Advanced econometrics : Medical expenses

Bruguet
Bulliard
Pion

October 2021

1 Introduction

Medical insurances are really interested in forecasting medical costs, in that way they could build more accurate actuarial tables which allow to set the price of yearly medical premiums higher or lower depending on the expected treatment costs. But modeling medical expenses can also be useful for states in countries where the basic medical care is covered by public services.
However, medical expenses can be hard to estimate because the most costly diseases are rare and random. Still, some conditions are more prevalent for certain segments of the population. For instance, lung cancer is more likely among smokers than non-smokers, and heart diseases may be more likely among the obeses. The goal of this study is to use patient data to estimate the average medical care expenses for these population segments.

Data relies on demographic statistics from the U.S. Census Bureau and were designed for the book **Machine Learning with R (Lantz, 2013)**. It is composed of 1 338 observations, each one is a beneficiary enrolled in the insurance plan in 2013, with features indicating characteristics of the patients as well as the total medical expenses charged to the plan for the year. More precisely, the available variables are :

- *charges* : the annual medical expenses in US dollars
- *age* : the age of the primary beneficiary, excluding those above 64 years old since they usually are covered by the government
- *sex* : either male or female
- *bmi* : Body Mass Index (BMI), it provides an estimation of the proportion between heigth and weight for each individual. An ideal BMI is between 18.5 and 24.9.
- *children* : number of children covered by the insurance plan
- *smoker* : a dummy variable that takes yes if the insurred regularly smokes and no otherwise
- *region* : place of residence in the U.S., there are four geographic categories : northeast, southeast, southwest and northwest.

An intuitive reflection will easily lead to think that smokers, older beneficiaries and those with a higher BMI will tend to have higher medical expenses. This econometric study should help us to infirm or confirm these hypotheses.

2 Data description

Our database in composed of 2 continuous quantitative variables (bmi and charges), 2 discrete quantitative variables (age and children) and 3 qualitative variables (sex, smoker and region). In order to implement econometrics models, we'll binarize the qualitative variables. sex and smoker contains two modalities, 'yes' or 'no', so we'll transform those in 0 and 1. region contains 4 modalities 'northeast', 'northwest', 'southeast' and 'southwest', so we'll create 4 variables for each of these modalities.

Table 1: Descriptive table of data

	charges	**sex***	**bmi**	**children**	**age**
count	1338	1338	1338	1338	1338
mean	13270.42	0.51	30.66	1.09	39.21
std	12110.01	0.5	6.1	1.2	14.05
minimum	1121.87	0	15.96	0	18
1st quartile	4740.29	0	26.3	0	27
median	9382.03	1	30.4	1	39
3rd quartile	16639.91	1	34.69	2	51
maximum	63770.43	1	53.13	5	64
	smoker*	**northeast***	**northwest***	**southeast***	**southwest***
count	1338	1338	1338	1338	1338
mean	0.2	0.24	0.24	0.27	0.24
std	0.4	0.43	0.43	0.45	0.43
minimum	0	0	0	0	0
1st quartile	0	0	0	0	0
median	0	0	0	0	0
3rd quartile	0	0	0	1	0
maximum	1	1	1	1	1

* : Binary variable

On this table, we can look at basic statistics over our ten variables. We'll comment on the interesting values.

Our explained variable, charges, has a mean of 13270.42. This means that in average in our sample, individuals have been charged with a total of 13 270 dollars in 2013. 75% of the sample paid more than 4740 dollars while the maximum value is 63 770 dollars. For the explanatory variable sex, the mean informs us that our sample contains 51% of males. The minimum bmi is 15.96 among all individuals. 50% of them have a body mass index superior to 30.4. The individual with the maximum value has 53.13. In our sample, 75% of the individuals have less than 2 children. The maximum number of children is 5. The average age is 39.21 years old. The individuals in the database are between 18 and 64 years old. About the variable smoker, there is 20% of smokers among the individuals. Finally, concerning the dummy variables of the variable region, the majority (27%) of individuals live in the south-east.

Figure 1: Correlation matrix

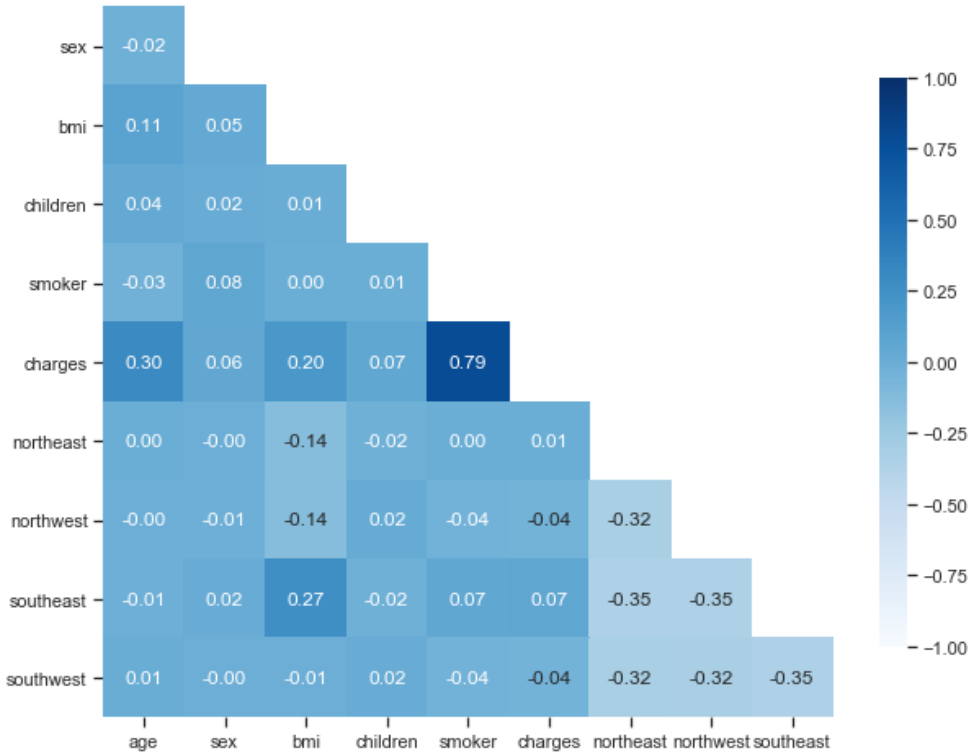
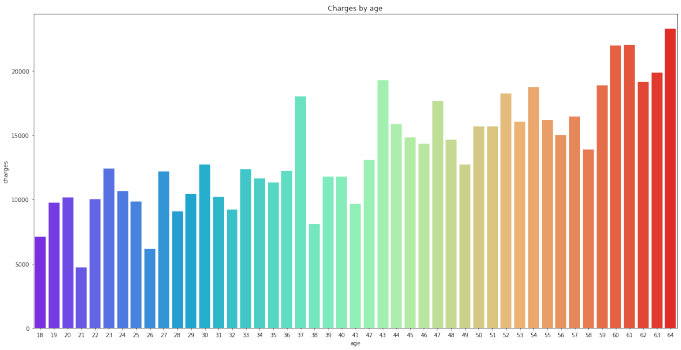


Figure 1 shows the matrix of correlation of our variables. As we can see, the strongest correlation concerns our explained variable, the medical expenses, with the fact that the individual smokes or not. The two variables are strongly positively correlated, which means that smoking has a great positive impact on our medical bill. We can explain this with the number of cardiovascular and respiratory diseases that can occur with the frequent consumption of tobacco.
The age of the individual is also midly correlated with the medical expenses. The correlation is also positive, which is again not a strange result. While an individual is growing old, diseases are more frequent which can lead to higher medical expenses as time goes by.
Finally, medical expenses seem to be wealky correlated with the bmi. The body mass index being a metric of a healthy body, a positive correlation between these two variables is also not a surprising result. The higher the bmi, the higher the chance to develop diseases and medical expenses.

Figure 2: Charges by age



With **Figure 2**, we can see the relation between age and charges. As we have seen with the correlation between these two variables, there is a tendency of increasing medical expenses as the age grows. The age with the lower medical bills is 21 while 64 years old individuals pay the most.

Figure 3: Charges by bmi

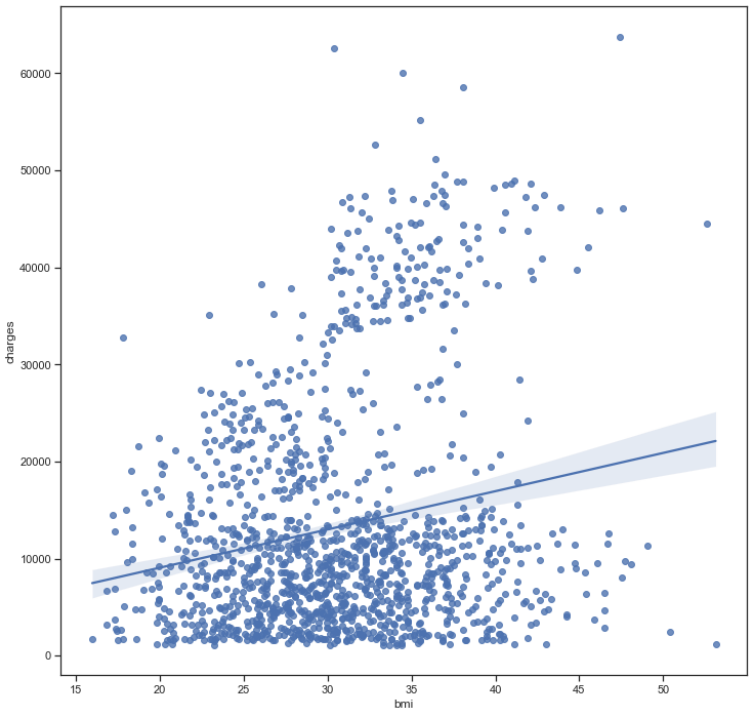


Figure 5 displays the relation between the number of childrens and our explained variable. Individuals with the highest medical bills seem to have 2 or 3 childrens. On the contrary, people with 5 childrens have the lowest medical expenses. This must be due to the fact that families with a lot of childrens can have financial support for medical expenses.

Figure 4: Charges by smoking habits

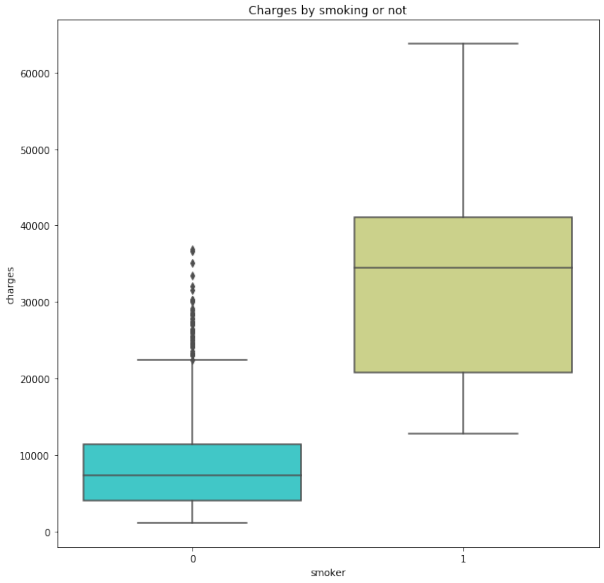


Figure 4 allows us to see the difference in charges between smokers and non-smokers. Here, there is no doubt that smoking has a lot of impact on medical expenses. The distribution of the two boxplots is totally different and smokers pay a lot more in medical bills than non-smokers.

In **Figure 3**, we plot each bmi observation and the corresponding charges. We also plot the linear regression line in order to have an idea of the relation between the two variables. As we have seen with the weak correlation coefficient between them (0.2), there is a slight positive relation between the two variables. When the body mass index increases, medical expenses increase a bit.Then, the two boxplots in **Figure 6** allow us to tell the difference in medical expenses between males and females. As we can see, males seem to pay more than females. This result is quite surprising. In fact, most studies show that women pay on average more health care costs than men. Since women live longer than men, they are more likely to need more medical treatment than men. Maybe our small sample isn't representative of the reality.

Figures 7 and 8 look at the distribution of the dependant variable charges. On original data the distribution is a right-skewed distribution. However, once it is transformed to log variable, the distribution becomes a Normal distribution centered in $log(9)$.

Figure 5: Charges by number of childrens

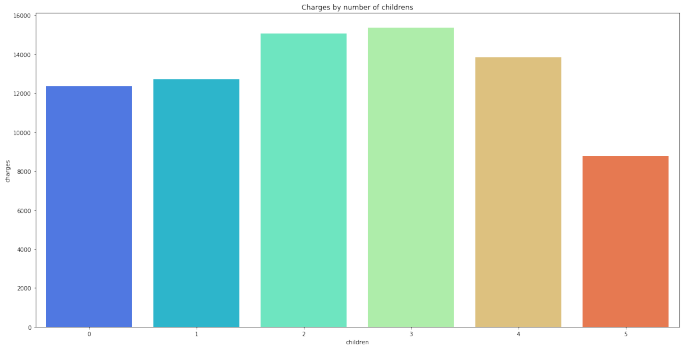


Figure 6: Charges by sex

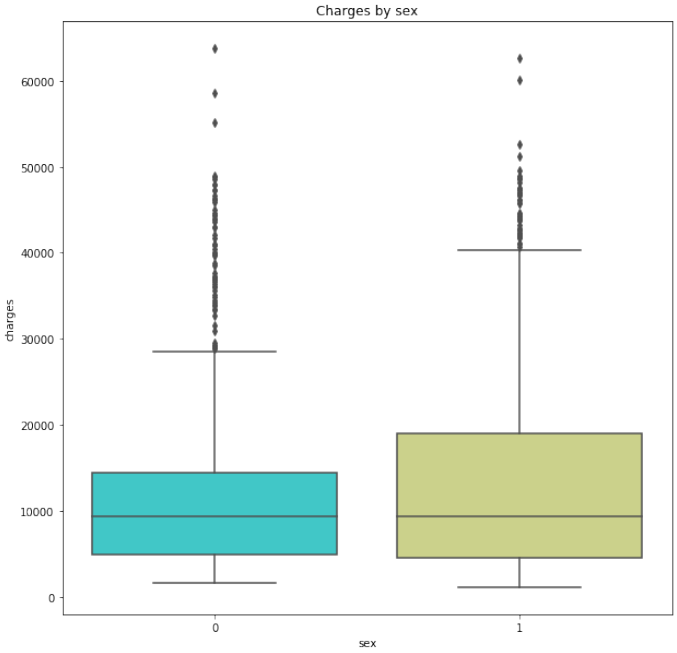


Figure 7: Skewed distribution

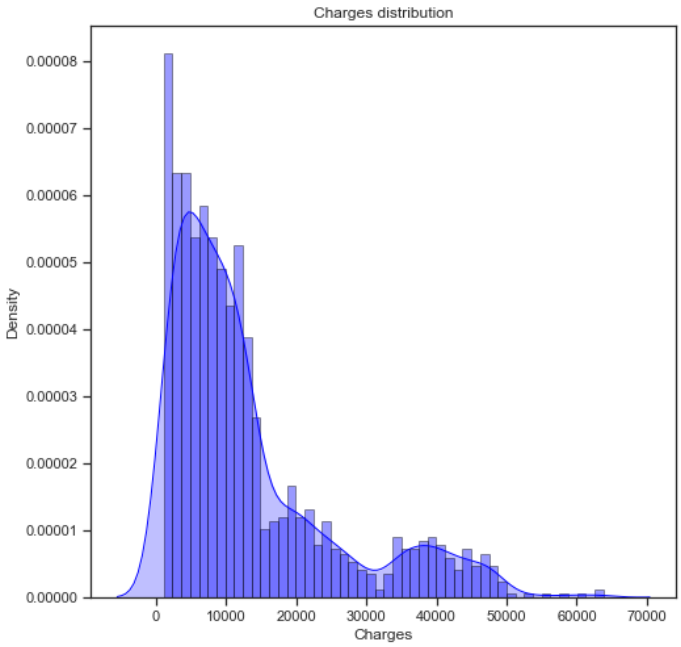
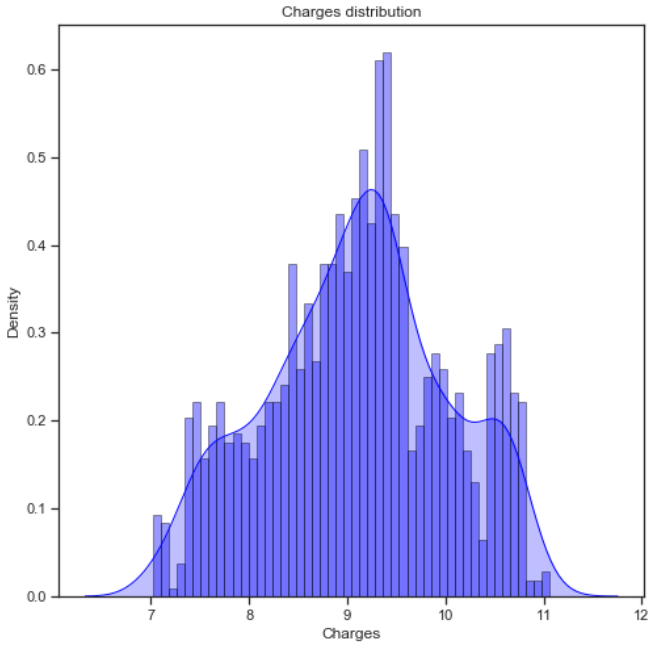


Figure 8: Normal distribution



3 First parametric model

3.1 Modeling

We denote our model as a log-level linear regression. This allows us to conveniently interpret the coefficients in terms of semi-elasticities. This specification induces that an increase of one unit increase y by $(\beta_1 * 100)\%$. Moreover it allows to have a normal distribution of the dependant variable.
Moreover, to avoid the dummy variable trap, we include an intercept but we keep only 3 region out of 4 for the geographical variable. That way there is no multi-collinearity between these variables.

Our initial model is the following :

$$ln(charges) = \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 bmi + \beta_4 children + \beta_5 smoker + \beta_6 region + \epsilon$$

We consider X the matrix of dependants variables and U the vector of errors, the usual assumptions are supposed to hold :

- H_1 : Model is correctly specified
- H_2 : $E(U|X) = 0$
- H_3 : The variance of the error is constant over observation $V(U|X) = \sigma^2 I$
- H_4 : There is no strict multicollinearity between the explanatory variables i.e. the matrix X'X is of full rank and invertible.
- H_5 : The error term is not correlated with the independent variable $cov(X, U)$

We assume the hypotheses we made are satisfied, we neglect endogeneity and heteroskedasticity or auto-correlation. We perform OLS to have a benchmark and a first intuitive estimation.

3.2 Estimations and interpretations

First of all, our linear regression has only one variable for the sex and smoker and three geographical variables. It means that the female, non smoker and the region northeast are take into account in the intercept, they are part of the reference group, when all other variables are equal to zero. However our intercept is here not interpretable but it is important to keep it for statistical purpose.[1]
Then looking at the t-Student test, except the northwest, all variables are statistically significant, i.e. different from zero, with a confidential interval of 95%. Moreover the ones that seem to contribute more to influence the medical expenses are the age, the smoking habits and having children. Indeed, one more year leads to an increase of 3% of the expenses. Having one more child induces an increase of 10% in the annual medical expenses. Finally, if the beneficiary smokes, this increases the potential medical expenses of 155%.
The geographic emplacement have very low impact compared to these three first variables. However, the sex and weight also have some influence. Being a male allows you to reduce the medical expense of 7% during a year, compared to females. Then, having one more point of BMI is equivalent to increasing the medical expenses of 1% along the year.

[1] We also estimate the model without intercept by OLS and the value of the intercept was refelcted in the geographical variables, that have become uninterpretable

Table 2: First linear model estimated by OLS

Coefficients	**Estimate**	**Std**	**t-stat**
intercept	7.0305	(0.072)	97.112
age	0.0346	(0.001)	39.655
sex	-0.0754	(0.024)	-3.091
bmi	0.0134	(0.002)	6.381
children	0.1019	(0.010)	10.085
smokeryes	1.5543	(0.030)	51.33
northwest	-0.064	(0.035)	-1.827
southeast	-0.157	(0.035)	-4.481
southwest	-0.129	(0.035)	-3.681
Residuals	Q1	Median	Q3
	-0.1983	-0.0492	0.0660
R-squared	0.7679		

The R-squared of this model is 0.7679, thus around 77% of the expenses seems to be explained by these variables. Looking at the residuals computation, we have a residual median of -0.04. This correspond to the median error between the true value and the predicted value :

$$log(y) - log(\hat{y}) = -0.04$$
$$log(\frac{y}{\hat{y}}) = -0.04$$
$$\frac{y}{\hat{y}} = exp(-0.04)$$
$$\frac{y}{\hat{y}} = 0.96$$
$$\hat{y} = \frac{y}{0.96}$$

In other words, the model over-estimate the expenses at a 4% level for 50 percent of the predicted observations. On the other hand, 50 percent of errors fall within the the first and third quartile, so the majority of predictions were between 19% over the true value and 6% under the true value.
Plotting the residuals distribution allows to have a better view of the situation(**Figure 9**).
We have a right-skewed distribution of residuals, they are not following a normal distribution as it should do, even if it can be noticed that the distribution is center around 0. This is non gaussian residual and is an hint that we missed some non-linearities in the way we model our regression equation.

The model can surely be improve, first we assume all our initial assumption were true and we neglect major classical modelisation problem. The main point we focus on in this study is some non-linearity relationship that should be taken into account, as shown by the distribution of the residuals.

4 Nonparametric model

We previously estimated a parametric model. We had to specify the model and the hypotheses. However, our modelisation is arbitrary and is based on our intuitive view of the topic, our own view of how to model health care insurance charges with the covariates at our disposal. We could have missed some non-linearities. To test if this is the case, we will now use a non-parametric model. Here we do not make any assumption about the shape of the function. We will implement a Generalized Additive Model (GAM) model. Make assumption that our model can be decomposed as a sum of several functions of one variable. We will apply spline method.

We will implement a GAM analysis using spline, the goal is to find a non-linear relationship between the independents and the dependent variable. The GAM estimator is additively separable. This is an iterative estimation process where we assume it can be decomposed as a sum of several functions of arbitrary shape.

A linear model will be :

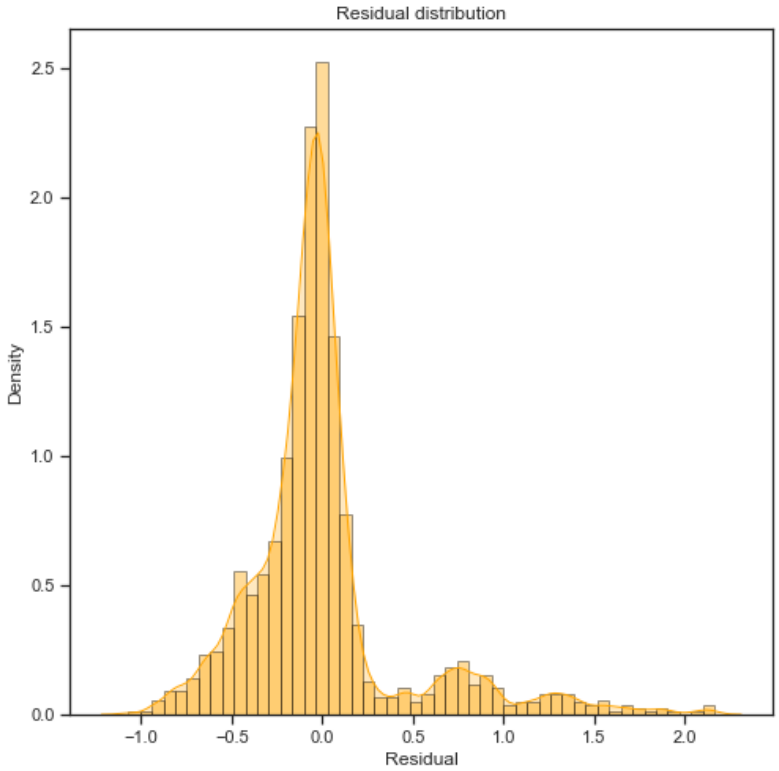
$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

With GAM what changes is the presence of a smoothing term that replace the coefficient of the independent variable :

$$y = \beta_0 + f(x_1) + \epsilon$$

The additive separability hypothesis is central to understand the pros and cons of the method. As GAM allows to only estimate regression of dimension 1, we may omit interaction terms. The main advantage is that it allows us to understand and visualize the partial relationship between the covariates and the dependent variable. Indeed, as we have a sum of complex functions of one variable, we can easily plot the 2D graphs for each covariates. Another disadvantage of the GAM model is that we can't easily interpret the parameters,

Figure 9: Distribution of residual from the linear model estimated by OLS



we can't make any inference. Indeed, we release the assumption that increasing the value by one always has the same effect on the predicted outcome. The interpretation is less clear.
Moreover, complex functions such as polynomials have unpredictable tail behavior which is bad for extrapolation. If we have new data out of the sample, our model will be powerless to predict.
We look for the form of the smooth function f(), it can be something such as x_1^2, a polynomial of degree 2 or something more complex. We use smoothing spline to reduce the influence of the number of knots chosen, because when there are too many knots, the spline is not smooth. Thus the criterion function is penalized. The penalty term is λ that allows to reduce the fluctuation between the predicted $\hat{y}$ and y. The goal is to select λ such that the distance between the predicted value and the real value of y is a minimum. We show the formula of smoothing splines which is convenient to illustrate the tradeoff

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int (s"(x))^2 dx$$

The first term shows closeness to the data and the second term penalize the curvature of the function and λ is the measure of the trade-off. We minimize the residuals to fit the data and the penalization imposes smoothness. Then, we have a trade-off between the smoothness of the function and the fitting of data. The higher the value of the penalization term lambda, the smoother the curve. The choice of this smoothing parameters can be done by cross-validation and that is want we do by using *Python* to estimate the GAM estimator, with PyGam librairy. There are several ways to estimate splines. Here, we will use univariate splines which is available in the librairy.
We estimate :

$$lcharges = \beta_0 + f_1(bmi) + f_2(age) + f_3(smoker) + f_4(sex) + f_5(children)$$

And suppose that the region variables have a linear relationship with the dependent variable.

Table 3: GAM estimation

Coefficients	**EDF**	**Signification**
intercept	0.0	0
$f_1(bmi)$	7.7	0
$f_2(age)$	7.5	0
$f_3(smoker)$	0.9	0
$f_4(sex)$	1.0	0.001
$f_5(children)$	3.8	0
Lamnda	784.76	

To interpret the summary results of a GAM estimator, first if the EDF here is close to 1 there is linearity. If not, it can have non-linearity and the larger is the EDF the more wiggly (not linear) the fitted model is.
Then looking at the t-student, there is non linearity if the H0 hypotheses is rejected with :

$$\begin{cases} H_0 & \text{Linear} \\ H_1 & \text{Non-linear} \end{cases}$$

Thus, the sex and the smoker variables have a linear relationship while BMI, AGE and CHILDREN seemed to be non linear. We have plotted the respective non linearities given by GAM. In the next section we will discuss about how to estimated these relations.

4.1 Function plot from the GAM

Figure 10: Smooth function of BMI

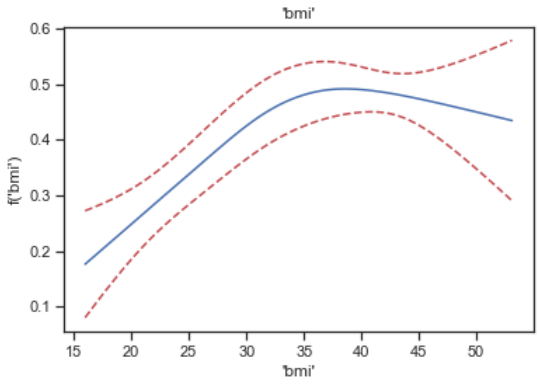


Figure 11: Smooth function of AGE

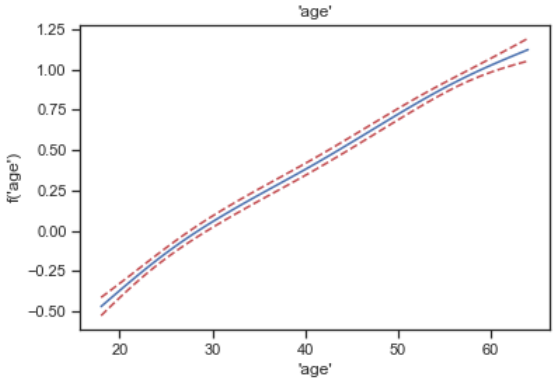
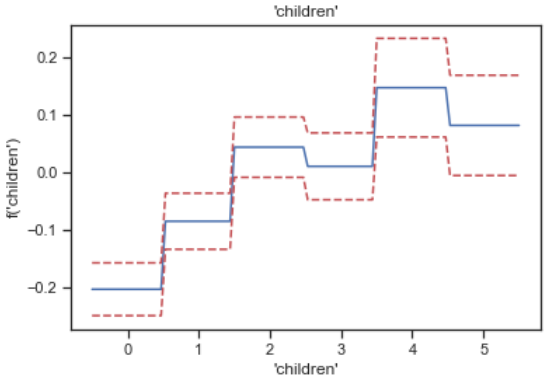


Figure 12: Function of CHILDREN



5 Comparison between the two approaches and improvement of parametric modelling

As we have seen in the previous questions, the GAM model revealed missed non linearities for 3 of our variables : bmi, age children. Here, we're trying to find another parametric model that would give better results than our first one, by taking into account non linearities. To do so, we'll transform the variables to fit as best as we can their non linearities.

In order to find the best transformation for our variables, we must look for simple known functions to approximate the partial effect plot. Here for bmi, we can see that the logarithmic function seems to fit pretty well with the partial effect plot. In our new parametric model, we will add log(bmi) as a variable to replace bmi, to account for its non linearities. Since our explained variable is already in logarithmic form, the predicted coefficient of our model should be interpreted as an elasticity.
For our variable age, we can see that the curve is 'wiggly'. There are lots of ups and downs, as in the sinus function.

Figure 13: Estimation of f(bmi)

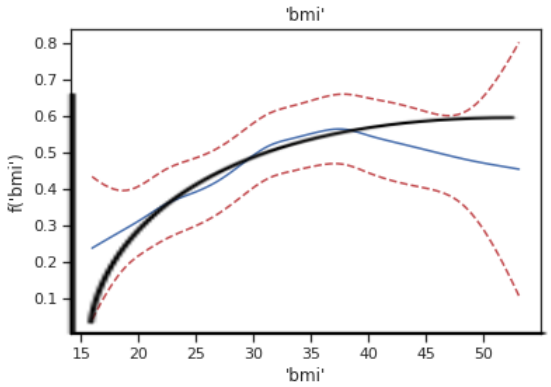


Figure 14: Wave of a sinus function

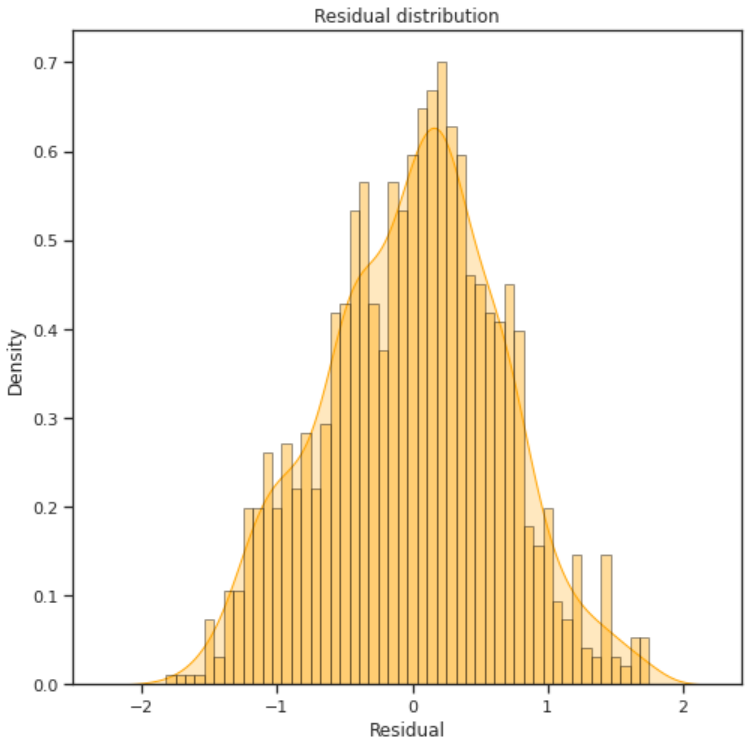
To transform this variable, we will take the sinus of **age** to replace **age** in our parametric model. This way, the data should be more fitted than with a straight regression line from a classic OLS model. Here, the main drawback is the lack of interpretability. But with this transformation, we'll see that the residuals of our new parametric model are much more gaussian than before, so we decided to keep it that way.
In order to take into account the non linearities of the variable **children**, we have tried 3 different approaches: don't transform the variable, taking it's sinus and taking it's log. For each one, we've compared the results and we've decided to keep the logarithmic transformation. Indeed, replacing children by log(**children**) gives the best results in terms of normality of the residuals, and allows us to interpret the result as we've seen with **bmi**.

Table 4: Second linear model estimated by OLS

Coefficients	**Estimate**	**Std**	**t-stat**
intercept	6.3238	(0.310)	20.370
sage	0.0162	(0.026)	0.631
sex	-0.0999	(0.036)	-2.785
lbmi	0.7258	(0.092)	7.894
lchildren	0.2650	(0.032)	8.260
smoker	1.5285	(0.045)	34.303
northwest	-0.077	(0.051)	0.132
southeast	-0.212	(0.052)	-4.120
southwest	-0.1437	(0.052)	-2.784
Residuals	Q1	Median	Q3
	-0.4483	0.0457	0.4425
R-squared	0.498		

The model now under-estimate the expenses at a 4% level for 50 percent of the predicted observations. On the other hand, 50 percent of errors fall within the the first and third quartile, so the majority of predictions were between 44% over the true value and 44% under the true value.
Plotting the residuals distribution allows to have a better view of the situation.

Figure 15: Distribution of residual with the non-linear variables



By looking at the estimated coefficient, the age becomes non-signifiant once it is passed to a sinus function. Then, the smoker parameter is still the strongest one, if an individual smokes it increase the expenses of 153%. Then having one more children or one more point of BMI respectively leand to an increase of 0.26% and 0.72% of the medical expenses.
The main point is that once the non-linearity is introduce, the residuals distribution becomes Gaussian and centered in zero. Thus the approximation of non-linearity made with the GAM model seems to be rather good.
The R-Squared of the model is now 0.49 thus only 49% of the medical expenses can be explained by this model. However even if the first estimation had a better R-Squared, it was probably biased. Considering the available data and the randomness medical expenses can have, it seems reasonable that the estimation of the model is around 50%.

Using the GAM model to find non-linearity in our independent variable modelisation allow us to respect classical hypotheses about the estimation by OLS. Our estimations are less biased in the second parametric analyse we propose. However, we didn't have check the satisfaction of all the fives hypotheses we assumed. Thus the estimation can be considered as a better one but surely not the best one possible.

6 Appendices

6.1 Table of contents

Contents

6.2 References

Choi.M (2017) . The insurrance data set in Kaggle. Kaggle.

Lantz.B (2013). "Machine Learning with R". 172-183.

Shafi.A (2021)."What is a Generalized Additive Model ?". Towards Data Science.

Shen. I (2018)."Building interpretable models with Generalized additive models in Python " . Medium.

Tenand.M . "Point sur les spécifications en LOG". Paris School of Economics.