Examining the Determining Factors Behind Medical Insurance Costs

Minhyuk Choi, Nikhilesh Kumar, Andrew Kanovsky, Richard Osikowicz

December 9, 2021

PART 1: INTRODUCTION

Medical costs are an essential expenditure across the population. By investigating the relationship between a number of characteristics and medical costs, we hope to illustrate how medical costs can vary for different groups of people. This information is vital to understanding how people consume healthcare and what ways seemingly unrelated factors can affect healthcare costs both on an individual and societal level. The paper contains analysis of the issue in the following order. First, it will examine the related studies and research on the relationship between medical costs and other several personal factors such as gender, diabetes, and smoking. These studies will provide the general idea and methodology of this analysis. Then, we will briefly explain the model we used in this analysis. In this part, we explore whether the data of explanatory variables that will be used in the regression model is valid by looking at the Gauss-Markov's Multiple Linear Regression (MLR) Assumptions. Lastly, we will provide the source of the data and summarized statistics for the audience, and the interpretation of the results will be presented by examining the statistical attributes such as parameter estimates, F- statistics, number of observations, and more.

PART 2: LITERATURE REVIEW

To begin, we examined existing literature to see if prior evidence existed concerning the relationship between medical costs and various factors. Kim et. al. find a positive relationship between medical costs and age. However, the data also shows that these trends vary by sex: medical costs increase for men were higher than those for women of the same age[1]. The base medical expenditures, however, between young men and young women, were higher among young women. Thus, the likely reason for a slower rate of increase among aging women over aging men is a higher initial cost. Based on this paper, we should observe a positive correlation between age and medical costs as well as a positive correlation between womanhood and medical costs, which will be observed in the dummy variable of men vs. women.

Ramsey et al. found that, among beneficiaries of employer health plans, costs were higher for employees with diabetes compared to employees without diabetes, all else held equal [2]. The incremental cost associated with diabetes was about $4,410. Because diabetes is closely related to high BMI, we should observe a similar relationship in our data between BMI and cost -- higher BMI should be associated with higher costs.

Barendregt et al. find a positive relationship between smoking and medical costs at a given age but find a positive relationship between medical costs for the entire population and the cessation of smoking (as nonsmokers live longer, and thus consume more healthcare when older) [3]. In our data, this should present as a positive correlation between smoking and costs, as we are not examining costs at a population-wide level but rather on a case-by-case basis.

PART 3: THEORETICAL ANALYSIS

Our regression analysis examines healthcare costs with the independent variables. The original econometric model we planned to use is as follows:

$$\boldsymbol{charges = \beta_0 + \beta_1 age + \beta_2 sex + \beta_3 bmi + \beta_4 children + \beta_5 smoker + \beta_6 southwest + \beta_7 southeast + \beta_8 northwest + \beta_9 smokerbmi + \beta_{10} bmi2 + u}$$

In order to determine the compounding effect of smoking and BMI on medical cost, we generated an interaction term: ***smokerbmi*** (smoker* bmi). Since both smoking and an increased BMI cause numerous health issues in conjunction, such as increased incidences of heart attacks, we believe this interaction term will cause a better fit in our model to explain health care costs. We also generated a quadratic term, ***bmi2*** (bmi^2), which is because with BMI, a distinct trend emerges at the ends of the spectrum. Those with an extremely low or extremely high BMI will often have more health problems associated with being at either end of the spectrum. Therefore, the effect of adding an additional point of BMI on health issues will exhibit diminishing effects until a turning point and then exhibit increasing effects, in a quadratic fashion. We expect this term to be positive to reflect the U-shape this effect has. We also generated ***lcharges,*** which is log(charges), because the values we are using are very high and small variations are hard to see without using a log-linear relationship.

Using this dataset, we will regress ***charges*** on ***age***, ***sex***, ***bmi***, ***children***, ***smoker***, ***southwest***, ***southeast***, ***northwest***, ***smokerbmi***, and ***bmi2***. The explanation of each variable is in Figure 1 below.

Figure 1. [4]

VARIABLE	LABEL
charges	Individual medical costs billed by medical insurance.
age	Age of primary beneficiary.
sex	Insurance contractor gender. =1 if the contractor is a male, =0 if the contractor is a female.
bmi	Body Mass Index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight, ideally 18.5 to 24.9.
children	Number of children covered by health insurance/Number of dependents.
smoker	Smoking. =1 if the contractor is a smoker, =0 if the contractor is a non-smoker.
region	The beneficiary's residential area. Divided to four regions: Southwest, Southeast, Northwest, and Northeast. Northeast is not included in order to serve a role as the control group.
southwest	=1 if the beneficiary's residential area is in Southwest, =0 if not.
southeast	=1 if the beneficiary's residential area is in Southeast, =0 if not.
northwest	=1 if the beneficiary's residential area is in Northwest, =0 if not.
smokerbmi	=smoker * bmi
bmi2	$=bmi^2$

There are five total assumptions for multiple linear regression models: the Gauss-Markov Assumptions. Before we run the regression model, checking that the assumptions are valid is essential.

MLR.1 from the Gauss-Markov assumptions states that the parameters to be used in the multiple linear regression should be linear. Our model is in form of $Y = \beta_0+\beta_1X_1+\beta_2X_2 + \ldots + \beta_kX_k + u$. Since the model could be written in form of $Y = \beta_0+\beta_1X_1+\beta_2X_2 + \ldots + \beta_kX_k + u$, it

satisfies the MLR.1. While there is one quadratic term, ***bmi2***, the model is still linear in parameters. MLR.2 states that the samples for the model should be randomly sampled from the population. The dataset we used for this regression included data that was randomly sampled from four different regions of the country and thus satisfies this assumption. The third assumption, or MLR.3, denotes that there should be no perfect collinearity among independent variables from the sample. This could be tested by generating the VIF values for each variable, shown below.

STATA Input: *vif*

Figure 2.

Variable	VIF
bmi2	59.35
bmi	59.32
smokerbmi	25.54
smoker	25.21
southeast	1.65
southwest	1.53
northwest	1.53
age	1.02
sex	1.01
children	1.00

The VIF values from Figure 2 show the degree of correlation between a variable and any other variables in the model. The VIF values mean that if the value is 1, then it means there is no correlation between a variable and any other variables. If the value lies between 1 and 5, then there would be mild correlation (which can be neglected) between the variable and any other variables. If it is higher than 5, it means there will likely be a correlation between the variables, and we need to drop the variable in order to proceed with the model we are currently using. The VIF values of the variables in the model vary widely, with some close to one and others as high as 59.35. The first four variables, specifically, have very high VIF values. However, this is expected, as the first four variables are interaction and quadratic terms, and they are based on

other terms in the regression. Therefore, the high VIF values are not out of the ordinary, and the model overall satisfies the MLR.3 assumption.

MLR.4 indicates that the error term, ***u***, has zero conditional mean for all given independent variables. To figure out whether we satisfy the assumption, we checked correlation between the error term and all independent variables.

STATA Input: *corr u age sex bmi children smoker southwest southeast northwest smokerbmi bmi2*

Figure 3.

	u
u	1.000
age	0.0000
sex	-0.0000
bmi	-0.0000
children	0.0000
smoker	0.0000
southwest	-0.0000
southeast	0.0000
northwest	-0.0000
smokerbmi	0.0000
bmi2	-0.0000

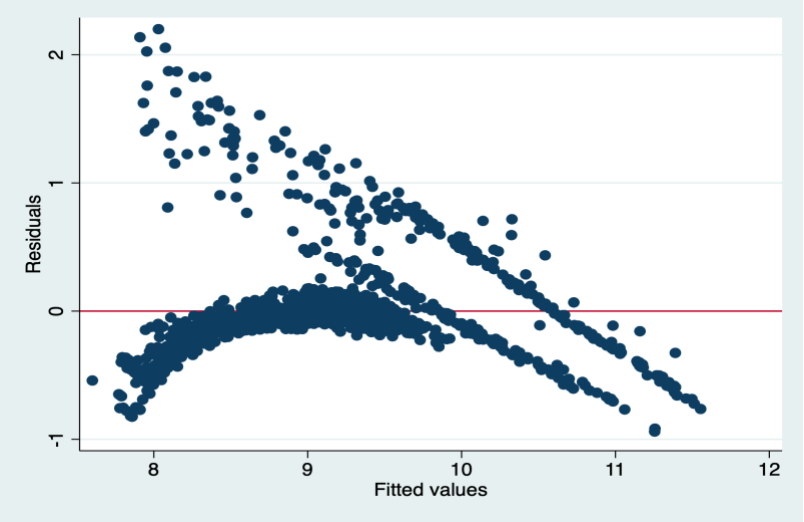
As shown in Figure 3, all of the correlation between the error term and other independent variables are zero. This means that the error term has a zero conditional mean for all given independent variables. Thus, MLR.4 is also satisfied.

MLR.5 tells us that the conditional variance of the error term, ***u***, is constant no matter what values the explanatory variables included in the model have. If MLR.5 is satisfied, the

model exhibits homoskedasticity. To check for homoskedasticity, we generated a scatter plot of residuals vs. fitted values (Figure 4). From this plot, we could immediately tell that the data exhibited heteroskedasticity. In order to prove this statistically, we ran a Breush-Pagan test on STATA at a significance level of 0.01. As the results are shown in Figure 5, the p-value for the Chi-Squared statistic in this test was 0.00; thus, we have sufficient evidence to reject the null hypothesis and conclude that there exists heteroskedasticity.

As we have recognized that heteroskedasticity exists in the data, we will use robust standard error in order to obtain unbiased standard errors of OLS coefficients under heteroskedasticity. Furthermore, we take a log for dependent variable ***charges*** and name it ***lcharges*** in order to remove heteroskedasticity.

Figure 4.



STATA Input: *estat hettest*

Figure 5.

Breusch–Pagan/Cook–Weisberg test for heteroskedasticity	
Assumption	Normal error terms
Variable	Fitted values of ***charges***
Null Hypothesis	Constant variance
Chi-Squared Test Statistic	82.70
p-value	0.0000

PART 4: EMPIRICAL ANALYSIS AND RESULTS

insurance.csv is the dataset used in this analysis [4]. The dataset is from the book called "Machine Learning with R" by Brett Lantz [5]. This dataset originally contains 9 columns and 1388 rows. For the qualitative variables such as ***sex***, ***smoker***, and regions (***southwest***, ***southeast***, and ***northwest***), we modified the data into the numeric values where they were originally in text. Due to the large size of the dataset, the summary of the dataset is in Figure 6, and the regression results are in Figure 7.

STATA Input: *summarize*

Figure 6.

Variable	Obs	Mean	Std. dev.	Min	Max
age	1,338	39.20703	14.04996	18	64
sex	1,338	.5052317	.5001596	0	1
bmi	1,338	30.6634	6.098187	15.96	53.13
children	1,338	1.094918	1.205493	0	5
smoker	1,338	.2047833	.403694	0	1
southwest	1,338	.2428999	.4289954	0	1
southeast	1,338	.2720478	.4451808	0	1
northwest	1,338	.2428999	.4289954	0	1
lcharges	1,338	13270.42	12110.01	1121.874	63770.43

STATA Input:

gen lcharges = log(charges)

*gen smokerbmi = smoker * bmi*

gen bmi2 = bmi^2

reg lcharges age sex bmi children smoker southwest southeast northwest smokerbmi bmi2

Figure 7.

age	.0346578*** (.008416)
sex	-.087267*** (.023537)
bmi	.0471304*** (.014783)
children	.1033115*** (.0097304)
smoker	.1495958 (.1455818)
southwest	-.1426135*** (.0337977)
southeast	-.1593416*** (.0338207)
northwest	-.077395** (.0336998)
smokerbmi	.0458431*** (.0046503)
bmi2	-.0006958*** (.0002325)
n	1338
$F(10, 1327)$	484.42
R^2	.7850

*Legend: ***p < .01 **p < .05 *p < .10*

The coefficients of our regression analysis are largely in line with our expectations. The positive coefficient on the ***age***, ***children***, and ***bmi*** terms are all positive, which track with

existing literature: the older and more overweight someone is, the more healthcare he or she will likely have to consume (as age and obesity are linked to poorer health outcomes). Having a large family, as well, likely leads to higher healthcare costs, as the number of hospital visits for the family increases with the number of children one has. The positive coefficient on our ***smokerbmi*** term makes sense as well, as the effect of smoking is magnified with an increasing BMI. Obesity is a comorbidity with smoking, so the cost of healthcare should increase for smokers with increasing BMI. This interaction term renders our smoker term invalid, which was unexpected. However, the effect of smoking on healthcare costs with no comorbidities such as obesity may be too minor to be detected in the analysis.

The terms ***sex*** and ***bmi2*** have negative coefficients. The ***sex*** variable is inline with expectations. The ***sex*** variable =1 for men and =0 for women, so our regression results imply that the status of womanhood is associated in an increase in healthcare costs. This relationship has been demonstrated in the literature previously. Our ***bmi2*** term, on the other hand, is not as expected, as we predicted that this term would be positive. However, after further inspection we realized that this term having a negative coefficient illustrates the diminishing positive correlation between BMI and costs as BMI increases. This makes sense, as severe obesity vs. moderate obesity likely does not change medical costs much when compared to a healthy person vs. a moderately obese person.

Additionally, all of the geographical region variables have negative coefficients. This makes sense because the control variable is the Northeast region. The Northeast region has the highest healthcare costs in the country, on average, so it follows that someone from outside of this region would pay less in health care costs than someone from the Northeast.

STATA Input:

reg lcharges age sex bmi children smoker southwest southeast northwest smokerbmi bmi2, vce(robust)

Figure 8.

*Legend: ***p < .01 **p < .05 *p < .10*

age	.0346578*** (.0009779)
sex	-.087267** (.0235782)
bmi	.0471304*** (.0132912)
children	.1033115*** (.008741)
smoker	.1495958 (.1323314)
southwest	-.1426135** (.0331695)
southeast	-.1593416** (.0355896)
northwest	-.077395** (.034345)
smokerbmi	.0458431*** (.0042968)
bmi2	-.0006958*** (.0002038)
n	1338
$F(10, 1327)$	354.19
R^2	.7850

This result with robust standard error shows that all the variables except ***smoker*** are still statistically significant at 5% because p-values for all variables except ***smoker*** are lower than .05. This means we can say with 95% accuracy that the relationship we found between each of our independent variables, excluding ***smoker***, and the log of health care charges are not due to

chance. The R^2 value of 0.7850 (Figure 8) shows that 78.5% of the variance in the log of health care charges is explained by our model. We believe this indicates that our model is sufficient to be used to predict health care costs on average and for individuals because there are many other factors that cannot be directly measured due to HIPPA laws and protection of private medical records. For example, someone who has undergone multiple surgeries, of course, would spend more money on health care costs. However, in the scope of this analysis, a medical history is not a feasible data for us to collect, so we believe that we have achieved the best model possible.

Although there exists heteroskedasticity, it does not mean that the regression itself is biased, and with our inclusion of robust standard errors, our tests are still valid. Therefore, the F-statistic for overall significance shown in the above table is a valid indicator of our regression's overall significance. Our F-statistic is 354.19 (Figure 8), which is far greater than the critical value at 0.01 of 2.40 given 10 numerator degrees of freedom and 1137 denominator degrees of freedom. This means that our independent variables are jointly significant and improve overall fit compared to a model with none of them. Additionally, the t-values for each individual variable's significance are also valid using robust standard errors.

Speaking to the effect of each of the individual variables, we can see that the coefficients we obtained in our regression analysis (Figure 8) can be used to show effects each variable has on medical costs. For ***age***, we observed that each additional year someone has in age, health care costs will increase by 3.47%. For ***sex***, we observed that if a person is male, his health care costs will decrease by 8.73%. For ***children,*** we observed that for each additional child a person has, health care charges will increase by 10.3%. For ***southeast***, ***southwest***, and ***northwest,*** health care costs decrease by 15.9%, 14.3% and 7.73% for people hailing from those regions, respectively. This intuitively makes sense since all of the regions would have lower healthcare costs compared to the northeast, with the southeast having the largest decreasing effect as it is the poorest region in the country.

Since ***smoker*** is included in an interaction term, its interpretation is slightly different from the other independent variables. This interpretation is that if a person is a smoker, then his or her health care costs will increase by 100*(.150 * 0.046(BMI))% at a given BMI. For example, for someone with a BMI of 21, smoking will increase healthcare costs by 116%. For someone with a BMI of 35, smoking will increase healthcare costs by 176%. This, in particular,

illustrates the detrimental comorbid effects that smoking and having a higher BMI have on one's health as healthcare costs increase *dramatically* with these two factors.

Finally, for ***bmi*** since it is both included in an interaction term and a squared term its interpretation is also slightly different from the other independent variables. This interpretation is that at a BMI of 0, each additional BMI point will increase healthcare costs by 4.7%. However, with each additional BMI point, this effect will decrease by 0.06%, so at a BMI of 20, each additional BMI point will increase healthcare costs by 3.5%. Additionally, if one is a smoker, health care will increase by an additional 4.6% on top of the aforementioned effect.

PART 5: CONCLUSION

Our primary goal in these analyses is to determine which factors significantly affect the cost of an individual's healthcare. These factors include biological traits (age, sex) and general health/lifestyle factors (smoking, bmi, location, number of children). Surprisingly, the only variable that we tested that proved to be insignificant at even a 10% level was ***smoker***. Despite being the only factor that is entirely a lifestyle choice and one that has been proven to negatively impact an individual's health, it did not affect healthcare costs by a significant margin. However, the interaction term of ***smoker*bmi*** was significant at the 1% level. This would suggest that insurance companies take a more holistic approach to determining how lifestyle choices affect the pricing of their plans. While smoking may provide some risks that could increase costs, it is when examined in conjunction with other potential health risks (obesity, etc.) that these increases actually come into effect. One possible reason for this observed effect could be that the risk for cardiovascular diseases becomes extremely significant when risk factors are combined (e.g. smoking and obesity) but can be difficult to interpret when only one (smoking) is present.

While our overall model proved to be significant, these analyses show that predicting healthcare costs is often tricky and can depend on numerous factors outside of the basic characteristics we included. For example, prior health complications and dispositions (i.e. family history of cancer, major car accident) were not included in this model and can take any multitude of forms. Because these factors do not take any definite value and often cannot even be converted into indicator variables (say, for example, the severity of a prior accident), they are extremely difficult to represent in an econometric model. For these reasons, it is important that we interpret not just the significance of individual factors but also what the model as a whole can

tell us. And as observed with the example of smoking, the model shows us that insurance companies are primarily focused on overall risk rather than the presence of individual risk factors. While some are certainly significant (bmi and its corresponding risks such as obesity), they do not tell the whole story of how insurance is priced. Additionally, we were able to demonstrate how a number of factors can help create a baseline for how one would expect their insurance costs to change given certain predictable traits (static in the case of gender and location, increasing at a constant rate in the case of age, and increasing by a set amount with each child).

In the modern world, healthcare is an industry at the forefront of society, with cost being a major talking point in each of the political, social, and economic spheres. By understanding what factors certain companies utilize to derive their insurance costs, both consumers and employers can better select healthcare options that suit their needs. But ultimately, not everything can be broken down and quantified, meaning proper research and education remain paramount to obtaining the best coverage possible.

PART 6: WORKS CITED

Works Cited

[1] Barendregt, Jan J., et al. “The Health Care Costs of Smoking: Nejm.” *New England Journal of Medicine,* 12 Feb. 1998, https://www.nejm.org/doi/full/10.1056/NEJM199710093371506.

[2] Kim, Hyun, et al. “Aging, Sex, and Cost of Medical Treatment.” *Journal of Occupational & Environmental Medicine*, vol. 55, no. 5, 2013, pp. 572–578., https://doi.org/10.1097/jom.0b013e318289eeda.

[3] Ramsey, S., et al. “Productivity and Medical Costs of Diabetes in a Large Employer Population.” *Diabetes Care*, vol. 25, no. 1, 2002, pp. 23–29., https://doi.org/10.2337/diacare.25.1.23.

[4] B. Lantz, “Chapter 6, insurance.csv” in Machine learning with R: Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R, Birmingham: *Packt Publishing*, 2015. https://docs.google.com/spreadsheets/d/1PuHJl5qZBJvMyhqKf0aHHljdYKE0WC4aLaQNhq6KI0E/edit?usp=sharing