National Research University Higher School of Economics
Data Science and Business Analytics

Comprehensive Medical Cost

Elizaveta Anikina
Group 212

Table of contents

The Problem and Its Significance

Healthcare expenditure represents a significant portion of personal and national budgets worldwide. Understanding the factors that drive these costs is crucial for individuals, insurers, and policymakers. This research focuses on identifying the key variables that influence individual medical expenses. The primary variable of interest is the cost individuals incur for medical care. The goal is to analyze how various demographic and health-related factors impact these costs, enabling more accurate predictions and informed decision-making in healthcare policy and personal health insurance planning.

The dataset that we have chosen offers a wide collection of individual-level data, including variables such as age, sex, BMI, number of children, smoking status, region, and individual medical costs. The diversity of the dataset provides a rich basis for exploring the complexities of healthcare expenditure, allowing us to uncover patterns and relationships that might not be evident in more aggregated data, offering insights into how different characteristics correlate with medical costs.

Scientific literature

"An Empirical Study on the Determinants of Health Care Expenses in Emerging Economies" published in BMC Health Services Research. This study explored the impact of factors like economic growth, aging population, industrialization, agricultural activities, and technological advancement on healthcare costs in emerging countries. It utilized datasets from 22 emerging countries and employed methods like Quantile regression and Pooled Mean Group causality tests. The study found a complex interplay of these factors on healthcare costs, with some factors like industrialization and the aging population having a varied impact at different levels of healthcare spending.

"Factors Affecting Hospital Costs and Revenue: Integrating Expert Opinions and Literature Review" from Emerald Insight. This study aimed to identify and classify factors affecting hospital costs and revenue (HCR) by integrating experts' opinions and a literature review. It identified 22 new factors as determinants of HCR and categorized them into seven main groups and 22 subgroups. This study provides a comprehensive overview of factors affecting HCR, useful for hospital budgeting and financial planning.

Source Description

The Medical Cost Dataset is an extensive collection of data specifically designed to analyze factors affecting individual healthcare costs. [1]

Key Features of the Dataset

- Variables Included: The dataset comprises several key variables instrumental in studying healthcare costs. These include age, sex, Body Mass Index (BMI), number of children, smoking status, region, and individual medical costs. These variables provide a holistic view of factors influencing healthcare expenses.
- Data Type and Structure: The data is structured in a tabular format, which facilitates ease of analysis and manipulation using various statistical tools and software.
- Sample Size: The dataset contains a substantial number of records, providing a robust sample for statistical analysis and ensuring the reliability of the research findings.

Relevance to the Study

- Comprehensiveness: The dataset's comprehensive nature, encompassing a range of demographic and health-related characteristics, is ideal for an in-depth exploration of the factors affecting healthcare costs.
- Diversity: It covers a diverse population, making the findings more generalizable and applicable to a broader context.
- Applicability: The variables included in the dataset are directly relevant to the research questions posed in this study, making it an appropriate choice for this research.

Data Source Reliability

- Credibility: Hosted on Kaggle, a reputable platform for data science and analytics, the dataset is considered reliable and widely used in academic and research settings.
- Data Integrity: The dataset is well-maintained and regularly updated, ensuring data quality and relevance.

Data Description

The Medical Cost Dataset is a rich resource for analyzing healthcare costs on an individual level. Core Components of the Dataset:

- Age: This variable represents the age of the individuals in the dataset, providing insight into how age may correlate with healthcare costs.
- Sex: The dataset includes gender information, allowing for analysis of differences in healthcare costs between males and females.
- BMI (Body Mass Index): BMI is included as a key health indicator, offering a perspective on how weight-related factors may influence medical expenses.
- Children: The number of children/dependents covered by the health insurance plan is recorded, pertinent for understanding the impact of family size on healthcare costs.
- Smoker Status: This variable indicates whether the individual is a smoker, a significant factor in health risk, and potentially in medical expenses.
- Region: Geographic region is noted, providing an opportunity to explore regional variations in healthcare costs.
- Charges: The primary variable of interest, represents the individual medical costs billed by health insurance.

Data Characteristics:

- Format and Structure: The dataset is structured in a tabular format, suitable for analysis with standard statistical and data analysis tools.
- Sample Size and Diversity: It encompasses a diverse sample of individuals, covering a range of ages, sexes, BMI levels, and other factors, making the dataset representative of a broader population.
- Data Quality: Given its availability on a reputable platform like Kaggle, the dataset is assumed to be of high quality, with reliable and valid data entries.

Applicability for the Study:

- Relevance: The variables included directly relate to the research objectives of identifying the factors influencing healthcare costs.
- Analytical Potential: The dataset's structure and content are conducive to various statistical analyses, including regression modeling, to identify and quantify the relationships between these variables and healthcare costs.

Exploratory Data Analysis

To evaluate the dataset, we conducted an Exploratory Data Analysis (EDA). To begin with, we considered it necessary to find the minimum and maximum values of all necessary

variables using the basic pandas functions. Further we decided to find outliers - to solve this problem, we built a density plot based on charges, and then used 3 sigma analysis. The results of the analysis showed that it is necessary to trim the sample by about 50,000 - these results coincided with the visual analysis, so we decided to trim the sample by the value of 50,000 charges.

Next, we introduced dummy variables in those columns where it was required, and got the original dataset.

Firstly, we built a correlation heatmap to check if the selected parameters correlate with each other. We got low correlation values, which means that all variables can be left.

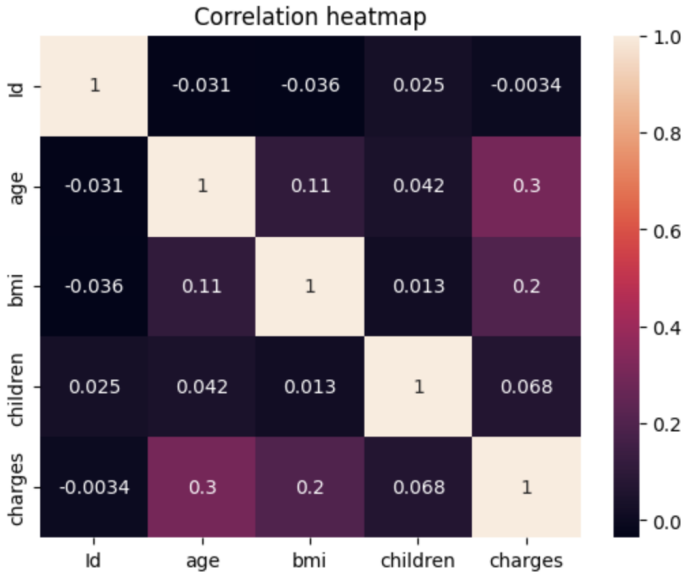


Figure 1

We decided to check the effect of age on the amount of medical expenses. We built a scatterplot, and we saw - firstly, in general, spending on medical care increases with age, but secondly, we noticed an interesting feature that society is divided into three conditional layers, which, regardless of age, spend more or less. The lower layer is formed by people whose expenses are the least, the average is the average, the upper layer is the largest expenses. We see that the main part is concentrated in the lower layer. We will continue to explore this feature in the following sections.

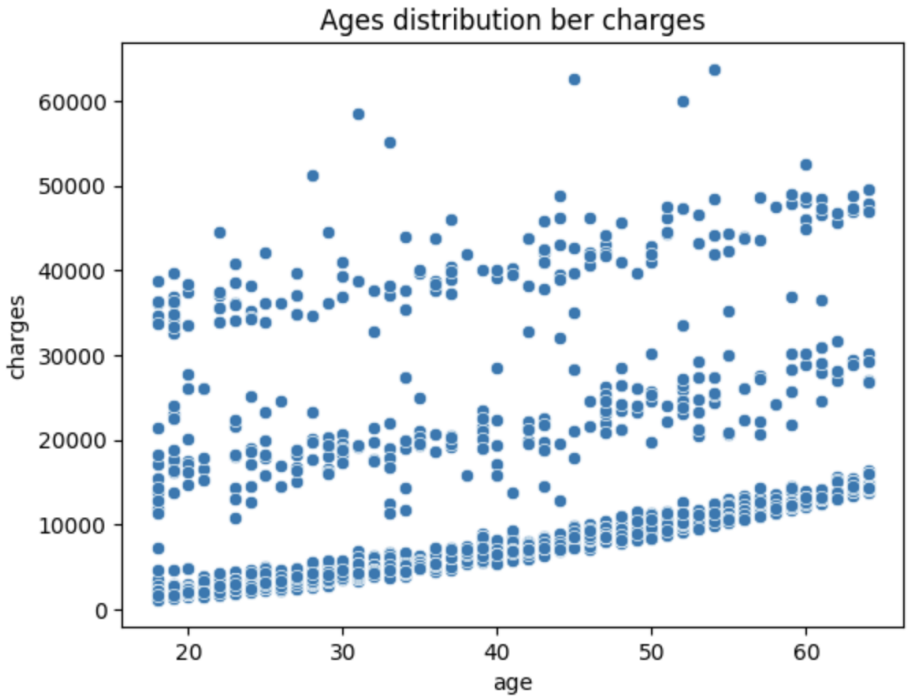


Figure 2

We assumed that the more children a woman has, the more her medical insurance charges should be. However, having plotted the graph, we found such a pattern only after 3 children - from 0 to 3 children - men spend more than women, where 4 children spend the same, where 5 - women spend more. We decided to check this hypothesis in the following discussion.

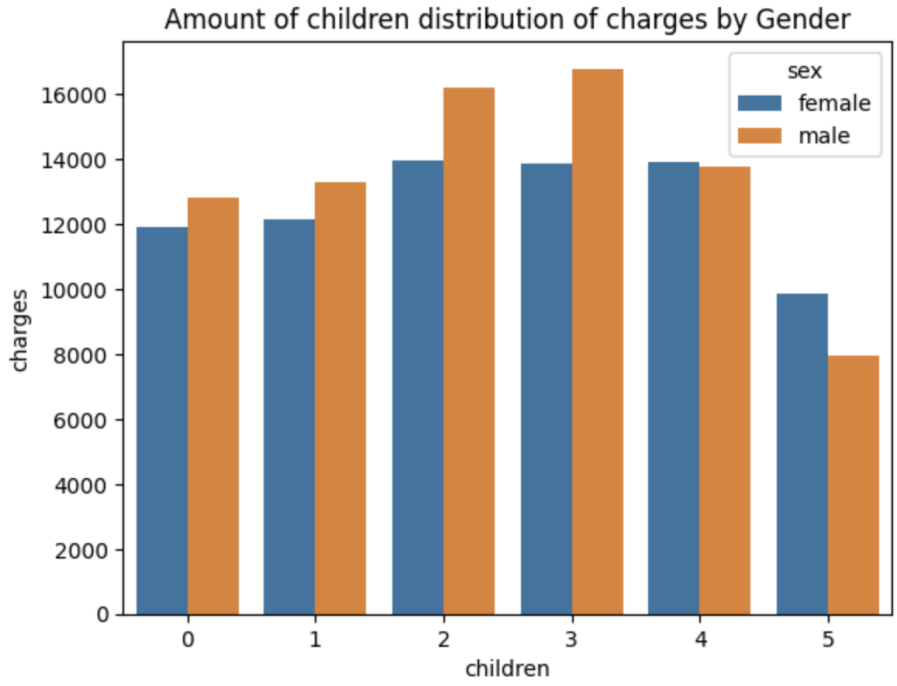


Figure 3

Next, we identified the following dependencies:

- Smokers have a significantly higher median medical insurance costs than non-smokers. We found out this by simply calculating the values and building a visual boxplot. It can be concluded that smokers spend more than non-smokers.

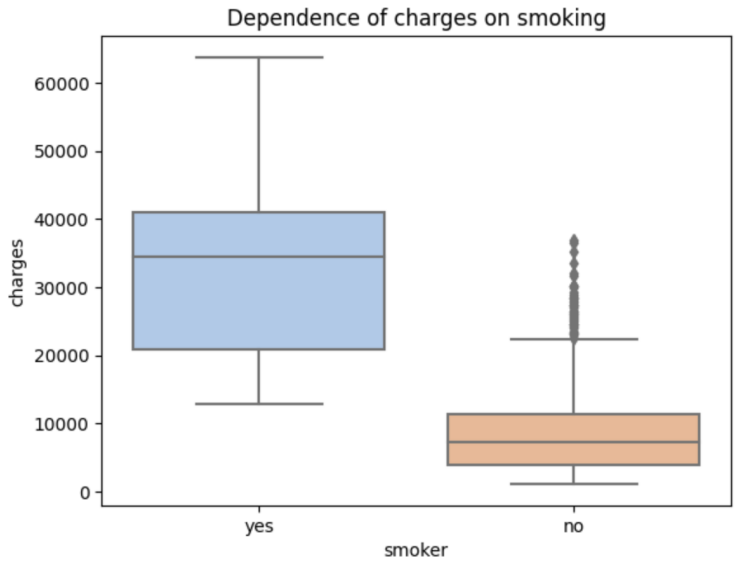


Figure 4

- Women have a greater spread in expenses than men, although the median is about the same level - boxplot.

- In the regions, health insurance costs are practically the same, so the difference in region variables is practically insignificant.

Our assumptions regarding future results

Now let's move on to our assumptions. Based on the already conducted EDA, the following assumptions can be made about the coefficients of our future model.

1. The coefficient of the age variable will be positive - the older the age, the more medical insurance charges will be

2. It can be assumed that the coefficient of the bmi variable will be positive. However effect of this coefficient may be less significant, since the distribution of bmi relative to charges can take the form of a parabola: people with low and high bmi have more charges than those with normal bmi.

3. As for children, according to the repeated graph(fig 3), it can be seen that the distribution has the shape of an inverted parabola, so we assume that the coefficient is likely to be negative

4. The variable sex_male will presumably have a negative coefficient, since men have statistically better health, so they are less likely to seek medical help then women.

5. Smoker_yes has a positive coefficients since smokers have statistically more health problems, therefore they spend more on health insurance than non-smokers(fig 4)

6. As for the regions, we assume that the regions are not significant at all in this model

Model Description

At the beginning we introduced this baseline model (linear regression using Ordinary Least Squares with the assumption of heteroscedasticity and robust standard errors (HC3) as a result) , which in future will be tuned:

$$\text{charges} = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{bmi}) + \beta_3(\text{children}) + \beta_4(\text{sex\_male}) + \beta_5(\text{smoker\_yes}) + \epsilon$$

In order to ensure that all the variables have significant effect and really explains the target variable, we conducted several F-tests for testing whether unrestricted models have better performance compared to restricted ones.

For region

We reject the null hypothesis,
F statistics (0.804) and p-value (0.492) with F critical (11, 980) = 1.75 at 5% sign. level
so region has no effect on our model, as we assumed earlier, that is why we can remove this variable from our model.

For sex

We failed to accept the null hypothesis,
F statistics (8.539) and p-value (0.004) with F critical (11, 977) = 1.75 at 5% sign. level
so sex has no significant effect on our model, but despite these results we decided to leave this variable, as in combination with the number of children for women it may lead to high significance.

For children

We accept the null hypothesis,
F statistics (0.843) and p-value (0.47) with F critical (11, 980) = 1.75 at 5% sign. level
so the number of children has an effect on our model according to the F-test, but we assume bias due to absence of division by sex (we are convinced that for women this variable has a crucial effect, while for men it's really pretty useful and doesn't predict medical charges. So the existence of 'men' sex (in roughly equal proportion) may lead to biased results.

Omitted variables

For checking the correctness of chosen functional form and presence of omitted variables the Ramsey test was performed, so we figured out that our model requires inclusion of higher-order terms, as we rejected the H0.

Ramsey RESET Test p-value 6.834922985587154e-18
Ramsey RESET Test F-statistic 79.04897294286812

To determine for which variables we may add polynomial terms we should carefully inspect our data.
Taking into consideration previous EDA and constructed scatter plot, we may conclude that the 'bmi' variable has a non-linear relationship with the target variable, that is why adding its polynomial term will be appropriate in this case.

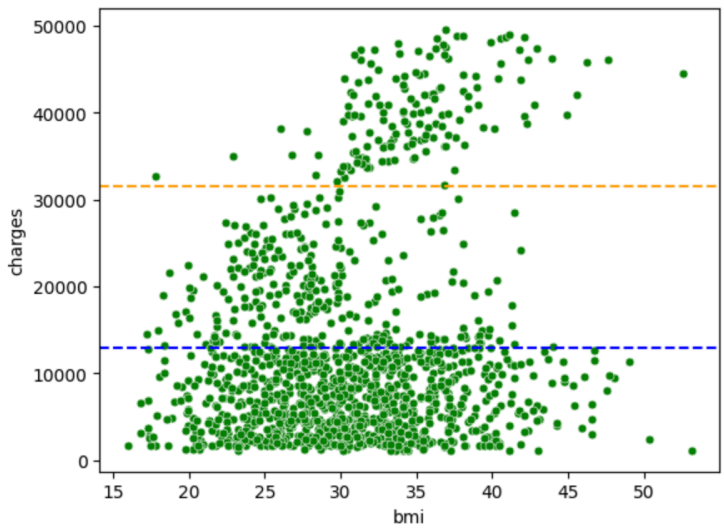
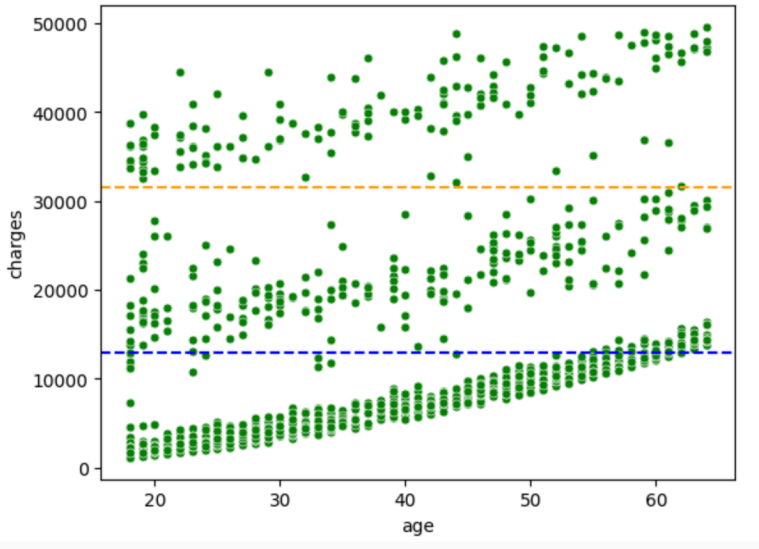


Figure 5

Moreover we constructed a hypothesis that we have three groups in our dataset which may be divided into different income groups. It is pretty visualized on a scatter plot of age (there is

approximately linear relation between age and charges but the respondents are divided into three layers where charges linearly change with age , but on the different levels). So we decided to try new models, where each will inspect different social groups.
Models were constructed for the following groups : charges <= 13000; 13000 < charges <= 31500 ; charges > 31500.



It is important to mention that in improvements we suggests to add parameter for indicating three levels of income for future using one of this models instead of merged one , as it provides much better performance

We have tried three separate models and accuracy scores were pretty high (R^2 equal to 0.901, 0.944, 0.915 respectively).

Heteroscedasticity

For checking the model for presence of heteroscedasticity we conducted White test.
Based on the results we reject Ho in favor of Ha,
LM Statistic: 111.79494605742155
LM-Test p-value: 1.3168994382549265e-07
F-statistics: 2.6687775945603702
F-Stat p-value: 3.933252016248433e-08
so it indicates the presence of heteroscedasticity. To deal with this issue, as models may have issues, such as inefficient estimation, biased standard errors, compromised test statistics, and prediction problems the most common solutions may be implemented: logarithmic/semi-logarithmic functional form or weighted regression, which we have tried.
However, despite these efforts, we have encountered significant challenges. The modifications we implemented did not lead to substantial improvements in the p-value and F-

statistics as per the results from the White test. This outcome suggests that the heteroscedasticity in our model is of a more complex nature and may not be readily resolved through conventional methods, but due to the usage of robust standard errors we may leave all as it was.

Interpretation of economic model

$$\text{charges} = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{bmi}) + \beta_3(\text{children}) + \beta_4(\text{sex\_male}) + \beta_5(\text{smoker\_yes}) + \beta_6(\text{bmi\_squared}) + \epsilon$$

Based on the hypotheses outlined in our project, we can address the questions as follows:

Our regression shows that

1. The 'age' coefficient is positive, which aligns with our hypothesis.
2. 'bmi' has a positive coefficient, and 'bmi_squared' has a negative coefficient, suggesting the parabolic relationship we hypothesized.
3. The coefficient for 'children' is positive, which does not coincide with our hypothesis.
4. The coefficient for 'sex_male' is negative and not statistically significant, indicating a possible deviation from our hypothesis.
5. The 'smoker_yes' variable has a positive coefficient, which is in line with our hypothesis.
6. The regression output does not include 'regional' effects, so we cannot assess that hypothesis.

Economic interpretation of the estimated coefficients:

Age: A one-year increase in age is associated with a $260 increase in charges, economically representing the higher healthcare costs expected as individuals age.

BMI: A one-unit increase in BMI is associated with a $625 increase in charges, reflecting the economic costs associated with higher BMI, possibly due to associated health risks.

Children: Each additional child is associated with a $455 increase in charges. Economically, this could reflect the incremental costs of healthcare coverage for larger families.

Sex_male: The negative coefficient (though not statistically significant) suggests no substantial economic difference in charges between the sexes, under the model.

Smoker_yes: Being a smoker is associated with an approximately $23,750 increase in charges, a significant economic burden reflecting the high costs associated with smoking-related health risks.

Conclusion

The majority of our initial assumptions coincide with the obtained results and the model provides an adequate accuracy score (R^2 = 0.754), but as was mentioned for higher accuracy (approximately R^2 = 0.9) more variables in the initial dataset should be added.

Discussion (Drawbacks and Possible improvements)

1. One limitation of our model may entail the presence of potential bias, as the number of children is unlikely to significantly impact the target value for males, while it can have a substantial influence for females. Despite this, we are unable to exclude either gender or the variable representing the number of children from the model.

OLS Regression Results

Dep. Variable:	charges	**R-squared:**	0.754
Model:	OLS	**Adj. R-squared:**	0.752
Method:	Least Squares	**F-statistic:**	316.2
Date:	Sat, 09 Dec 2023	**Prob (F-statistic):**	3.60e-226
Time:	20:12:16	**Log-Likelihood:**	-10073.
No. Observations:	998	**AIC:**	2.016e+04
Df Residuals:	991	**BIC:**	2.020e+04
Df Model:	6		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.592e+04	4225.293	-3.768	0.000	-2.42e+04	-7640.410
age	260.3672	13.479	19.317	0.000	233.950	286.785
bmi	625.7676	275.628	2.270	0.023	85.547	1165.988
children	455.0039	153.308	2.968	0.003	154.526	755.482
sex_male	-82.3830	374.110	-0.220	0.826	-815.625	650.859
smoker_yes	2.375e+04	653.372	36.353	0.000	2.25e+04	2.5e+04
bmi_squared	-5.5998	4.435	-1.263	0.207	-14.293	3.093

Omnibus:	203.324	**Durbin-Watson:**	2.005
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	422.364
Skew:	1.151	**Prob(JB):**	1.93e-92
Kurtosis:	5.204	**Cond. No.**	2.10e+04

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
[2] The condition number is large, 2.1e+04. This might indicate that there are strong multicollinearity or other numerical problems.

2. The second possible improvement is to add variables responsible for social status and approximate income in order to choose one of the three models trained on the respectful data for higher accuracy.