Multivariate Regression

ISMAIL SEVER

Table of Content

Description of Research

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). Age, sex, bmi, children, smoker, region are predictors (or independent) variables, and expenses is an outcome variable (or dependent).

We will import this dataset into our RStudio and look for answers for following studies.

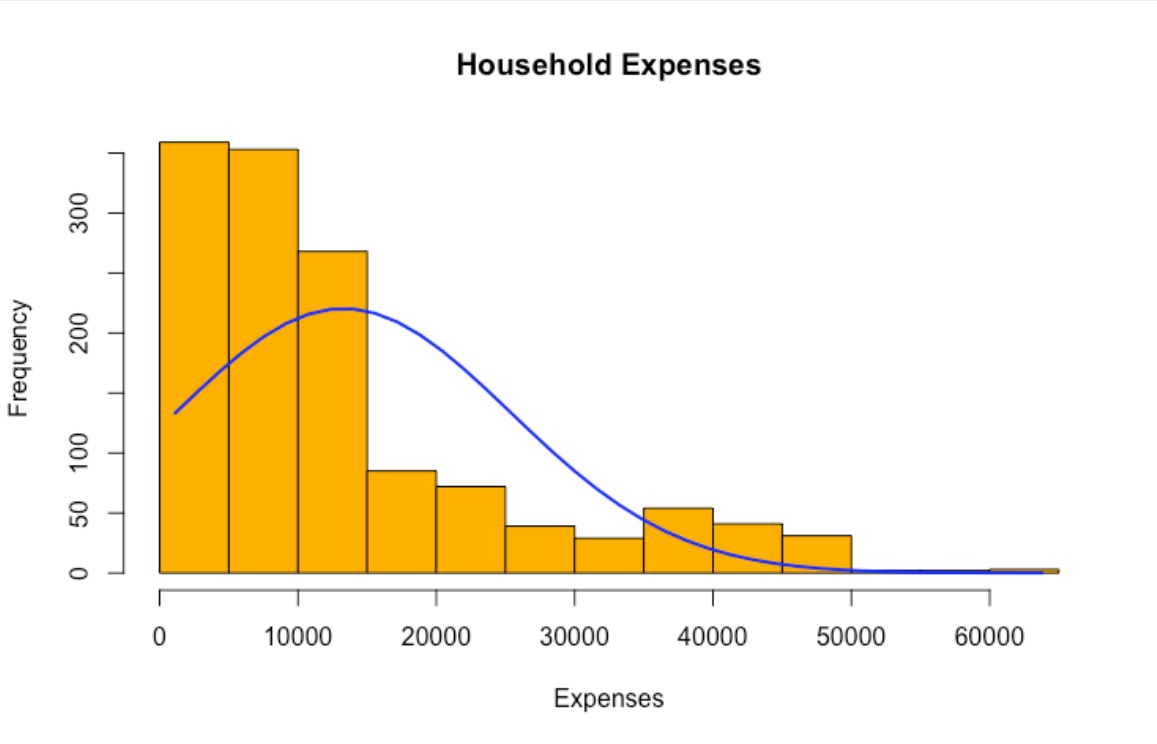
- Mean, median, min, max, standard deviation, ... of the dataset
- Histogram of the dependent variable (expenses)
- T-test whether the mean for expenses is equal to 10,000
- Build a simple linear model and check whether smoking has effect on expenses
- Build a multiple linear model and check whether all predictors has effect on expenses

Basic Statistics

- The youngest person in the dataset is 18 years old and biggest is 64 years old. Average age of people in the dataset is 39.21 and median is 39. Since mean and median are close, we can say the data in age column distributed symmetrically.
- Mean and median of BMI column is 30.67 and 30.4 respectively. We can say that more than 50% of the people are obese. (Obesity = BMI of 30 or greater)
- Average of household expenses is $13270.42 with least $1122 and highest ~$63770. Median is $9382 which is not very close to the mean, which shows that the data is not normally distributed.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
age	1	1338	39.21	14.05	39.0	39.01	17.79	18	64.0	46.0	0.06	-1.25	0.38
sex*	2	1338	1.51	0.50	2.0	1.51	0.00	1	2.0	1.0	-0.02	-2.00	0.01
bmi	3	1338	30.67	6.10	30.4	30.50	6.23	16	53.1	37.1	0.28	-0.06	0.17
children	4	1338	1.09	1.21	1.0	0.94	1.48	0	5.0	5.0	0.94	0.19	0.03
smoker*	5	1338	1.20	0.40	1.0	1.13	0.00	1	2.0	1.0	1.46	0.14	0.01
region*	6	1338	2.52	1.10	3.0	2.52	1.48	1	4.0	3.0	-0.04	-1.33	0.03
expenses	7	1338	13270.42	12110.01	9382.0	11076.02	7440.81	1122	63770.4	62648.6	1.51	1.59	331.07

Histogram



- The curve shows that the data is not normally distributed. We can also use Shapiro-Wilk test to determine the normality of the data. When we do that, we got the p-value less than 0.05 which proves that the distribution is not normal.
- The distribution has a long right tail, it is right skewed.

T-test

Research: We want to check whether the mean for expenses is equal to 10,000.

Step 1: Ho is our null hypothesis which is "The average expense is 10,000" - Ho: μs = 10,000

Ha is our alternative hypothesis if Ho is concluded to be untrue. - Ha: μs ≠ 10,000

Step 2: The significance level = 0.05 (p-value)

Step 3: We are going to use two-tail test, because there is an indication "the mean for expenses is equal to 10,000" in the question.

```
        One Sample t-test

data:  MultiRegDataset$expenses
t = 10, df = 1337, p-value <2e-16
alternative hypothesis: true mean is not equal to 10000
95 percent confidence interval:
 12621 13920
sample estimates:
mean of x
    13270
```

Conclusions: Since p-value of 2e-16 is much lower than 0.05 confidence interval, therefore we reject the null hypothesis that μ = 10,000. Another point is that the mean 10,000 is not between the confidence intervals which are 12621 and 13920. This also proves that we reject the null hypothesis.

Simple Linear Regression 1

We have a dataset called MultiRegDataset. We will import this dataset into our RStudio and build a linear model to check whether the predictor “smoker” is significantly associated with outcome variable expenses.

Null Hypothesis: $\beta=0$, co-efficient β of the predictor is zero and not statistically significant

Alternative Hypothesis: $\beta \neq 0$, co-efficient β of the predictor is not equal to zero and is statistically significant

Simple Linear Regression 2

```
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        229    36.8   <2e-16 ***
smokeryes      23616        506    46.7   <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:  0.62,      Adjusted R-squared:  0.619
F-statistic: 2.18e+03 on 1 and 1336 DF,  p-value: <2e-16
```

The prediction equation:

(expenses) = 8434 + 23616(smokeryes)

Part I (Residuals)

The median is -919. The difference between median and min-max values are huge, so it seems a normal distribution.

Part II (Model Outcome)

Since p-value is 2e-16 which is very smaller than our significance value 0.05, we reject the null hypothesis that the predictor variable is not significantly associated with outcome variable. So, smoking has effect on expenses.

Part III (Model Performance)

R-squared value is 0.62 which is high. This means the model fit very well the data. Higher is the R-squared value, better is the model.

Multiple Linear Regression 1

We have a dataset called MultiRegDataset. We will import this dataset into our RStudio and build a multiple linear model to check whether all predictors(6) are significantly associated with outcome variable expenses.

Null Hypothesis

H0 : age = sex = bmi = children = smoker = region = 0

Alternative Hypothesis

Ha : at least one $\beta i \neq 0$ (for i = 1, 2, ..., 6)

Multiple Linear Regression 2

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

Residuals:
   Min     1Q Median     3Q    Max
-11303  -2851   -980   1384  29982

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11941.6      987.8  -12.09  < 2e-16 ***
age                256.8       11.9   21.59  < 2e-16 ***
sexmale           -131.4      332.9   -0.39  0.69325
bmi                339.3       28.6   11.86  < 2e-16 ***
children           475.7      137.8    3.45  0.00057 ***
smokeryes        23847.5      413.1   57.72  < 2e-16 ***
regionnorthwest   -352.8      476.3   -0.74  0.45898
regionsoutheast  -1035.6      478.7   -2.16  0.03069 *
regionsouthwest   -959.3      477.9   -2.01  0.04492 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6060 on 1329 degrees of freedom
Multiple R-squared:  0.751,     Adjusted R-squared:  0.749
F-statistic:  501 on 8 and 1329 DF,  p-value: <2e-16
```

Part I (Residuals)

The median is -980. The difference between median and min-max values are huge, so it seems a normal distribution.

Part II (Model Outcome)

Except sexmale and regionnorthwest, p-values of age, bmi, children, smoker, regionsoutheast, and regionsouthwest are all smaller than 0.05. We reject the null hypothesis that the coefficients of all predictors variable is zero. We can also say that sexmale and regionnorthwest is not significantly associated with expenses.

Part III (Model Performance)

Adjusted R-squared value is 0.749 which is high. This means the model fit very well the data. Higher is the Adj R-squared value, better is the model.

The prediction equation: (expenses) = -11941.6 + 256.8(age) – 131.4(sexmale) + 339.3(bmi) + 475.7 (children) 23847.5(smokeryes) – 352.8(regionnorthwest) – 1035.6(regionsoutheast) – 959.3(regionsouthwest)

Conclusions

```
Residual standard error: 7470 on 1336 degrees of freedom
Multiple R-squared:  0.62,      Adjusted R-squared:  0.619
F-statistic: 2.18e+03 on 1 and 1336 DF,  p-value: <2e-16
```

```
Residual standard error: 6060 on 1329 degrees of freedom
Multiple R-squared:  0.751,     Adjusted R-squared:  0.749
F-statistic:  501 on 8 and 1329 DF,  p-value: <2e-16
```

Higher is the Adj R-squared value, better is the model.

Both simple and multiple linear models produced high R- squared results. However multiple linear model (0.749) gave a better result than simple linear model(0.62). Based on the comparison, we suggest that multiple linear model fits better.

We showed that smoking has effect on expenses.

We also showed that age, bmi, children, smoker, regionsoutheast, and regionsouthwest has effect on expenses, but sexmale and regionnorthwest doesn't have effect on expenses. We can use backward elimination or forward selection which is useful to eliminate unrelated predictors, by doing that we can best model for our dataset.

Appendix

```
#View Dataset
MultiRegDataset
#Check the data type
str(MultiRegDataset)
#Set calculations to 3 digits
options(digits=3)
#Descriptive Statistics (Mean, Median, ...)
describe(MultiRegDataset)
#Histogram of Expenses
x=MultiRegDataset$expenses
h<-hist(x, breaks=10, col="orange", xlab="Expenses",
      main="Household Expenses")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
#Check the Normality of Data
shapiro.test(MultiRegDataset$expenses)
```

```
#One Sample t-test - two-tail
t.test(MultiRegDataset$expenses, mu=10000)
#Build Simple Linear Model
simple.fit<-lm(expenses~smoker, data=MultiRegDataset)
LinearModel<-simple.fit
#Summary of Key Statistics of the Model
summary(LinearModel)
#Create Multivariate Regression
Multimodel <- lm(expenses~., data = MultiRegDataset)
#Summary of Multivariate Regression
print(Multimodel)
summary(Multimodel)
```

Thank you