Data Mining mini projest

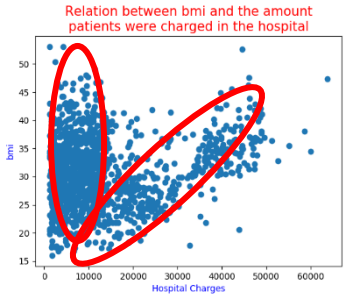
Katayoon Sadeghi mehr Feb 2020

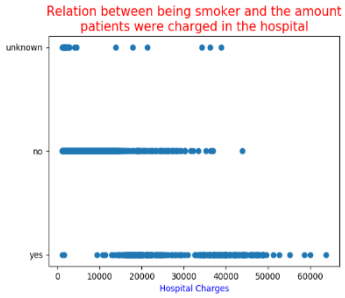
Linear Regression: Medical Cost

The dataset of insurance contains 10 columns. **ID**, **age**, **sex**, **bmi** of 1380 patients, whether they have **partner** or **children**, if they are **smoker**, the **region** they live, and hospital **charges** as dependent variable. Age has 18, smoker 32 and charges 15 missing values.
mean is used to replace age missing values, "unknown" to replace smoker missing values and the 15 rows including missing values of y=charges are dropped. After cleaning we have 1365 rows of data. Encoding on categorical variables, sex with 2 levels male=1, female=2, , partner with 2 levels yes=1, no=0, smoker with 3 levels unknown=0, no=1, yes=2 and region with 4 levels northeast=1, northwest=2, southeast=3 and southwest=4 is done.

ID	age	sex	bmi	children	smoker	region	charges	partner	sex_C	smoker_C
10001	27	female	31.4	0	yes	southwest	34838.87	yes	2	2
10002	39	female	39.71	0	yes	northeast	13143.34	no	2	2
10006	23	female	28.31	0	yes	northwest	18033.97	no	2	2

Correlation and association between variables:





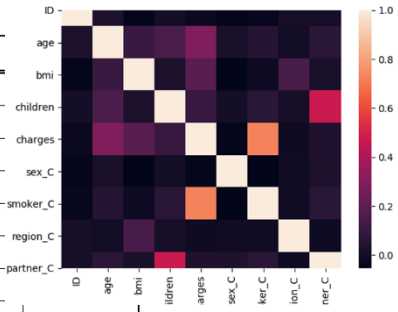
As the age increases, the amount of charges also increases.

For the amount of less than 10,000, bmi doesn't show a specific association with charges, but as the amount increases, patients tend to have higher bmi.
Smoker patients have a bigger variation over their hospital charges and the highest charges belong to smokers.

Being smoker and the amount of charges has the highest correlation.

Also number of children and having a partner is correlated.

	ID	age	bmi	children	charges	sex_C	smoker_C	region_	
ID	1								
age	0.0250	1.0000							
bmi	-0.0423	**0.1061**	1						
children	0.0037	**0.1525**	0.0273	1					
charges	-0.0249	**0.2873**	**0.1859**	**0.1006**	1				
sex_C	-0.0417	0.0194	-0.0500	0.0012	-0.0447	1			
smoker_C	-0.0363	0.0517	-0.0151	0.0626	**0.7244**	-0.0533	1		
region_C	0.0110	-0.0011	**0.1520**	0.0100	-0.0082	-0.0055	-0.0058	1	
partner_C	0.0091	0.0642	0.0202	**0.4705**	0.0335	0.0326	0.0633	-0.0155	1



Linear Regression Model:

X = [age, bmi, children, sex, smoker, region, partner]

y = charges

y = -26562.1157 + 200.596 age + 335.825 bmi + 384.039 children - 44.8978 sex_C + 19319.8 smoker_C – 376.116 region_C - 1133.59 partner_C

R2 (Train)= 0.6043 R2 (Test)=0.6621

Children has the highest pvalue which makes it insignificant and the next insignificant X is bmi.

y = -17126.4998 + 221.689 age - 276.906 sex_C + 19300.7 smoker_C - 135.451 region_C - 709.801 partner_C

X	P-Value
age	0.000
sex_C	0.000
smoker_C	0.000
region_C	0.000
partner_C	0.010

With 1365, **5-fold** regressions were developed to check the validation on the model. The values are 0.6272, 0.5068, 0.6075, 0.5257, 0.6265 with average of **0.5787.**

R-Square from final test model is **0.6442. R2-Adj= 0.6481**

$$\bar{R}^2 = 1-\left(1-R^2\right)\left[\frac{n-1}{n-\left(k+1\right)}\right]$$