Politecnico di Milano

Course in Nonparametric Statistics

Application Oriented Project

February 3, 2021

Great deals for a great insurance

Evaluating the right premium for health insurance in the United States

Authors:
Valentina Buccioni (969816)
Martina Garavaglia (967257)
Anna Iob (963336)
Veronica Mazzola (963106)

Abstract

Our project purpose is to create a web page which evaluates the fairness of health insurance premium for US citizens, based on physical characteristics and lifestyle habits. Firstly, we analyze our dataset variables, investigating which ones are the most significative for our aim using nonparametric tests, such as ANOVA and permutation tests. Secondly, we build the best fitting model with Generalized Additive Models (GAMs), conducting a stepwise procedure, and we make premium predictions. Lastly, we implement the web application with the R package "Shiny" to assess prices for our web audience.

Contents

1 Introduction

1.1 Our aim

The purpose of this research is to develop an innovative solution to check the fairness of health insurance premiums in the US.
Sickness insurances are so expensive since are exposed to adverse selection risk. It occurs when individuals with greater health care needs are more likely to purchase health insurance than healthier people.
Moreover in the United States the costs of health care are constantly increasing because of the expensive and specialized treatments that are offered [1]. Therefore, is extremely important for US citizens to choose the best health insurance.
Our web app "US Health Insurance Calculator" provides a performant tool to help users to make their best choice according their medical needs.

1.2 Description of the dataset

For the aim of our research, we have chosen a dataset that compares annual health insurance premium between different beneficiaries from United States.
The dataset [2] we have used is available on "Machine learning with R", a book by Brett Lantz.
It has 1338 observation and 10 variables:

- *Charges*: individual medical costs billed by health insurance (insurance premium, from 1122$ to 63770$);
- *Age*: age of primary beneficiary (from 18 to 64 years old);
- *Sex*: insurance contractor gender (female or male);
- *Smoker* : dichotomous variable (equal to 1 if the claimer smokes, 0 otherwise);
- *Children* : number of children covered by health insurance (from 0 to 5 children);
- *Region*: the beneficiary's residential area in the US (northeast, southeast, southwest, northwest);
- *Bmi*: body mass index, providing an understanding of body composition (value between 15 kg/m^2 and 54 kg/m^2).

Additional variables:

- *Steps*: variable indicating the average walking steps per day of policy holder (from 3000 to 10000 steps);
- *Insuranceclaim*: if the policy holder made a claim the variable is 1, otherwise is 0;
- *Coverage*: dummy variable created by us that is 1 if the insurance holder has children covered, 0 otherwise.

1.3 Statistical approach

Parametric assumptions do not hold in our sample: in general, there is not a linear relationship between the response *charges* and the other predictors and moreover residuals are not normally distributed in any of our models. For this reason, we developed our research using nonparametric methods.

2 Data visualization

Among our previously described dataset covariates, *charges* represents a reasonable choice of response variable for our research.
We started looking at its distribution according to our two continuous variables: *age* and *bmi*. However, the variable that mainly affect the response is *smoker*. The impact of this variable on the distributions of *charges* with respect to *age* and *bmi* can be seen in Figure 1.

Figure 1: charges vs age and charges vs bmi

Using depth measure methods, we furthered the study of the two continous variables.
Bmi has a quite symmetric distribution and for this reason we can compute its depth using both the Tukey and the Mahalanobis depth. Both the definitions give as median the value 30.80 for *bmi*, instead the median for *charges* is 9414.92 with the Tukey depth and 13390.56 with the Mahalanobis.
For what concerns *age*, data are quite uniformly distributed across the possible values from 18 to 64 and in the bivariate distribution (*age,charges*), the Tukey median is 41 for *age* and 7954.5 for *charges*.
Detection of outliers has been made still using the depth approach and representing data through bagplots. As it can be seen in Figure 2, in the lower right-hand corner there are two outliers in the distribution of *charges* with respect to *bmi*. These outliers have low charges and high bmi. As expected, both are not smokers, indeed

smoking raises the costs of insurance. They come from the same region (southeast), they are both males and they are young.
We decided to not remove them from our dataset since obesity does not always directly affect the health insurance premium, which instead depends on the whole health state of the person.

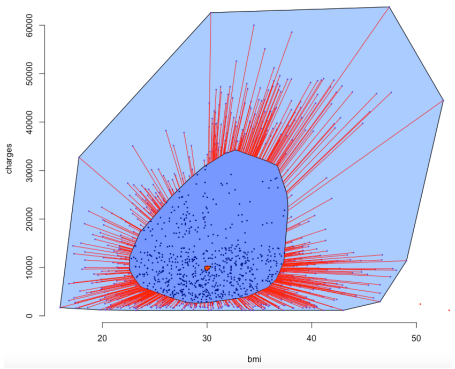

Figure 2: Bagplot charges vs bmi

We proceeded the analysis with the study of the categorical variables. *Children* is a very relevant and tricky variable; indeed, it describes how many people are covered by the health insurance of a single person and so it affects the response, but at the same time it has not correlation with other variables that instead describe only the single policy holder. This creates some issues because the target variable *charges* is the result of the claims of all people covered and so it is more difficult to give interpretability to the other predictors.
As expected, we found with a permutation ANOVA test that the distributions of *charges* in the different groups of *children* is not the same (see Figure 3). We used the permutation method since the distribution of the sample is not normal.
Next, we tried to run the same test only in the 5 groups which actually have a family health insurance, so only for *children* greater than 0. The result was interesting since we can accept the null hypothesis with a p-value of 33%. Therefore, we discovered that people who take out a family health insurance pay on average the same. This allowed us to create a new dummy variable, named *coverage*, that takes the value 1 if the person applies for a family plan and 0 if he asks for an individual plan.
Coverage represents well the differences between the costs of health insurances. Indeed, the plan's deductible and out-of-pocket maximum, that is what will be paid at most during a policy period for health care services, are based on whether you have an individual or family coverage. The deductible and out-of-pocket maximum for a family plan are greater than those for an individual plan, no matter how many people the plan covers [3].
Finally, we have done Mann-Whitney U tests and permutation tests for the remaining variables. The test statistics we used for the permutation tests is the absolute value of the difference between the sample means of every group, except for *sex*, for which we used the absolute value of the difference between the sample median since it is a more robust estimator. As a result, we found out that neither *sex* nor *region* affect the response, while *insuranceclaim* and *smoker* have a significant impact on *charges*, as anticipated. We also did a permutation ANOVA for *steps* and discovered

that the response is affected by the number of steps per day.
With regard to this results, the next step is to find the best fitting model for our data.

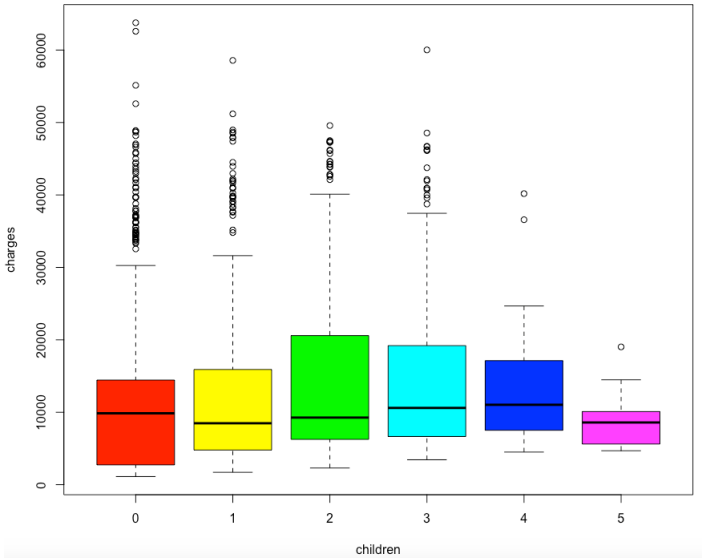

Figure 3: Original data

3 GAM

3.1 Model selection

We have used Generalized Additive Models (GAMs) to construct the best fitting model for our data. We have made this choice because assumptions for parametric multiple regression do not hold, as stated in the introduction section.
In particular, we want to study our response variable *charges* with respect to the others: quantitative variables are *age*, *bmi* and *steps* and dummy variables are *smoker*, *insuranceclaim* and *coverage*.
We do not consider the variables *region* and *sex* because, as we have discussed in the previous chapter, are not significant for our research.
The initial model is the very complete model with all variables and all possible interactions. Looking at this first summary, we observe the R-squared adjusted parameter equals to 86.3%.
By testing the residuals with a Shapiro test we have observed they are not gaussian. Therefore we have used a backward selection procedure giving less importance to the p-value of each term than the R-squared adjusted. Indeed, we have observed the former only to determine the removal order, wheareas the latter to choose firmly whether to keep the variable or not.
Thus we have removed the interactions between quantitative variables: the R-squared adjusted remains unchanged, so we have deduced they are not significant. Same happens for the two dummy variables insuranceclaim and coverage and for the following interactions: each quantitative variable and *insuranceclaim*, *steps* and *coverage*, *age* and *smoker*. So the final model is:

$$charges \sim age + bmi + steps + smoker + bmi : smoker$$
$$+steps : smoker + age : coverage + bmi : coverage$$

In this model we have used cubic splines (cr) for all terms except for *bmi*, *steps* and *smoker*. Indeed the first two terms are linear since the equivalent degrees of freedom equals one, meanwhile smoker is a significant dummy that appears only as a parametric term.

For reasons concerning the distribution of *steps* we have reduced the number of spline knots from 9 to 5, corresponding to the number of distinct density regions of its scatterplot.
By observing the summary, the R-squared adjusted is still 86.3% and all variables are significant, except for *bmi* and *steps* that we have decided to keep for hierarchical reasons.

3.2 Interpretation of regressors

In order to understand the behaviour of the regressors of our model, we have decided to make a qualitative evaluation on their plots exploiting the properties of additive models and adding one dummy variable at a time.

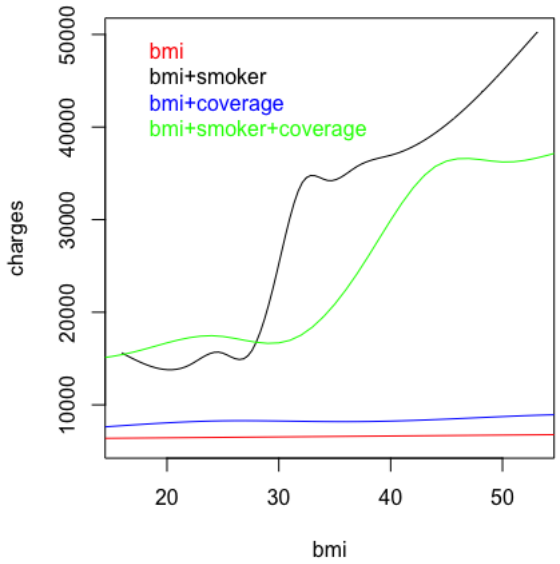

Figure 4

Figure 4 shows the trend of *bmi* in relation to *smoker* and *coverage*. We can see that *bmi* has a linear upward behavior, with little slope. Indeed, weight gain does not affect much charges as smoking does. *Bmi* in connection with *smoker* has a strong non-linear behavior and it has a steep increase from value 26/27 kg/m^2 and less steep one from 32/33 km/m^2, which means that a little weight gain or

loss impacts more on an overweight smoker's premium then on an obese's one. The interaction increases *charges* from a bmi of 30 kg/m^2, the threshold value between overweight and obese category. Moreover, the interaction between *bmi* and *coverage* has a small effect on the growth of *charges*. When both dummy variables are equal to 1, we have an increase of the insurance premium in the obese category and then a stabilization. For the values greater than bmi equals to 30 coverage has a negative impact on the premium, maybe because people with children and overweigth have a low income and so their insurance premium is cheaper.

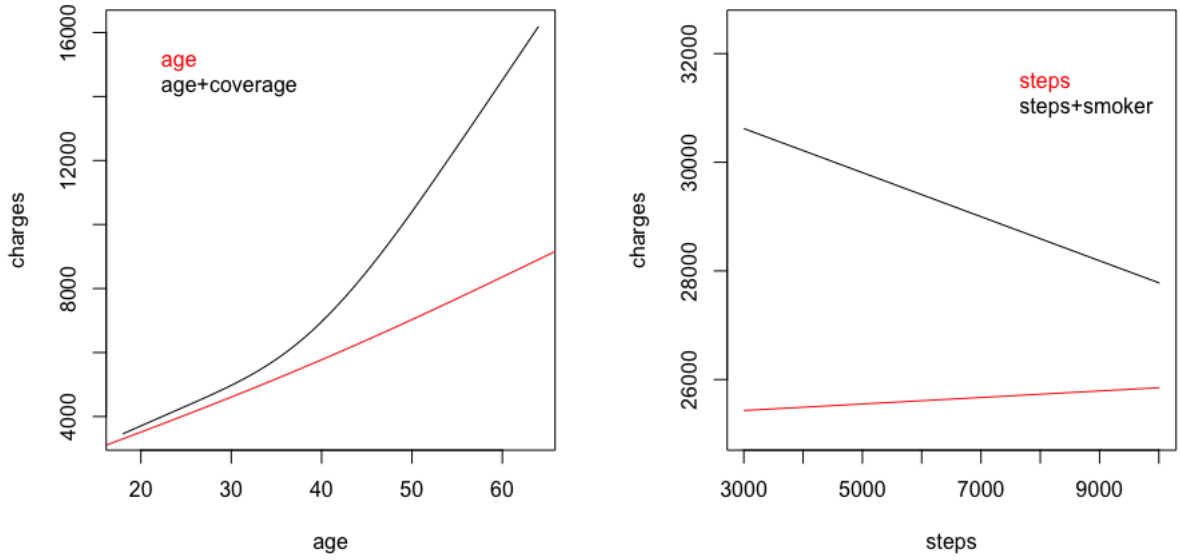

Figure 5: *Plot of age and steps*

In Figure 5 (left-hand side) we can investigate the trend of *age* in relation to *coverage*. *Charges* has an upward linear behavior because people are more prone to have health issues as *age* increases. The same happens in correlation with *coverage*. Indeed, the oldest component of the insured with children covered may be subjected to a strong increase in the premium, maybe because there are holders with children with disabilities.

As we can see in Figure 5 (right-hand side), *steps* has a premium increase, but too small to be significant. Smoker holders pay way more than non smokers, but there is a decreasing linear trend as people with a greater daily number of steps have a less sedentary and healthier life.

4 Prediction

In order to make prediction of our GAM model and to determine whether a price is fair or not, we compute some confidence intervals around different predicted values obtained with the *predict.gam* function.
We have thought to implement a split conformal or a bootstrap method to evaluate

prediction intervals, but then rejected this idea: for our objective the web application must be both fast and precise in assessing results, but these methods applied to our model did not attain these features. Meanwhile *predict.gam* has allowed us to obtain a pointwise prediction ensuring high performance, so we have opted for it. Initially we have decided to compare a smoker with a not smoker and results, reported in Table 1, confirmed that smoker is a decisive factor for our model. Moreover, they revealed that the first one pays a right premium according to his situation, instead the second one should pay a little less than he does.

Insurance holder	**Pointwise prediction**	**Lower bound**	**Upper bound**	**Real charges**
NS, O, WC	7178 $	6327 $	8028 $	7537 $
S, O, C	38505 $	37121 $	39890 $	42213 $

Table 1: Smoke impact in health insurance computations.
Legend: NS=not smoker, S=smoker, O=obese, C=with children, WC=without children

Then we compared a normal weight insurance holder with one that has a high bmi and, as can be seen in Table 2, we obtained different pointwise predictions. However, both these values are not in their confidence intervals: the first is just outside its range, instead the second is significantly far away from the predicted value. More precisely the normal weight insurance holder is paying a little less than he should, the other is paying a lot more.

Insurance holder	**Pointwise prediction**	**Lower bound**	**Upper bound**	**Real charges**
NW, NS, WC	5668 $	4848 $	6487 $	4772 $
O, NS, WC	11006 $	10116 $	11897 $	15602 $

Table 2: Comparison between high and normal bmi in terms of health premium.
Legend: NS=not smoker, O=obese, NW=normal weight, WC=without children

Moreover, as concrete examples, we wanted to investigate how much some american celebrities should pay. In particular we decided to report the case of Lizzo, a singer with a high bmi, Beyoncé and George Clooney, acquiring their information from a website [4]:

Celebrity	Age	Bmi	Steps	Smoker	Coverage	Premium
Lizzo	32	38.80	4000	No	Single	5400 $
Beyoncé	39	21.40	6000	Yes	Family	21320 $
George Clooney	59	24.00	6000	Yes	Family	25980 $

Table 3: Data about Lizzo, Beyoncé and George Clooney

We got the results reported in Table 4:

Celebrity	Pointwise prediction	Lower bound	Upper bound
Lizzo	5721 $	4928 $	6514 $
Beyoncé	18692 $	16630 $	20753 $
George Clooney	28170 $	26488 $	29854 $

Table 4: Prediction for some american celebrities

5 Implementation of the software application

5.1 Web app objective

The final product of our research is "US Health Insurance Calculator" web page [5], which offers the possibility for the individual user to check the fairness of his paid premium or price quotation of his offered health insurance plan. This application aims to be performant in many aspects: fast in loading, precise in assessing prices and attractive to the user. The analysis reported in the previous chapters allows us to ask the user all necessary information on which the prediction model is based, evaluate the predicted fair premium and make concrete suggestions to reduce health insurance costs.

5.2 Why choosing Shiny?

Shiny is a R package used to build interactive apps straight from R. One of the main reasons it has been chosen to implement our web app is because Shiny has provided us an easy to learn and time saving alternative to JavaScript, thanks to an extensive documentation of the package and a long list of web app examples on its site [6]. Moreover, compared to other web site builders, this choice has allowed us to perform real-time predictions not at expense of an attractive user interface.

5.3 How it works, a brief scheme

In order to clarify better the web app functioning, here is reported a brief summary of an example session:

1. When the app is opened a green value box, which directs the user to open the sidebar, an empty table and a disclaimer are displayed.

2. User opens sidebar, inserts personal information and submits it by clicking "Check your premium" button.

3. After a limited loading time, on the dashboard page the following value boxes are displayed:

   - First one shows user's submitted information on how much he is charged by his insurance.
   - Second one shows predicted charges obtained by our prediction model. The value box colour is based on the comparison between user's actual charges and lower and upper bounds of the confidence interval evaluated.
   - Third one shows a comment about the result obtained.
   - Conditionally on the fact whether the user smokes or not, a purple value box is displayed, suggesting not smoking to reduce user's insurance premium.
   - Conditionally on the body mass index category, a light blue value box is displayed, indicating a healthy weight compared to one's height reduces charges premium.

4. Meanwhile a table called "Your info submitted" shows input values for the prediction model.

5. If user wants to insert a completely new set of information, opens sidebar and clicks "Clear" button. Then all previous results are not displayed anymore and the page becomes as it was in 1. stage.

5.4 Actual implementation

A Shiny web app is built by running a shinyApp() function. Its arguments are ui (user interface) and a server function. The first one is nothing but the dashboard page and describes the input/output content to be displayed. The second one defines the server-side logic of the Shiny application, which generally involves creating functions that map user inputs to various kinds of output.
Our dashboard page has three parts: a header, a sidebar and a body. In particular, in the sidebar selectizeInput() and sliderInput() collect dichotomous and continuous user information respectively. At the end of the sidebar there are "Check your premium" and "Clear" buttons. Whereas, in the dashboard body, content is displayed in a row styled layout and in each fluidRow() a value box, a table and a text output are defined. Two things are worth to mention: at the beginning the style of the disclaimer and the desktop view are set; uiOutput() defines a reactive object that indicates a specific requirement, which, if it is met, another ui object is created (in our case the additional suggestion value boxes).
Inside the server function, firstly two reactive datasets are initialized, in order to store input data and prediction results respectively. Then is described what happens when the "Clear" button is clicked: the last row of both datasets is deleted. Afterwards is defined the reactive event of the "Check your premium" button and the one of a data table: input content is stored inside one of the datasets and is

displayed in the data table as it was stored. Then, thanks to renderText() function, the disclaimer is stated.
The remaining part of the server function is dedicated to value boxes. The first one shows how much the user is charged, if information is inserted, or an indication on submitting it if not. The second and the third one, instead, are just two green boxes when user has not inserted input data yet.
Inside the second box reactive event, firstly the GAM is created, based on our given dataset, stored in insurance.rda and loaded at the beginning. Then predictions are obtained and stored inside the second reactive dataset. Afterwards the actual value box is built: it shows the predicted charges amount and its colour describes whether user's charges are good or not compared to the predicted ones. In particular, the colour criteria are:

- Red if user's charges are outside of the confidence interval;
- Green if user's charges are inside the confidence interval and lower than predicted ones;
- Orange if user's charges are inside the confidence interval and greater than predicted ones.

The third value box displays a comment on the result obtained and a suggestion on checking the insurance plan coverage.
Then are stated the conditions for the display of additional value boxes. The first one appears only if the user is a smoker and informs him how much his premium would be if he quits this habit. The other value boxes display occurs when the user's body mass index category is overweight (between 24.9 kg/m^2 and 30.0 kg/m^2) or obese (over 30.0 kg/m^2). In both cases a prediction of a normal body mass index (21.6 kg/m^2) is showed, suggesting reducing theirs.
At last, the shinyApp() receives ui and the server functions.

5.5 Four adjectives to describe our web app

- **User friendly**
  Some little details such as the colour choice based on the prediction interval, the initial instructions for the user which direct him to the sidebar and icons [7] based on what is suggested make the web app effectively intuitive. Moreover, since slow loading frustrates site visitors [8], its fast speed is a key aspect we have considered when selecting the prediction method.

- **Clear**
  The comparison of user's and predicted charges is the first information showed in the main dashboard and an additional comment confirms the result. Moreover, the data table shows the user whether he has inserted his information correctly or not.

- **Useful**
  The user's interest on knowing his fair insurance premium is moved by the desire of avoiding wasteful expenses, which are greatly present in the health care system [9]. Our web application seems a good solution for this.

- **Non-judgemental**
  The suggestion value box colours (purple for smoker and light blue for a particular bmi category) do not imply any critic message on the user. Red and orange colours are chosen only for value boxes related on charges. Moreover, the disclaimer advises to talk with a professional since the web app is only for information purposes.

6 Conclusion

Our research shows that the US health insurance premium is impacted by the holder's physical characteristics, such as body mass index and age, lifestyle habits, like smoking and sedentary behaviour, and coverage type. Beneficiary's gender and location do not affect its price.
These results are used to develop an efficient service for the individual US citizen, which communicates his right premium according to his personal information.

Bibliography

[1] Stanley Eakins Giancarlo Forestieri Frederic Mishkin. *Istituzioni e mercati*. Pearson.

[2] https://www.kaggle.com/easonlai/sample-insurance-claim-prediction dataset.

[3] https://www.bcbsm.com/index/health-insurance-help/faqs/topics/buying-insurance/family-size-impact cost.html.

[4] https://celebrityinside.com.

[5] https://valentinabuccioni.shinyapps.io/USHealthInsuranceCalculator/.

[6] https://shiny.rstudio.com/.

[7] https://fontawesome.com/icons?d=gallery.

[8] https://www.ericsson.com/assets/local/mobility-report/documents/2016/ericsson-mobility-report-feb 2016-interim.pdf.

[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2690367/.