Sean Toon

IT-460-18500-M01 Machine Learning 2024 C-5

9/18/2024

2-1 Lab Assignment and Report Brief: R Linear Regression

The data set that is being use for this lab is titled “insurance” and it contains the variables, age, sex, BMI, children, smoker/nonsmoker, region and charges. The age, children, BMI and charges variables are numeric features, and the sex, smoker, and region variables are all factor-type. All the values are stored in a CSV (comma separated values) file.

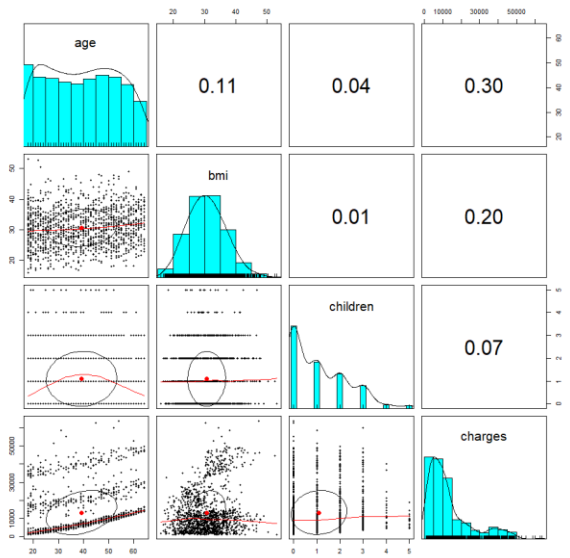
One example of an R command that I used to help my understanding of reading data from a CSV in this assignment is the following:

```
> summary(insurance$bmi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.96   26.30   30.40   30.66   34.69   53.13
```

I used the summary() command to display the bmi values stored in insurance.csv. The bmi values that were returned seem to be in the format of listing the min and max values, the 1st and 3rd quartiles, and the median and mean of all the bmi values in order.

A command that I used in the assignment and the scatter plot matrix that was generated is the following:

```
> pairs.panels(insurance[c("age", "bmi","children", "charges")])
```



By generating an enhanced scatter plot matrix, patterns in the data were a lot easier to recognize than looking at numeric variables generated in a correlation matrix and in a basic scatter plot matrix. The scatter plot matrix generated above provided histograms to depict the distribution of values for each feature, better visualizations of correlation strength with correlation ellipses, and loess curves for a better generalization of the relationship between the x and y axis variables.

Some more commands that I executed for the assignment are:

```
> insurance$bmi2 <-insurance$bmi^2
> summary(insurance$bmi2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  254.7   691.5   924.2   977.4  1203.7  2822.8
>
```

With the commands I was able to square each of the values of the bmi variable in the insurance database. Although I did not use these values in my project, squaring the values of a certain variable can be useful for a task such as improving a regression model. If one wishes to

improve model performance, one must create another beta value in this way to complete a regression equation.

More commands that I executed in my assignment:

```
> ins_model <- lm(charges~age + children + bmi + sex + region, data = insurance)


> summary(ins_model)

Call:
lm(formula = charges ~ age + children + bmi + sex + region, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max
-14768  -6966  -4918   6738  47284

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -6525.80    1840.87  -3.545 0.000406 ***
age               242.25      22.27  10.878  < 2e-16 ***
children          555.69     257.96   2.154 0.031406 *
bmi               314.60      53.53   5.877 5.28e-09 ***
sexmale          1311.61     621.52   2.110 0.035018 *
regionnorthwest -1025.98     891.33  -1.151 0.249916
regionsoutheast    69.99     895.41   0.078 0.937712
regionsouthwest -1603.32     894.46  -1.792 0.073280 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11350 on 1330 degrees of freedom
Multiple R-squared:  0.1264,    Adjusted R-squared:  0.1218
F-statistic:  27.5 on 7 and 1330 DF,  p-value: < 2.2e-16

>
```

The commands above are used for interaction effects of data. The command is useful if one wants to evaluate if any features have a combined outcome on the dependent variable. In this case, I wanted to see if age, children, bmi, sex, and region had any combined effects on the model.