Modeling Medical Cost Differences Across the US

Kyle Smith & Jon Duea

*Final Submission Date here*

*Stat 311 Regression Analysis, Spring 2025*

## Introduction:

For our topic, we chose the research question: “Do the factors that influence medical costs remain consistent across different regions?”

This question is significant because it highlights the variability in medical costs billed by insurance across the United States, potentially driven by demographic and lifestyle differences. Understanding these variations can reveal whether people face cost disparities based solely on their region, and highlight the influence specific factors have on the costs people face around the U.S.

To answer this question, we use a publicly available dataset that records total annual medical costs billed by insurance (charges) for 1338 simulated beneficiaries (individuals with an insurance policy) together with six predictors (age, sex, BMI, children, smoker, region).

The approach we will take to answer this question is to first conduct exploratory data analysis (EDA) to inform our model building process, build regression models incorporating insights from our EDA, validate our model assumptions before making any conclusions, and then compare competing models for efficacy and to determine whether the effects of the predictors remain consistent across different regions.

## Data Description:

The dataset we will be using for our analysis was obtained from [*Medical Cost Personal Datasets*](https://www.kaggle.com/datasets/mirichoi0218/insurance/data)(Choi, n.d.)available on Kaggle, which was originally curated for use in the book *Machine Learning with R* and was “created for this book using demographic statistics from the U.S. Census Bureau, and thus approximately reflect real-world conditions.” (Lantz, 2013)*.* Although not explicitly stated in the book’s description of the dataset, it is assumed that it was generated based off 2013 demographic statistics from the U.S. Census Bureau.

Included in the dataset is complete information with no values missing for 1338 examples of beneficiaries currently enrolled in an insurance plan and covers the following variables:

* Charges (USD/year)
* Age (integer years, )
* Sex (male/female)
* BMI (Body-Mass Index,
* Children (number of dependents)
* Smoker (yes/no)
* Region (of United States: northeast, northwest, southeast, southwest)

This dataset represents a snapshot of individuals from varying demographic profiles (age, sex), health indicators (BMI, smoking status), family sizes (number of children), and geographic locations within the US. The associated charge for everyone reflects the medical expenses billed to their insurance and provides the basis for exploring how these factors influence medical costs collectively and individually.

Exploratory Data Analysis

A screenshot of a table

AI-generated content may be incorrect.

Snapshot of first five records of the dataset.

Exploration of our datasets revealed many insights for us and highlighted some issues that will need to be accounted for when conducting our analysis. Appendix A contains the detailed results from our analysis, and the accompanying JMP workbook (insurance.jmp) can be found in our [GitHub repository for this project](https://github.com/jmduea/Modeling-Medical-Cost-Differences-Across-the-US) (Duea & Smith, 2025) if you would like to run the analysis for yourself. All figures and tables included in this report and the appendices have an associated script in insurance.jmp of the same name that will reproduce them.

For a descriptive summary of the numeric variables in our data, see appendix A, figure A1. Below we highlight four key insights about our data uncovered during our analysis:

* **Charges**: Analysis of the distribution of charges shows outliers and a rightward skew, applying the log transformation transforms charges to a more normal distribution (see appendix A, figure A2). Additionally, when looking at charges by region and trying to visually estimate whether the assumption of homoscedasticity is satisfied we can see that it is not and we will have to take the necessary steps to address the heteroscedasticity in our modeling efforts (see appendix A, figure A3).
* **BMI**: The average BMI for an individual in our dataset (see appendix A, Figure A1) is 30.66, which classifies them as obese, according to info about adult BMI categories provided by the CDC (Centers for Disease Control and Prevention, 2024). Looking at charges across different BMI values (see appendix A, figure A4) we see the assumption of homoscedasticity violated here as well, with a noticeable increase in variance in charges at the 30 BMI mark, suggesting that individuals that would be classified as obese have increased variance in medical costs compared to those in healthy weight ranges.
* **Smoker:** The difference between charges between smokers and non-smokers is significant, with smokers facing over 3 times the medical costs of a non-smoker on average; furthermore, charges that would be considered outliers for non-smokers fall firmly within the range of normal values for smokers (see Appendix A, figure A5). Considering the average individual in our dataset falls into the obese category and research from the NIH indicating that individuals who are both obese and smoke face worse health outcomes than others (Stewart, Cutler, & Rosen, 2009), we looked to see if that interaction might be part of the reason we saw such a sharp increase in variance of charges for obese individuals (see appendix A, figure A6). Based on this we hypothesize the interaction between BMI and smoking seems to be significant in influencing medical costs, with the interaction being more influential for obese individuals.
* **Region:** Average costs across the regions are similar, with one notable exception being the southeast region (see appendix A, figure A7). Looking at the prevalence of smokers and obese individuals across the different regions provides an explanation for why this might be, and we find that the southeast region has the highest prevalence of smokers (see appendix A, figure A8), and obese individuals (see appendix A, figure A9). With this in mind, we deemed it was important to account for the impacts of smoking and obesity before drawing any conclusions about how much a region influences the effect of our predictors on medical costs.

Based on these insights, we conclude with three actions we should take when preparing our models:

1. Taking the log transformation of charges is necessary to address the skew and unequal variances present in our data, so we will be interpreting our models as multiplicative models.
2. Incorporation of an indicator variable for BMI 30 will help us better model the effects of BMI and its interaction with smoking.
3. Incorporation of interactions between smoking and our BMI terms to account for the major effects they have on each other in influencing medical costs.

See appendix A, figure A10 and figure A11 notes for a detailed breakdown and supporting visualizations for this.

## Methods:

We model medical charges using multiple linear regression, and because raw charges are highly skewed, violate constant-variance and normal-error assumptions, we fit multiplicative models with . Considering the nature of medical costs and the factors that influence them, it intuitively follows that percentages changes in charges are more appropriate than additive changes in charges. The accompanying JMP workbook (insurance.jmp) includes scripts for model fit along with their accompanying diagnostics so that, if desired, you can run the models yourself. We hypothesize three models, progressively adding more terms and comparing them for efficacy.

### Model 1 (main effect terms, BMI x Smoker interaction):

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
| Northeast is the base level for region | | |
|  |  |  |

This model considers all the main effects of our predictors, along with the known interaction effect between BMI and smoking, providing a baseline for us to compare the other models against. The output from JMP for this model can be found in the script group Model 1 of the insurance.jmp workbook as well as in Appendix B.

### Model 2 (extends model 1 with quadratic term for age and smoker interactions):

terms defined the same as model 1 with one added new qualitative variable.

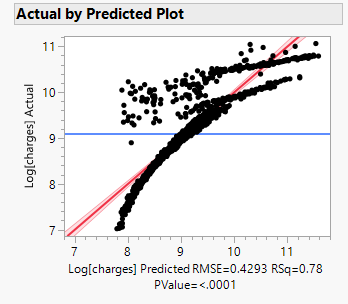
This model is meant to test whether the terms (i.e. obesity indicator and smoker interactions) as well as adding the allowance for a curve in the relationship between age and charges will contribute significantly to the prediction of medical costs. The addition of the quadratic term is based on the hypothesis that the relationship between age and costs is not linear. In order to avoid issues of collinearity with the new quadratic term, age was standardized through JMP’s standardize method which subtracts the group mean and divides by the standard deviation of the group. Testing whether the new terms added are statistically significant in contributing information for the prediction of charges will be assessed using a partial F-test. The output from JMP for this model can be found in the script group Model 2 of the insurance.jmp workbook as well as in Appendix C.

### Model 3 (extends model 2 with region interactions over every previous term):

This model is meant to provide us with the information needed to make an informed conclusion about whether the factors that affect medical costs remain consistent across the different regions of the US. We acknowledge that the more hierarchical and complete approach of taking all k-way interactions between our qualitative and quantitative predictors would allow us to have more certainty in our conclusion but due to the large number of terms (100+), interpretation of that model is not practically useful and a trade off between statistical robustness and practicality was made. To determine whether there are any significant interactions between region and our predictors from model 2, we’ll perform a partial f-test to see if the new terms added are statistically significant in contributing information for the prediction of charges like we did for model 2. See the script group Model 3 of the insurance.jmp workbook as well as Appendix D for the outputs from JMP for this model.

## Results:

### Model 1:

A screenshot of a computer

AI-generated content may be incorrect.

Appendix B, Figure B1 (above) and Appendix B, Figure B2 (right)

Our baseline model fits the data reasonably well, with nearly all terms being significant. Looking at appendix B, figure B1 we can see that our model does not account for all the relationships between our predictors and charges though. This model accounts for about 78% of the variance in charges after adjusting for our predictors (appendix B, figure B3) and is considered to be statistically useful for prediction of charges based on the global F-test (F=534, p < .001, appendix B, figure B4). Normal distribution and constant variance assumptions are satisfied to the best of our ability but studentized residuals show the presence of high leverage outliers (see appendix B figures B5-B7 for residual plots). Due to the nature of the data, and the robustness of regression to outliers no action is taken to exclude outliers at this point.

### Model 2:

A graph with black dots and red line

AI-generated content may be incorrect.A screenshot of a computer screen

AI-generated content may be incorrect.

Appendix C, Figure C1 (above) and Appendix C, Figure C2 (right)

The addition of the quadratic term for age, obesity indicator, and smoking interactions for age and obesity have improved on model 2. Comparing figure B1 and figure C1 we can see that model 2 more accurately reflects the relationship between our predictors and charges, accounting for about 83% of the variance in charges after adjusting for our predictors (appendix C, figure C3). This model is statistically useful for prediction of charges based on the global F-test (f=472.341, p<.0001, appendix C, figure C4) as well. Assumptions around the error component are approximately satisfied by this model (see appendix D figures B5-B6) and looking at the studentized residual plot, we see that the residuals are now more tightly contained within of 0, but the presence of high leverage outliers is still there (see appendix C, figure C7).

Visually and based on the value we can see that adding the extra terms do significantly contribute information to the prediction of charges beyond the first models capability, this is validated by a partial F-test (see appendix C, figure C8 for full output).

A close-up of a number

AI-generated content may be incorrect.

Formally:

Given an F-statistic of 79.23, and a p-value less than .001 (p < .001), the null hypothesis is rejected at the .01 significance level (α = .01). The inclusion of the new terms significantly improves the prediction of charges.

### Model 3:

A graph with black dots

AI-generated content may be incorrect.

Appendix D, Figure D1 (above) and Indicator Function Parameterization can be found in Appendix D, Figure D2

This model appears to be slightly better at predicting charges at first glance when comparing figure D1 and figure C1 but looking at the indicator function parameterization most of the new terms show no statistical significance (see appendix D, Figure D2). Additionally, this model accounts for the same amount of variance in charges after adjusting for our predictors as model 2 (83%, see appendix D, Figure D3). This model is statistically useful for prediction of charges based on the global F-test (f=143.5069, p<.0001, appendix D, figure D4) and the assumptions about the error seem to be as close to satisfied as we could reasonably get with the data we had as the other 2 models (see appendix d, figures D5-D7).

The question we were hoping to answer with this model was whether region had a significant impact on our other predictors, that is, whether the factors that influence medical costs are consistent across regions. Looking at the effect summary (appendix D, figure D8) the answer would appear to be **yes** the factors that influence medical costs are consistent across different regions of the U.S. if you were to look at any one interaction individually; however the partial f-test was also performed across all the terms added, which signals that at least one of the added terms effects are significant (at the .05 significance level). (see appendix D, figure D9 for the full output).

A close-up of a card

AI-generated content may be incorrect.

From appendix D, figure D9, partial f-test statistics table

Formally:

Given an F-statistic of 1.47, and a p-value less than .05 (p =.04), the null hypothesis is rejected at the .05 significance level (α = .05). The inclusion of the new terms significantly improves the prediction of charges, suggesting that at least one of the factors that effects medical costs does not remain consistent across the regions of the U.S.

### Model comparison and recommendations:

|  |  |  |  |
| --- | --- | --- | --- |
| Model |  | RMSE | BIC |
| 1 | 0.78205 | 0.429282 | 1603.295 |
| 2 | 0.831639 | 0.377298 | 1288.831 |
| 3 | 0.833599 | 0.375096 | 1476.929 |

Summary table for all 3 models summary of fit statistics

Out of the three models fit, we would recommend using model 2 for the prediction of medical costs across the US. Model 2 and model 3 provide very similar predictive capabilities, with model 3 showing slightly higher and lower RMSE, but at the cost of increased complexity (15 terms in model 2 versus 47 terms in model 3). This is reflected in the BIC for the 2 models, which based on that criteria would put model 2 firmly in the best model spot.

Considering our selection of model 2 as our preferred model for predicting medical costs, this is practically equivalent as saying that the factors that influence medical costs are consistent across the different regions of the U.S. which is what (based on our data and model selection) is the practical answer to the question. It is important to note that the partial F-test for model 3 did suggest that at least one of the interaction terms for region and our predictors was significant but when assessing the effect of our parameters through the parameter effect summary none of them came back as significant.

## Conclusions and Discussion:

Our project examined whether the factors driving medical costs are the same across different regions of the US, using data simulated from U.S. Census bureau statistics from around 2013. We looked at total annual medical charges alongside personal factors like age, sex, body mass index (BMI), number of children, smoking status, and region.

The initial analysis highlighted a few clear patterns. Notably, smokers and individuals with high BMI (in the obese range) tend to incur **significantly** higher medical charges on average compared to non-smokers and those at healthy weights. Regional differences in average costs were relatively small overall. The southeast region stood out with slightly higher average charges, but this was explained by that region having more smokers and obese individuals compared to the other regions. In other words, the higher costs seen in the southeast were likely to do with the higher prevalence of costly risk factors, rather than the region itself causing higher expenses.

Through model building and comparison, we ended up picking the model that balanced predictive accuracy and simplicity as the preferred model for predicting medical costs based on the factors available in our data. This model assumed that region had no impact on the other factors that effect medical costs, and performed comparably to one that did which lead us to the conclusion that, for the most part, and for all practical purposes region is not important in determining how much the impacts of other high risk factors such as obesity or smoking will contribute to average medical costs.

In the end, our analysis indicates that the major driver of higher medical costs – particularly smoking and obesity – are consistent nationwide. Any regional cost difference seems to stem from how common those risk factors are in each region, rather than the region altering how those factors impact medical expenses.

# References

Centers for Disease Control and Prevention. (2024, 3 19). *Adult BMI categories*. Retrieved from CDC: https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

Choi, M. (n.d.). *Medical Cost Personal Datasets*. Retrieved from www.kaggle.com: https://www.kaggle.com/datasets/mirichoi0218/insurance/data

Duea, J., & Smith, K. (2025, 28 4). *Modeling-Medical-Cost-Differences-Across-the-US*. Retrieved from https://github.com: https://github.com/jmduea/Modeling-Medical-Cost-Differences-Across-the-US

Lantz, B. (2013). Example - predicting medical expenses using linear regression. In B. Lantz, *Machine Learning with R* (pp. 361-388). Packt Publishing. Retrieved from https://www.everand.com/book/253051595/Machine-Learning-with-R

Stewart, S. T., Cutler, D. M., & Rosen, A. B. (2009, 12 3). Forecasting the Effects of Obesity and Smoking on U.S. Life Expectancy. *New England Journal of Medicine, 361*(23), 2252-2260. doi:10.1056/nejmsa0900459

# Appendix A

**Exploratory Data Analysis Figures and Tables**

Figure A1

*Descriptive summary of variables (n=1338)*

**A table with numbers and letters

AI-generated content may be incorrect.**

Figure A2

*Distribution of charges vs log(charges)*

A screenshot of a computer

AI-generated content may be incorrect.

Figure A3

*Distribution of charges (USD) by region of the US*

A diagram of distribution of charges

AI-generated content may be incorrect.

Note. Variance in charges across the four different regions of the US are not constant, with the most pronounced differences between the southeast and southwest regions.

Figure A4

*Charges (USD) across different BMI values*

A graph of red and blue dots

AI-generated content may be incorrect.

Note. Variance in charges nearly doubles at the 30 BMI mark, research into the interaction of obesity (BMI>30) and smoking (Stewart, Cutler, & Rosen, 2009) provides some context on why we might be seeing this in our data.

Figure A5

*Distribution of charges (USD) by smoking status*

A screenshot of a graph

AI-generated content may be incorrect.

Note. The cost differences between smokers and non-smokers is significant, as well as the variance in costs for the two groups.

Figure A6

*Charges (USD) across BMI values, colored by smoking status*

A graph of red and blue dots

AI-generated content may be incorrect.

Note. Comparing the smoking and non-smoking groups, the spread of charges remains constant across all BMI values for non-smokers. Charges for smokers increase across BMI values, with a marked increase after the 30 BMI mark.

Figure A7

*Average charges (USD) by region of the US*

A graph of blue rectangular bars

AI-generated content may be incorrect.

Figure A8

*Counts of individuals by region, grouped by smoking status*

A graph of different colored columns

AI-generated content may be incorrect.

Figure A9

*Counts of individuals by region, grouped by BMI > 30*

A graph of red and blue bars

AI-generated content may be incorrect.

Figure A10

*Raw Charges (USD) vs BMI, faceted by Region and colored by smoker status*

*A screenshot of a graph

AI-generated content may be incorrect.*

Note. This figure pulls together the four insights from our EDA. In all four regions blue points (smokers) sit consistently higher than red points (non-smokers), confirming the significant difference in costs for smokers. Each region also shows the spread of charges increasing past the obesity threshold (BMI>30), with obese smokers seeing the highest jump in charges. The southeast region is clearly an outlier as well, with the cloud of high BMI smokers being larger than the other 3 regions.

Figure A11

*Log(Charges) (USD) vs BMI, faceted by Region and colored by smoker status*

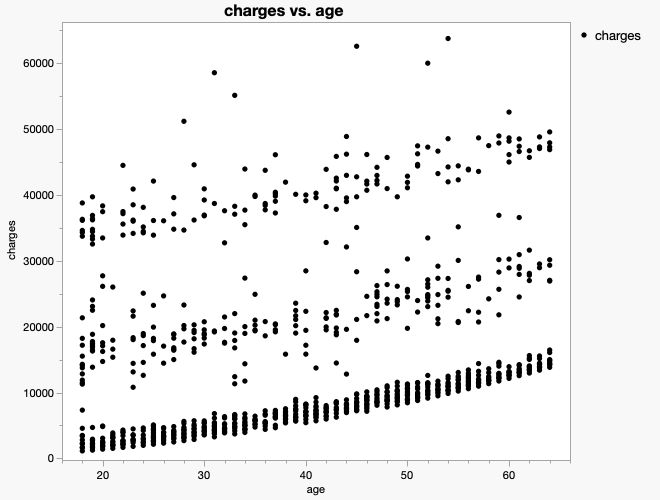
*A screenshot of a graph

AI-generated content may be incorrect.*

Note. This figure shows the same relationship as Figure A10, but on the logarithmic scale. The log transformation effectively evens out the spread for each region and satisfies the constant-variance assumption. The prevalence of high BMI smokers in the southeast region is still noticeable as well. This figure provides reassurance that the log transformation stabilizes variance without masking our key interactions identified.

Figure A12

*Charges in relation to age.*



# Appendix B

**Model 1 Fit and Diagnostics**

Figure B1

*Model 1 Actual by predicted plot*

*A graph with black dots and red lines

AI-generated content may be incorrect.*

Figure B2

*Model 1 Indicator Function Parameterization*

A screenshot of a computer

AI-generated content may be incorrect.

Figure B3

*Model 1 Summary of Fit*

*A screenshot of a computer

AI-generated content may be incorrect.*

Figure B4

*Model 1 Analysis of Variance*

*A screenshot of a graph

AI-generated content may be incorrect.*

Figure B5

*Model 1 Residual Normal Quantile Plot*

*A graph with a line drawn on it

AI-generated content may be incorrect.*

Figure B6

*Model 1Residual by Predicted plot*

*A screen shot of a graph

AI-generated content may be incorrect.*

Figure B7

*Model 1 Studentized Residuals*

*A screen shot of a graph

AI-generated content may be incorrect.*

# Appendix C

**Model 2 Fit and Diagnostics**

Figure C1

*Model 2 Actual by predicted plot*

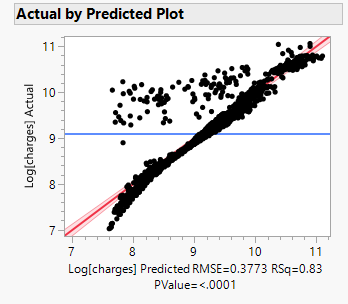


Figure C2

*Model 1 Indicator Function Parameterization*

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure C3

*Model 2 Summary of fit*

*A screenshot of a computer

AI-generated content may be incorrect.*

Figure C4

*Model 2 Analysis of Variance*

*A screenshot of a graph

AI-generated content may be incorrect.*

Figure C5

*Model 2 Residual Normal Quantile Plot*

*A graph with lines and numbers

AI-generated content may be incorrect.*

Figure C6

Model 2 Residual by Predicted plot

A screen shot of a graph

AI-generated content may be incorrect.

Figure C7

*Model 2 Studentized Residuals*

*A screen shot of a graph

AI-generated content may be incorrect.*

Figure C8

*Model 2 Partial F-test*

*A screenshot of a computer

AI-generated content may be incorrect.*

# Appendix D

**Model 3 Fit and Diagnostics**

Figure D1

*Model 3 Actual by predicted plot*

*A graph with black dots

AI-generated content may be incorrect.*

Figure D2

*Model 3 Indicator Function Parameterization*

*A screenshot of a computer

AI-generated content may be incorrect.*

Figure D3

*Model 3 Summary of fit*

**A screenshot of a computer

AI-generated content may be incorrect.**

Figure D4

*Model 3 Analysis of Variance*

*A screenshot of a graph

AI-generated content may be incorrect.*

Figure D5

*Model 3 Residual Normal Quantile Plot*

*A graph with a line drawn on it

AI-generated content may be incorrect.*

Figure D6

*Model 3 Residual by Predicted Plot*

*A screen shot of a graph

AI-generated content may be incorrect.*

Figure D7

*Model 3 Studentized Residuals*

*A screen shot of a graph

AI-generated content may be incorrect.*

Figure D8

*Model 3 Effect Summary*

*A screenshot of a computer

AI-generated content may be incorrect.*

Figure D9

*Model 3 Partial F-test*  
A white screen with a grid

AI-generated content may be incorrect.