R squared:** This value reflects how fit the model is. Higher the value better the fit. Adjusted R-squared value of our data set is 0.75.
  - The regression output shows that predictor variables age, bmi, children, smoker are statistically significant because their p-values are less than indicated significance level of 0.001 respectively.
  - On the other hand, sex variable is not statistically significant because its p-value (0. 69) is greater than the usual significance level of 0.05. region variable p-value is greater than usual significance level of 0.05 indicating it is not statistically significant.
  - It is standard practice to use the coefficient p-values to decide whether to include variables in the final model. For the results above, we would consider removing sex, and region variables. Keeping variables that are not statistically significant can reduce the model’s precision.

Prediction Equation:

Y (expense)= 11905.9+(256.8)*age+(339)*bmi+(475.7)*children+(-23847)*smoker

Conclusion Based on Findings

- To Answer Mr. John's Question on whether there is an effect of smoking on expenses, and what are the effect of all input variables on expenses, the following was concluded:

- It can be noted from the linear regression analysis that there is a significant association between whether a person smokes, or does not smoke and their household expenses.

- Almost 62% of variation in expenses can be explained by whether a person has a smoking habit.

- Looking further into the different aspects of smoking may provide Mr. John more insight into patterns and trends on household expenses.

- It can be noted from the multi-linear analysis that after smoking which has the highest coefficient and significant effect on household expenses in predictor equation, age is the second factor that has the largest effect on household expenses, followed by bmi, and number of children.

- The sex of the individual had little effect the total household expense, as it was found that sex is not a statistically significant variable in multi-linear regression analysis.

- The region also has little effect on the total household expense, as it was found that region is not a statistically significant variable in multi-linear regression.

Conclusion Based on Findings

- From the t-test analysis, it was found that the average expenses were higher than the null hypothesis prediction of 10000 dollars.
-
- There could be many factors contributing to this higher mean expense, including whether the person smokes, their age, bmi, and number of children.

- The data variables from basic statistics can be interpreted to be skewed or not skewed based on the skewness value:

•**Symmetric**: Values between -0.5 to 0.5
•**Moderated Skewed data**: Values between **-1 and -0.5** or between **0.5 and 1**
•**Highly Skewed data**: Values **less than -1** or **greater than 1**

- From basic statistics table, and skewness value of each variable it was noted that :

Age-highly skewed
Sex-symmetric
bmi-symmetric
Children-symmetric
Smoker-highly skewed
Region-symmetric
Expenses- highly skewed

- Mr. John should examine his dataset to get a less skewed data, or determine which factors are causing the increased skewness in some variables and not others to get more insight.

References

1. **DATA 1204 CLASSNOTES**
2. **https://www.r-bloggers.com/2020/11/skewness-and-kurtosis-in-statistics/**
3. **https://vitalflux.com/data-science-8-steps-to-multiple-regression-analysis/**
4. **https://www.tutorialspoint.com/r/r_multiple_regression.htm**
5. **http://analyticuniversity.com/step-by-step-linear-regression-in-r/**