Program

BITM-Data Analysis for Business Decision Making Graduate Certificate

COURSE

1204_STATISTICAL PRED MODELING

PROF. RITWICK DUTTA

ASSIGNMENT -06

MULTI-LINEAR REGRESSION

SUBMITTED BY

AYYAPPA MADU

100827123

Problem Statement

- To perform straight relapse, demonstrate for the impact of smoking on costs and multivariate relapse demonstrate for impact of all input factors on costs on MultiRegDataset to Mr. Hughes.

- **Key Statistics**

From the Key statistics, we will induce a few of the following:

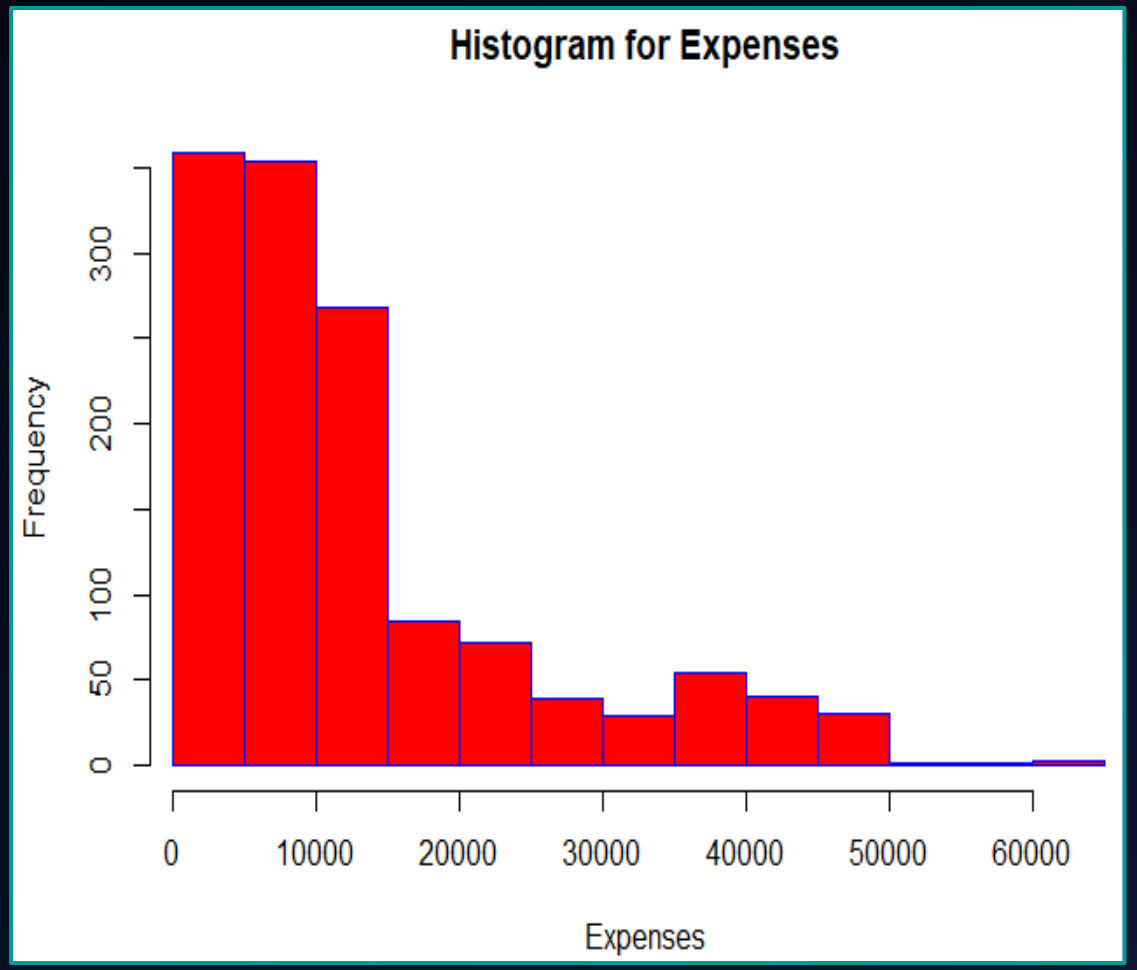
- In our information set, we have a total of 7 columns with 1338 rows.
- The ***most extreme expenses*** of the individual is **$ 63770.43**
- Individuals with an average age of 39, a min of 18, and a max of 64 years old ranging from 64 years ***expenses*** **$13270.42** on average and the minimum value of **$1121.87.**
- Median BMI is 30.40 when compared to its mean 30.67, shows there is not much deviation.

	vars	n	mean	sd	median	trimmed	mad	min	max	range
age	1	1338	39.21	14.05	39.00	39.01	17.79	18.00	64.00	46.00
sex*	2	1338	1.51	0.50	2.00	1.51	0.00	1.00	2.00	1.00
bmi	3	1338	30.67	6.10	30.40	30.50	6.23	16.00	53.10	37.10
children	4	1338	1.09	1.21	1.00	0.94	1.48	0.00	5.00	5.00
smoker*	5	1338	1.20	0.40	1.00	1.13	0.00	1.00	2.00	1.00
region*	6	1338	2.52	1.10	3.00	2.52	1.48	1.00	4.00	3.00
expenses	7	1338	13270.42	12110.01	9382.03	11076.02	7440.8	1121.87	63770.43	62648.56

	skew	kurtosis	se
age	0.06	-1.25	0.38
sex*	-0.02	-2.00	0.01
bmi	0.28	-0.06	0.17
children	0.94	0.19	0.03
smoker*	1.46	0.14	0.01
region*	-0.04	-1.33	0.03
expenses	1.51	1.59	331.07

Histogram - Expenses

- Information related to the costs appear exceedingly right-skewed
- People investing capacity is the most elevated between the run of 0-15000.
- There could be a sharp decrease after 15000 within the level of costs. Expenses producing to 60000 is the least.
- The most noteworthy check of the costs is around 180 and stands at 1000



T-test

- We can define the corresponding ***null hypothesis* (H0) and *alternative hypotheses* (Ha)** as follow:

Null Hypothesis:

H_0: $\mu s = 10,000$ *(Mean for the expenses is equal to 10,000)*

Alternative Hypothesis:

Ha: $\mu s \neq 10,000$ (Mean for the expenses is not equal to 10,000)

Significance level of $\alpha = 0.05$

- *Formula of one-sample t-test*
- The t-statistic can be calculated as follow:
- $t = \frac{m-\mu}{s/\sqrt{n}}$
- here,
  - m is the *sample mean.*
  - n is the *sample size.*
  - s is the *sample standard deviation* with n−1 *degrees of freedom.*
  - μ is the *theoretical value.*

- ***Mean expenses = 13270.42.***
- ***p-value < .05***
- The p-value of the test is 2.2e-16,
- which is less than the centrality level alpha = 0.05.
- We ***reject the null hypothesis*** which states the mean of the costs is rise to 10,000 whereas supporting the alternative hypothesis.

```
        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
```

Linear Regression

Definition

- In statistics, ***linear regression*** is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).
- Model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x) and their coefficients(Bo & B1) along with the constant (c).
- **Formula**
- ***y = B0 + B1*x + c***
- *y= Dependent variable ;* *c= intercept ;* *x = Independent variable ;*
- ***B0*** & ***B1*** *= coefficients*

Hypothesis

Null Hypothesis:

H_0: A relationship between smoker and expenses doesn't exist

Alternative Hypothesis:

Ha: A relationship between smoker and expenses does exist.

Significance level: We assume, the standard significant level of 0.05

Evaluation - linear regression model :

- p value of ***2.2e-16,which is lower than the stated significance level of 0.05***.
- ***We reject the null hypothesis,*** no relationship between the smoker and the expenses.
- We ***do not reject the Alternative Hypothesis*** i.e.; smokers will have an impact on the expenses.
- The residuals and the coefficients of the Linear regression ***model is a good fit.***

```
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
```

Multi Linear Regression

Definition

- Multiple linear regression is the extension of the simple linear regression, which is used to predict the outcome variable (y) based on multiple distinct predictor variables (x). With the help of three predictor variables (x1, x2, x3), the prediction of y is expressed using the following equation:

$$y=b0+b1*x1+b2*x2+b3*x3$$

- Here,
- y is a response variable.
- b0, b1, b2...bn are the coefficients.
- x1, x2, ...xn are the predictor variables.

Hypothesis

Null Hypothesis:

H_0: *age = sex = bmi = children = smoker = region = 0*

Alternative Hypothesis:

Ha: at least one βi ≠ 0 (for i = 1, 2, 3,4,5,6)

p- values of variables like age, bmi, smoker, children are highly significant
Reject H0 if F calculated > F criticalF Critical = F (df1,df2,α) = F(6,1333,0.05)= 2,0986

Decision (compare Fcalculated with Fcritical) F calculated = F Critical = 2.0986Since Fcalculated > Fcritical, ***we reject H0***.

Conclusion:

The model is a good fit, as at least one coefficients is not equal to zero.

```
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 < 0.0000000000000002 ***
age                256.8       11.9  21.586 < 0.0000000000000002 ***
sexmale           -131.3      332.9  -0.395             0.693255
bmi                339.3       28.6  11.864 < 0.0000000000000002 ***
children           475.7      137.8   3.452             0.000574 ***
smokeryes        23847.5      413.1  57.723 < 0.0000000000000002 ***
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: < 0.00000000000000022
```

Backward Elimination

- P – values for age, bmi, children, smoker is highly significant.
- R- square = 0.7509 i.e, is equal to full model.

```
Call:
lm(formula = expenses ~ age + bmi + children + smoker + region,
    data = MultiRegDataset)

Residuals:
     Min       1Q   Median       3Q      Max
-11365.0  -2839.4   -985.3   1375.5  29924.5

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)    -11993.31     978.75 -12.254 < 0.0000000000000002 ***
age               256.96      11.89  21.609 < 0.0000000000000002 ***
bmi               338.76      28.56  11.862 < 0.0000000000000002 ***
children          474.75     137.74   3.447              0.000585 ***
smokeryes       23835.24     411.84  57.875 < 0.0000000000000002 ***
regionnorthwest  -352.01     476.11  -0.739              0.459825
regionsoutheast -1034.93     478.53  -2.163              0.030738 *
regionsouthwest  -958.63     477.76  -2.007              0.045003 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6060 on 1330 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7496
F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 0.00000000000000022
```

Forward Selection

- p values for age, bmi, sex, children, smoker highly significant.
- Also, R – square value is same as full model

```
Call:
lm(formula = expenses ~ age + sex + bmi + children + smoker +
    region, 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 < 0.0000000000000002 ***
age                256.8       11.9  21.586 < 0.0000000000000002 ***
sexmale           -131.3      332.9  -0.395             0.693255
bmi                339.3       28.6  11.864 < 0.0000000000000002 ***
children           475.7      137.8   3.452             0.000574 ***
smokeryes        23847.5      413.1  57.723 < 0.0000000000000002 ***
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: < 0.00000000000000022
```

Remove Region from model

- R- square becomes 0.7497 which is less than full model.
- When we remove other variable like sex, R- square reduced to 0.7237 which is not good.
- Therefore, we can conclude that full model is best suitable for this dataset.

```
Call:
lm(formula = expenses ~ age + children + smoker + sex + bmi,
    data = MultiRegDataset)

Residuals:
     Min       1Q   Median       3Q      Max
-11834.4  -2916.4   -990.6   1372.1  29554.5

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -12055.16     951.23 -12.673 < 0.0000000000000002 ***
age             257.72      11.90  21.650 < 0.0000000000000002 ***
children        474.60     137.85   3.443             0.000593 ***
smokeryes     23822.31     412.51  57.750 < 0.0000000000000002 ***
sexmale        -128.68     333.35  -0.386             0.699540
bmi             322.45      27.42  11.761 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6070 on 1332 degrees of freedom
Multiple R-squared:  0.7497,    Adjusted R-squared:  0.7488
F-statistic: 798.1 on 5 and 1332 DF,  p-value: < 0.00000000000000022
```

Evaluation

- From summary of full model, forward and backward model,
- we found that p values of all variables are very significant
- R – square is same for all three strategies.
- we can conclude that all the parameters have an impact on the ***expenses*** except the ***sexmale*** variable and rest of the model seems a good fit.
- Therefore, ***we reject null hypothesis and accept that all variable affect expenses.***

Conclusion

- I would like to recommend both linear and multi linear regression models to Mr.Huges.
- Models seems to a good fit by comparing the p values with the significance levels,T-test,F-statistic results.
- All the parameters have correlation with the dependent variable ***expenses.***

THANK YOU