Understanding The Factors Affecting Health Care Charges Using SAS Procedures

By Shrinidhi Rajesh

03/16/2023

Abstract:

Healthcare costs are a significant concern in the United States, with an estimated $3.8 trillion spent in 2019. To better understand the factors contributing to these costs and develop more accurate pricing models, this study investigated the relationship between age, BMI, number of children covered, smoking status, gender, region, and medical charges among health insurance beneficiaries. The main goal of the study was to determine the clinical factors that contribute to health insurance charges, to identify which factors play a role in determining these costs, and to investigate significant differences in healthcare charges across different regions, based on the number of children covered by the insurance plan and between smokers and non-smokers, different age groups, and genders. The dataset consisted of the details of 1338 insurers and seven variables. The findings indicate that age, BMI, number of children covered, smoking status, gender, and region all have a significant impact on health insurance charges. These findings can help insurance companies and healthcare providers to develop targeted interventions and strategies to manage these factors and reduce healthcare costs for patients and their clients.

Table of Contents:

Background of the Study

Health insurance is a critical component of the healthcare system, providing financial protection to individuals and families against the high costs of medical care. However, the factors that determine health insurance charges can be complex and difficult to comprehend. Understanding how these factors contribute to medical charges is important for insurance companies and healthcare providers to accurately price their policies and provide better care to their customers.

Several studies have attempted to identify the factors that contribute to high health insurance costs, including age, gender, smoking status, BMI, and the number of dependents covered by insurance. For instance, Zhang et al. (2020)[5] came up with a conclusion that higher BMI was associated with higher healthcare costs, while Baker et al. (2019)[6] found that women tend to have higher healthcare costs than men. Similarly, Berman et al. (2019)[7] identified smoking as a significant contributor to higher healthcare costs in his study. Additionally, Gaskin et al(2015)[8] worked on Residential segregation and disparities in healthcare services utilization giving an insight into how segregation caused the usage of different healthcare services. Finding out why health insurance costs are very high has always raised a question and a huge void in this understanding. That is where this study comes into play. This project aims to investigate various factors that affect healthcare charges among health insurance beneficiaries. Specifically, this research will examine the impact of age, BMI, number of children covered by insurance, smoking status, gender, and region on medical charges. The research also identifies predictors of higher medical charges and examines the extent to which these predictors influence the likelihood of incurring higher charges. Additionally, the study examines if there is a significant difference in healthcare charges among individuals from different regions and whether the number of children covered by health insurance as dependents are showing a significant difference in the healthcare charges. Furthermore, the study will also investigate if there is a significant difference in medical charges between smokers and non-smokers, gender and among different age groups.

To explore these objectives, I obtained the health insurance charges dataset from Kaggle which consists of the details of 1338 insurers and seven variables [1]. By analysing this dataset, this research aims to gain a better understanding of the clinical factors that contribute to health insurance charges and to help reduce the burden of healthcare costs on insurers by aiding healthcare providers to come up with a better pricing model.

Methods

The dataset for the study was obtained from Kaggle [1] which comprises of 1338 health insurance beneficiaries and 7 different columns as follows:

- age: age of the primary beneficiary.
- sex: insurance contractor gender, female, male.
- bmi: Body mass index, providing an understanding of the 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.
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance.

Below is a detailed description of the column variables [Table 1]

Table1: Variables and their measurement

Variables	Description	Level of Measurement	Type
age	Age of primary beneficiary	Ratio	Continuous
sex	Insurance contractor gender, female, male	Nominal	Categorical
bmi	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height;	Ratio	Continuous
children	Number of children covered by health insurance / Number of dependents	Ratio	Count
smoker	Smoker or Non Smoker	Nominal	Categorical
region	The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.	Nominal	Categorical
charges	Individual medical costs billed by health insurance	Ratio	Continuous

SAS software was used in this study, the robustness of this tool aided in performing statistical analyses and provided a high degree of flexibility and customization in data manipulation and reporting.

The data obtained from Kaggle was loaded into SAS Software using the Procedure Import and the data were labelled for a better understanding of the variables in the output. After which missing values were checked using the Means procedure with the nmiss parameter, table 2 comprises of its result stating there were no missing values in this report.

Table 2: Checking for missing values:

The MEANS Procedure

Variable	Label	N Miss
age	age beneficiary	0
bmi	Body mass index	0
children	Number of children covered by health insurance / Number of dependents	0
charges	Individual medical costs billed by health insurance	0
sex_status	insurance contractor gender - 0=female, 1=male	0
region_value	Residential area in the US seperated from 0 to 3 based on each region	0
smoker_status	Smoking status - 0=No,1=Yes	0

Additional Variables were used by adding dummy values to the original variable value, the below is the list of additional variables.

- Sex_Status
- Age_group
- Region_value
- Smoker_Status
- Charges_separation
- Charges_seperationmedian

For instance, the sex_status consist of values 0 for females and 1 for males taken from the column sex in the original dataset.

The main objective of this study was to understand the relationship between various factors, such as age, BMI, number of children covered by insurance, smoking status, gender, region, and medical charges. To do this, I first tried to perform a multiple regression. To check the assumption of normality, a univariate procedure was used to plot a histogram. Although a bell curve was observed, the output caused some confusion whether the charges were symmetric or skewed towards the right tail [table 3]. To confirm the skewness, I plotted a QQ plot, which revealed a few outliers [table 4]. The means procedure was performed on continuous variables, age, bmi, charges [Table 5] and proc freq procedures were performed on categorical variables [Table 6]. The correlation matrix was created using the corr procedure and the result revealed that there is no autocorrelation between the variables in the study [Table 7]. The upper quartile region for charges was 16657.72, and there were 332 entries higher than Quartile 3, indicating potential outliers. These outliers can be relevant in practice because healthcare charges can vary widely between individuals due to factors such as pre-existing medical conditions, hospitalization duration, and the type of medical procedures required. Neglecting these outliers could result in a biased analysis and incorrect conclusions especially 332 entries on 1338 observations being removed will definitely cause a wrong prediction model. Therefore, it is important to consider these outliers in this study and their impact on the analysis. With these outliers in mind, I still performed multiple regression to check the result and found that the root MSE value was extremely high making the model fit questionable [table 8].

To account for the right skewness and the way the dataset is, I performed Gamma Regression. I used the SAS procedure Genmod and log as the link function [table 9]. To test the assumption of independence, I performed a scatterplot with the observation and residuals. The plot showed no pattern, ensuring that the assumption of independence was not violated [table 10].

To investigate which predictors contribute to higher medical charges among health insurance beneficiaries and to what extent. I used charges as the response variable but for this model, I split the charges into two making it dichotomous, using the median value as the cutoff. Any charges with value below or equal to median were assigned a dummy value 0 and above the median value was assigned as 1 creating a new variable called charges_seperationmedian. Using this as the response variable, I used all the other 6 variables apart from charges as predictors. Performed Logistic Regression method(descending) using Procedure Genmod and the selection as stepwise[table 11]. Further, I tested for the assumption of independence by using proc scatterplot procedure [table 12], The result showed random data points proving that the assumption of independence was not violated.

I then looked for the crucial factors contributing to higher insurance charges using Quartile 3 as cutoff. In general values above Q3 is categorized as outlier and is removed but in this data there were 332 entries greater than Q3 which is a large proportion in a dataset of 1338 observations. Also in the insurance industry it is pretty common to face right skewed data, thus analysing why these high charges incur for some gives valuable insights on understanding the factors responsible so they can be neutralized and maintained value. Health care charges are used as the response variable here and are split based on charges <= 16657.72 as the base level (0) and charges > 16657.72 as 1, to perform logistic regression. Using the stepwise selection method, the predictors age, BMI, number of children, and smoking status showed a significant relationship with higher medical charges, while gender and region were not statistically significant[table 13]. I performed Logistic Regression(descending) using Procedure Genmod with stepwise selection method. Further I tested for the assumption of independence by using proc scatterplot procedure[table 14], the result showed random datapoints proving that the assumption of independence was not violated.

To determine if there was a significant difference in mean healthcare charges among individuals from the four different regions (southwest, southeast, northwest, and northeast), I performed a one-way ANOVA with region as the independent variable and charges as the dependent variable. I first tested the assumptions of normality and homogeneity. I used univariate procedure to plot the histogram to check the assumption of normality failed. The result failed to produce a normalized distribution, thus I took the Log of charges to make the distribution normalized[table 15] Levene's test was performed to check the assumption of homogeneity and the p value indicated that the assumption of variance failed[table 16]. Thus to check for the difference in charges among regions I used Kruskal Wallis test, the non parametric alternate for ANOVA using the npar1way method [table 17].

To investigate if there was a significant association between the number of children covered by health insurance and individual medical costs billed by health insurance, I performed ANOVA using the proc GLM procedure, with the number of children covered as the independent variable and medical costs as the dependent variable. The assumption of normality failed when I when I used univariate procedure to perform an histogram, the distribution was right skewed [table 18], so I took log of charges and the result showed that the assumption of normality was not much violated. The Levene's test showed that the assumption of variance was violated[table19], thus Kruskal Wallis test was performed as an alternate[table 20].

To investigate the difference in medical charges between smokers and non-smokers, I tested the assumptions of normality and homogeneity using the proc univariate method [table 21]and Levene's test[table 22]. The assumption of normality failed so I took the log of charges and when the assumption of homogeneity failed, I used the non-parametric Wilcoxon test instead[table 23].

To investigate the difference in medical charges between gender. I tested the assumptions of normality and homogeneity using the proc univariate method [table 24]and Levene's test[table 25]. The assumption of

normality failed so I took the log of charges and when the assumption of homogeneity failed, I used the non-parametric Wilcoxon test instead[table 26].

Finally, to check if there is any difference in medical charges between age groups, I divided the age into two groups based on the median value age_category<=39 and age_category>39. Normality assumption was violated so I took the log of response variable[table27]. The Levene's test value was lesser than 0.05, indicating that the assumption of homogeneity was violated [table28]. I performed Wilcoxon test using the proc npar1way method with age group as the independent variable and charges as the dependent variable[table 29].

Discussions and Results:

To analyse the influence of different predictors on health insurance charges. I performed Gamma Regression with charges as a predictor variable and age, BMI, region, smoking status, gender, and children as independent variables, the results were as follows:

Model:
log(Predicted Charges) = 7.3683 + 0.0286(Age) + 0.0140(BMI) + 0.0837(Children) + 1.5004(smoking_status) - 0.0572(Gender) - 0.0396(Region)

The scale was at 4.0945, with a dispersion parameter of 0.2554 and AIC of 26383.4723. None of the predictors' Wald's Confidence Intervals ranged with a value of 0, suggesting that the model predicted significant values. The Pearson Chi-Square value is 623.1569, with 1331 degrees of freedom, and the corresponding value/df ratio is 0.4682. A low value/df ratio suggests a good fit of the model, indicating that the observed values are close to the expected values predicted by the model. The AIC is predicted to be 26383.4723 by the model, which is less than the AIC of the multiple regression (27112.046), indicating that the Gamma model is a good fit comparatively as well. Going forth with the estimations, the model suggested that all the predictors are significant with the interpretations as follows:

For increase in age by 1 year the expected charges increases by a factor of exp(0.0286)= 1.029013 holding all other variables constant, which suggests that the insurance charges increases at a rate of 2.901291% for every one year increase in age holding all other variables constant.

For every one unit increase in the bmi level the expected charges increases by a factor of exp(0.0140)= 1.014098 holding all other variables constant, which suggests that the insurance charges increases at a rate of 1.4 % for every one unit increase in the bmi holding all other variables constant.

For every one additional children(dependents) added to the insurance plan the expected charges increases by a factor of exp(0.0837)= 1.087303 holding all other variables constant, which suggests that the insurance charges increases at a rate of 8.7% for every added children to the insurance plan holding all other variables constant.

For the smokers, the expected charges increases by a factor of exp(1.5004)= 4.483482 holding all other variables constant, which suggest that for smokers the insurance charges increases at a rate of 348% compared to the non-smokers, holding all other variables constant.

For the gender, the expected charges decreases by a factor of exp(-0.0572)= 0.9444052 for males holding all other variables constant, which suggests that for male the insurance charges decreases at a rate of 5.56% compared to the females, holding all other variables constant.

Finally for the regions, the expected charges decreases by a factor of exp(-0.0396)= 0.9611738 compared to northwest which is the base, the northeast charges decreases by exp(-0.0396) 0.9611738, the south west charges decreases by 2* exp(-0.0396)= 1.922348, and the south east charges decreases by 3*exp(-0.0396)= 2.883521, holding all other variables constant. This suggests that for northeast, southwest and southeast regions the insurance charges decreases at a rate of 3%, 92% and 188%, respectively compared to the northwest region, holding all other variables constant.

The result of performing Logistic Regression using median as the cutoff for separating charges as a binomial, AIC was 1856.862. The goodness-of-fit measures, including Likelihood Ratio, Score, and Wald, were all below the significant level of 0.05, indicating that the model is fitted good. Summarizing the estimates,

The log odds of charges increases by 1.092 for every year increase in age and the log odds of charges increases at the rate of 9.2 percent for every one year increase in age. However, with only one predictor variable it is impossible to understand why the charges are higher, thus we might either need more observations or more predictors that can contribute or tell why there is higher medical charges.

Performed Logistic Regression with Quartile 3 as the cut off suggested that the AIC value was 1505.709 and the odds ratio confidence intervals did not include 0 or 1, indicating that the estimates were reliable. The goodness-of-fit measures, including Likelihood Ratio, Score, and Wald, were all below the significant level of 0.05, indicating that the model is fitted good. Summarising the result[Table]

For increase in age by one year the log odds of charges increases by 1.026 holding all other variables constant, the log odds of charges is increasing at the rate of 2.6% for each year increase in age holding all other variables constant.

For increase in bmi by one unit the log odds of charges increases by 1.059 holding all other variables constant, the log odds of charges is increasing at the rate of 5.9% for each unit increase in bmi holding all other variables constant.

For increase in each children i.e dependents added to the insurance increases by one the log odds of charges increases by 1.181 holding all other variables constant, the log odds of charges is increasing at the rate of 18.1% for increase in each children i.e dependents added to the insurance holding all other variables constant.

To check if the difference in charges existed across region, the Kruskal Wallis test yields a chi-square value of 4.7342 with 3 degrees of freedom, and a p-value of 0.1923. This suggests that there is no significant difference in the distribution of charges across the regions at the 0.05 significance level. However, since the p-value is close to the significance level, it is possible that a larger sample size could yield a significant result. Therefore, further investigation may be needed to fully understand the relationship between the variable "charges" and the variable "region".

The Kruskal Wallis test on identifying if there is difference in charges across insurers with number of children count to the insurance showed that the model was statistically significant, with a p-value lesser than 0.001. This indicated that the number of children covered by health insurance was associated with individual medical costs, but with an F-value of 9 accounting for very less variation. This indicates that the number of children covered by health insurance is associated with the individual medical costs billed by health insurance but with very less coverage of variation. For instance, individuals with 0 children covered by health insurance had a mean medical cost of $12,365.98, while those with 5 children covered had a mean medical cost of $8,786.04.

Performing Wilcoxon test on difference in charges between smoker status resulted in a p value lesser than 0.05 indicating that there is a strong evidence that there is a significant difference in medical charges between the smokers and non-smokers.

Wilcoxon test on difference in charges male and female resulted in a p value greater than 0.05 proving that there is no difference in charges among gender. Although from the regression model and the mean values of male and female there is a slight increase in charges for females compared to male the Wilcoxon test states that there is no difference in charges.

Finally using Wilcoxon test to check if there is any difference in charges among age groups, resulted in a pvalue greater 0.05 proving that there is no significant difference in medical charges between the age groups separated by the median. Although as age increases the likelihood of charges increase, may be separating the age group with median as the cut off for performing Wilcoxon states otherwise.

Conclusion

By understanding how different factors influence medical charges, insurance companies can better tailor their pricing models to reflect the actual risks associated with different demographic and health-related factors. This can lead to more accurate and fair pricing for insurance policies, and also help insurers manage their risk more effectively. Additionally, the analysis can also inform public policy discussions around healthcare costs and access by highlighting how different factors contribute to variations in medical charges.

Based on the logistic regression model results, healthcare providers and insurance companies can use the identified factors (age, BMI, children, and smoking status) to develop targeted strategies to manage these factors and reduce healthcare costs for their patients and clients. Also, helps the companies as well as the people to focus on promoting healthy lifestyle choices, such as encouraging smoking cessation.Overall, the logistic regression model can provide valuable insights for healthcare providers and insurance companies to understand the factors that contribute to higher medical charges and develop strategies to manage them.

The research question of comparing healthcare charges across different regions can address several real-world problems, such as identifying disparities in healthcare costs and access to healthcare services. Understanding the variation in healthcare charges across regions can help policymakers and healthcare providers in developing targeted interventions to improve healthcare affordability and accessibility. This also provide insights into the factors driving healthcare costs and help in identifying areas for cost reduction while maintaining quality healthcare services. In this case, no disparities on health care charges across different regions in US, proves that the charges were uniform throughout the country when it comes to health insurance. Overall, the results suggest that the number of children covered by health insurance is associated with the individual medical costs billed by health insurance. This information can be used by insurance providers and policymakers to develop targeted interventions to improve healthcare affordability and accessibility for families with children.

Identifying factors that contribute to differences in medical charges between smokers and non-smokers, will help highlight the importance of smoking cessation programs to reduce healthcare costs. Similarly, no significant difference in charges between males and females, it might indicate that certain gender-specific health issues is neutralized in their driving healthcare costs. Understanding how charges vary based on age can help identify potential areas for improvement in healthcare delivery and cost management. It can also inform public policy decisions related to healthcare access and affordability. Additionally, this analysis can help healthcare providers and insurers better understand how to allocate resources and design insurance plans that are equitable across age groups.

Limitations:

The upper quartile region for the charges was 16657.72, and there were 332 entries higher than the Quartile3, indicating potential outliers. These outliers can be relevant in practicality because healthcare charges can vary widely between individuals due to factors such as pre-existing medical conditions, hospitalization duration, and the type of medical procedures required. Neglecting these outliers could result in a biased analysis and incorrect conclusions. Therefore, it is important to consider these outliers and their impact on the analysis. But considering these outliers, the result of the model is definitely influenced by it. But in general for many fields and studies, when there are outliers they are checked if they are influencial if not are usually removed or altered.

Secondly the Smoker and Non Smoker results in quasi-complete separation, which can occur when a predictor variable perfectly or almost perfectly predicts the response variable. However, the limitation of quasi-complete separation for the smoker variable should be considered. This means that the large coefficient estimate for smokers may not be entirely accurate, as it could be affected by the limited data available for this variable therefore with more observations, the variable smoker can be used to predict accurately on why the high insurance charges occur.

Lastly, the insurance charges can be due to multiple clinical factors. The analysis provided is only concerning the variables in this dataset, when more accuracy is expected other factors have to be attained either through survey or by other online datasets.

REFERENCES

[1] DATA FROM: Kaggle https://www.kaggle.com/datasets/mirichoi0218/insurance

[2] ADDITIONAL RESOURCES: Cody and Smith book to understand the procedures

[3] An Introduction to Generalized Linear Models Dobson A.J., Barnett A.G.

[4] SAS https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/pgmsaswlcm/home.htm for ref on usage of different sas methods.

[5] Zhang, W., & Tao, W. (2015). Study on the influencing factors of medical insurance costs in China. American Journal of Public Health Research, 3(5A), 47-50.

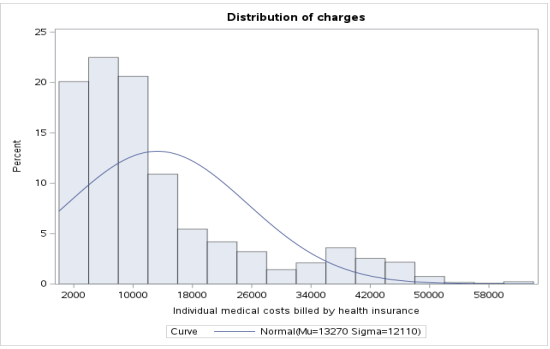
[6] Baker, L. C., & Wheeler, S. K. (1998). Managed care and technology adoption in health care: Evidence from magnetic resonance imaging. Journal of Health Economics, 17(3), 201-218.

[7] Berman, P. C., Giffin, R. B., & Nutting, P. A. (1984). The cost of universal health insurance in California. New England Journal of Medicine, 310(22), 1437-1442.

[8] Gaskin, D. J., Dinwiddie, G. Y., Chan, K. S., & McCleary, R. R. (2015). Residential segregation and disparities in health care services utilization. Medical care research and review, 72(2), 182-191. I have a column region in my dataset.

Tables and Plots

Table 3: Distribution of the Response Variable Charge



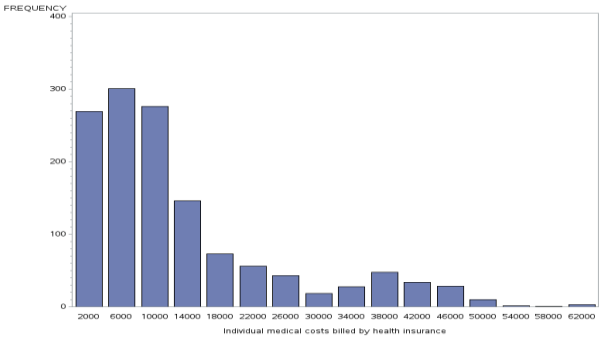
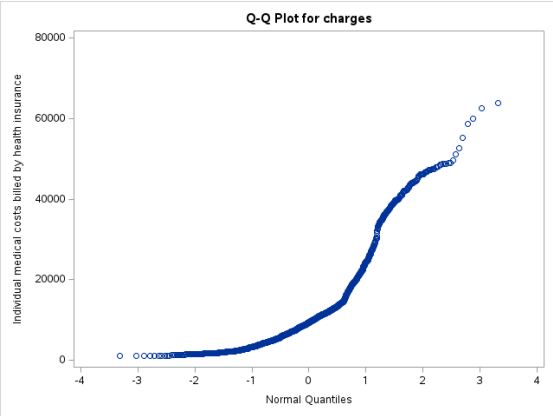


Table 4: Univariate Procedure for QQ Plot of the Response Variable Charges along with the box plot



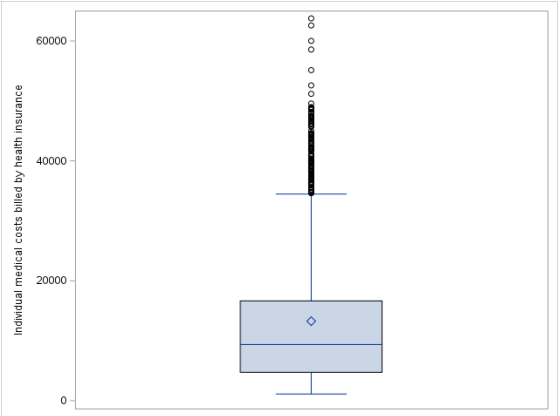


Table 5: Means Procedure for the Continuous Varaiable

The MEANS Procedure

Variable	Label	Mean	Median	Skewness	Std Dev	Variance	Lower Quartile	Upper Quartile
age	age beneficiary	39.21	39.00	0.06	14.05	197.40	27.00	51.00
bmi	Body mass index	30.66	30.40	0.28	6.10	37.19	26.29	34.70
children	Number of children covered by health insurance / Number of dependents	1.09	1.00	0.94	1.21	1.45	0.00	2.00
charges	Individual medical costs billed by health insurance	13270.42	9382.03	1.52	12110.01	146652372.15	4738.27	16657.72

Means Procedure classified by Gender

The MEANS Procedure

insurance contractor gender	N Obs	Variable	Label	N	Mean	Median	Skewness	Std Dev	Variance	Kurtosis
female	662	age	age beneficiary	662	39.50	40.00	0.03	14.05	197.52	-1.24
		bmi	Body mass index	662	30.38	30.11	0.25	6.05	36.55	-0.25
		children	Number of children covered by health insurance / Number of dependents	662	1.07	1.00	0.94	1.19	1.42	0.19
		charges	Individual medical costs billed by health insurance	662	12569.58	9412.96	1.73	11128.70	123848048.29	2.75
male	676	age	age beneficiary	676	38.92	39.00	0.08	14.05	197.41	-1.24
		bmi	Body mass index	676	30.94	30.69	0.32	6.14	37.70	0.13
		children	Number of children covered by health insurance / Number of dependents	676	1.12	1.00	0.94	1.22	1.49	0.22
		charges	Individual medical costs billed by health insurance	676	13956.75	9369.62	1.34	12971.03	168247513.29	0.82

Means Procedure classified by Smoking Status

The MEANS Procedure

Smoking status	N Obs	Variable	Label	N	Mean	Median	Skewness	Std Dev	Variance
no	1064	age	age beneficiary	1064	39.39	40.00	0.03	14.08	198.34
		bmi	Body mass index	1064	30.65	30.35	0.28	6.04	36.52
		children	Number of children covered by health insurance / Number of dependents	1064	1.09	1.00	1.00	1.22	1.48
		charges	Individual medical costs billed by health insurance	1064	8434.27	7345.41	1.54	5993.78	35925420.50
yes	274	age	age beneficiary	274	38.51	38.00	0.16	13.92	193.86
		bmi	Body mass index	274	30.71	30.45	0.31	6.32	39.93
		children	Number of children covered by health insurance / Number of dependents	274	1.11	1.00	0.65	1.16	1.34
		charges	Individual medical costs billed by health insurance	274	32050.23	34456.35	0.13	11541.55	133207311.21

Means Procedure classified by Region

The MEANS Procedure

Rhe beneficiarys residential area in the US	N Obs	Variable	Label	N	Mean	Median	Skewness	Std Dev	Variance
northeast	324	age	age beneficiary	324	39.27	39.50	0.02	14.07	197.94
		bmi	Body mass index	324	29.17	28.88	0.23	5.94	35.25
		children	Number of children covered by health insurance / Number of dependents	324	1.05	1.00	0.96	1.20	1.44
		charges	Individual medical costs billed by health insurance	324	13406.38	10057.65	1.49	11255.80	126693102.65
northwest	325	age	age beneficiary	325	39.20	39.00	0.08	14.05	197.45
		bmi	Body mass index	325	29.20	28.88	0.04	5.14	26.39
		children	Number of children covered by health insurance / Number of dependents	325	1.15	1.00	0.64	1.17	1.37
		charges	Individual medical costs billed by health insurance	325	12417.58	8965.80	1.68	11072.28	122595316.36
southeast	364	age	age beneficiary	364	38.94	39.00	0.07	14.16	200.64
		bmi	Body mass index	364	33.36	33.33	0.22	6.48	41.96
		children	Number of children covered by health insurance / Number of dependents	364	1.05	1.00	1.08	1.18	1.39
		charges	Individual medical costs billed by health insurance	364	14735.41	9294.13	1.25	13971.10	195191595.78
southwest	325	age	age beneficiary	325	39.46	39.00	0.05	13.96	194.88
		bmi	Body mass index	325	30.60	30.30	0.16	5.69	32.40
		children	Number of children covered by health insurance / Number of dependents	325	1.14	1.00	1.03	1.28	1.63
		charges	Individual medical costs billed by health insurance	325	12346.94	8798.59	1.68	11557.18	133568388.77

Table 6: Frequency procedure on categorical variables

The FREQ Procedure

insurance contractor gender				
sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
female	662	49.48	662	49.48
male	676	50.52	1338	100.00

Rhe beneficiarys residential area in the US				
region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
northeast	324	24.22	324	24.22
northwest	325	24.29	649	48.51
southeast	364	27.20	1013	75.71
southwest	325	24.29	1338	100.00

Smoking status				
smoker	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	1064	79.52	1064	79.52
yes	274	20.48	1338	100.00

Table 7: Correlation Matrix

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
charges	1338	13270	12110	17755825	1122	63770	Individual medical costs billed by health insurance
age	1338	39.20703	14.04996	52459	18.00000	64.00000	age beneficiary
bmi	1338	30.66340	6.09819	41028	15.96000	53.13000	Body mass index
children	1338	1.09492	1.20549	1465	0	5.00000	Number of children covered by health insurance / Number of dependents
smoker_status	1338	0.20478	0.40369	274.00000	0	1.00000	Smoking status - 0=No,1=Yes
sex_status	1338	0.50523	0.50016	676.00000	0	1.00000	insurance contractor gender - 0=female, 1=male
region_value	1338	1.54410	1.13089	2066	0	3.00000	Residential area in the US seperated from 0 to 3 based on each region

Pearson Correlation Coefficients, N = 1338 Prob > |r| under H0: Rho=0							
	charges	age	bmi	children	smoker_status	sex_status	region_value
charges Individual medical costs billed by health insurance	1.00000	0.29901 <.0001	0.19834 <.0001	0.06800 0.0129	0.78725 <.0001	0.05729 0.0361	0.05699 0.0371
age age beneficiary	0.29901 <.0001	1.00000	0.10927 <.0001	0.04247 0.1205	-0.02502 0.3605	-0.02086 0.4459	-0.00521 0.8489
bmi Body mass index	0.19834 <.0001	0.10927 <.0001	1.00000	0.01276 0.6410	0.00375 0.8910	0.04637 0.0900	0.26183 <.0001
children Number of children covered by health insurance / Number of dependents	0.06800 0.0129	0.04247 0.1205	0.01276 0.6410	1.00000	0.00767 0.7792	0.01716 0.5305	-0.01926 0.4816
smoker_status Smoking status - 0=No,1=Yes	0.78725 <.0001	-0.02502 0.3605	0.00375 0.8910	0.00767 0.7792	1.00000	0.07618 0.0053	0.05393 0.0486
sex_status insurance contractor gender - 0=female, 1=male	0.05729 0.0361	-0.02086 0.4459	0.04637 0.0900	0.01716 0.5305	0.07618 0.0053	1.00000	0.01612 0.5557
region_value Residential area in the US seperated from 0 to 3 based on each region	0.05699 0.0371	-0.00521 0.8489	0.26183 <.0001	-0.01926 0.4816	0.05393 0.0486	0.01612 0.5557	1.00000

Table 8: Multiple Regression – High root MSE Output depiction

The REG Procedure
Model: MODEL1
Dependent Variable: charges Individual medical costs billed by health insurance

Number of Observations Read	1338
Number of Observations Used	1338

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1.471419E11	24523646779	667.06	<.0001
Error	1331	48932340896	36763592		
Corrected Total	1337	1.960742E11			

Root MSE	6063.29877	R-Square	0.7504
Dependent Mean	13270	Adj R-Sq	0.7493
Coeff Var	45.69032		

Table 9 Gamma Regression for right skewed data

The GENMOD Procedure

Model Information		
Data Set	HEALTHIN.HEALTHINSURANCEDATA	
Distribution	Gamma	
Link Function	Log	
Dependent Variable	charges	Individual medical costs billed by health insurance

Number of Observations Read	1338
Number of Observations Used	1338

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	1331	339.9999	0.2554
Scaled Deviance	1331	1392.1464	1.0459
Pearson Chi-Square	1331	623.1569	0.4682
Scaled Pearson X2	1331	2551.5466	1.9170
Log Likelihood		-13183.7361	
Full Log Likelihood		-13183.7361	
AIC (smaller is better)		26383.4723	
AICC (smaller is better)		26383.5806	
BIC (smaller is better)		26425.0637	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	7.3683	0.0798	7.2119	7.5246	8531.87	<.0001
age	1	0.0286	0.0010	0.0267	0.0305	854.60	<.0001
bmi	1	0.0140	0.0024	0.0093	0.0186	34.48	<.0001
children	1	0.0837	0.0114	0.0613	0.1061	53.72	<.0001
smoker_status	1	1.5004	0.0345	1.4327	1.5681	1886.86	<.0001
sex_status	1	-0.0572	0.0272	-0.1105	-0.0038	4.42	0.0356
region_value	1	-0.0396	0.0125	-0.0640	-0.0152	10.10	0.0015
Scale	1	4.0945	0.1523	3.8067	4.4042		

Note: The scale parameter was estimated by maximum likelihood.

Table 10: Assumption of Independence for Gamma Regression

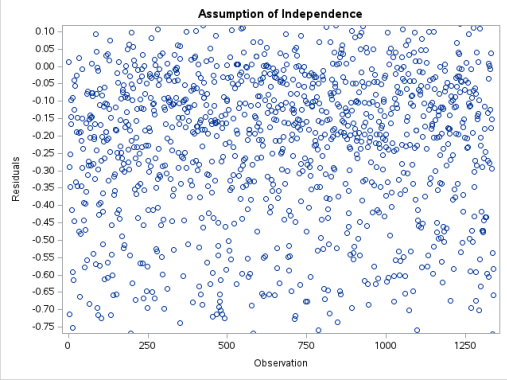


Table 11 Logistic regression on Charges using Median as the Cutoff value for separation

The LOGISTIC Procedure

Model Information	
Data Set	HEALTHIN.HEALTHINSURANCEDATA
Response Variable	charges_seperation_median
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1338
Number of Observations Used	1338

Response Profile		
Ordered Value	charges_seperation_median	Total Frequency
1	1	669
2	0	669

Probability modeled is charges_seperation_median=1.

Step 1. Effect age entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1856.862	1474.675
SC	1862.061	1485.072
-2 Log L	1854.862	1470.675

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	384.1873	1	<.0001
Score	351.5940	1	<.0001
Wald	289.7877	1	<.0001

Summary of Stepwise Selection								
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	age		1	1	351.5940		<.0001	age beneficiary

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.4450	0.2118	264.6241	<.0001
age	1	0.0880	0.00517	289.7877	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.092	1.081	1.103

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	78.7	Somers' D	0.590
Percent Discordant	19.8	Gamma	0.599
Percent Tied	1.5	Tau-a	0.295
Pairs	447561	c	0.795

Table 12: Assumption of Independence for Logistic Regression with Median as the cut off

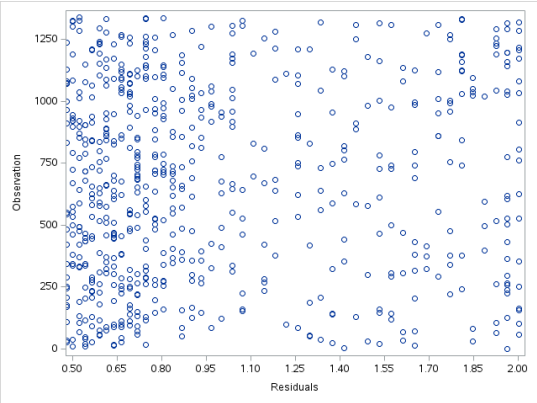


Table 13: Logistic Regression on Charges using Quartile 3 as the cutoff value

The LOGISTIC Procedure

Model Information	
Data Set	HEALTHIN.HEALTHINSURANCEDATA
Response Variable	charges_seperation
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	1338
Number of Observations Used	1338

Response Profile		
Ordered Value	charges_seperation	Total Frequency
1	1	334
2	0	1004

Probability modeled is charges_seperation=1.

Step 4. Effect children entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1505.709	692.476
SC	1510.908	718.471
-2 Log L	1503.709	682.476

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	821.2329	4	<.0001
Score	854.3852	4	<.0001
Wald	360.7508	4	<.0001

Summary of Stepwise Selection								
Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
	Entered	Removed						
1	smoker_status		1	1	844.0402		<.0001	Smoking status - 0=No,1=Yes
2	age		1	2	13.7304		0.0002	age beneficiary
3	bmi		1	3	10.8452		0.0010	Body mass index
4	children		1	4	4.0202		0.0450	Number of children covered by health insurance / Number of dependents

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.5860	0.6604	71.5525	<.0001
age	1	0.0261	0.00793	10.8087	0.0010
bmi	1	0.0565	0.0173	10.6267	0.0011
children	1	0.1658	0.0830	3.9937	0.0457
smoker_status	1	5.3356	0.2826	356.5857	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.026	1.011	1.042
bmi	1.058	1.023	1.095
children	1.180	1.003	1.389
smoker_status	207.598	119.320	361.188

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.3	Somers' D	0.786
Percent Discordant	10.7	Gamma	0.786
Percent Tied	0.0	Tau-a	0.295
Pairs	335336	c	0.893

Table 14: Assumption of Independence with Quartile 3 as the cut off

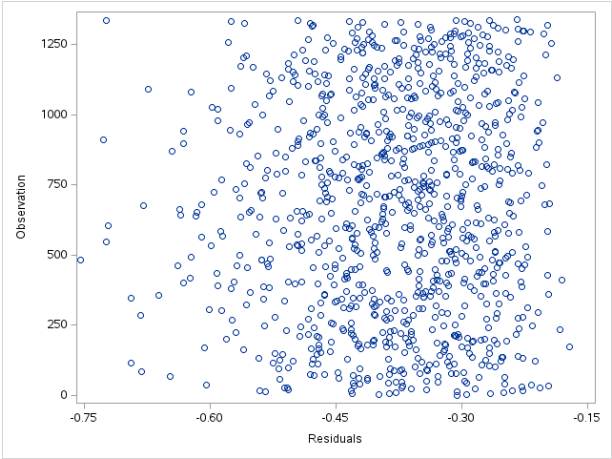
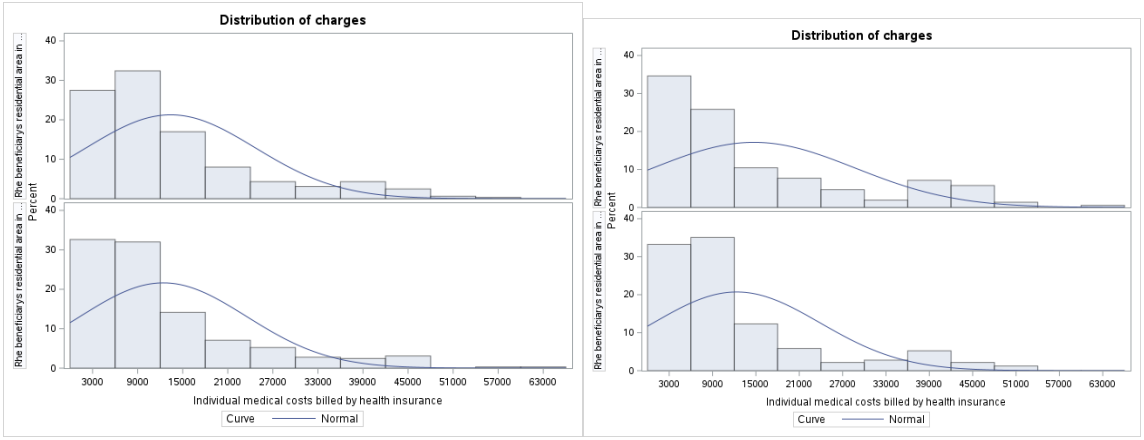
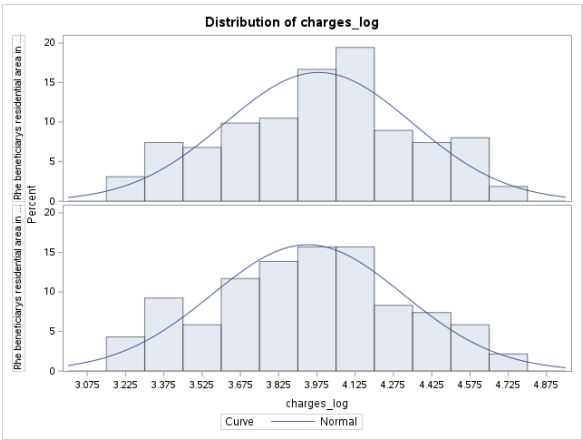


Table 15: Histogram before and after transforming charges to check for the assumption of normality





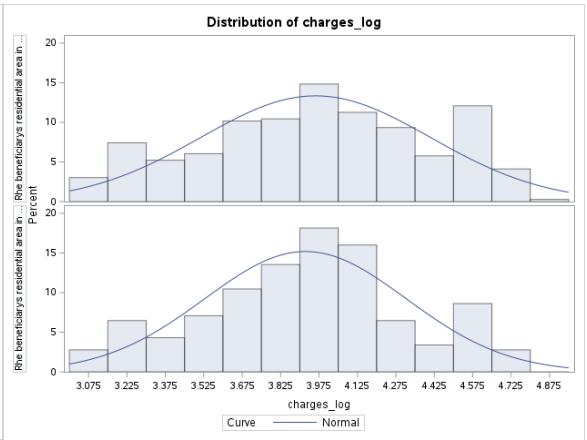


Table 16 Test for Homogeneity

The GLM Procedure

Levene's Test for Homogeneity of charges_log Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
region	3	0.9441	0.3147	9.29	<.0001
Error	1334	45.1842	0.0339		

Table 17: Kruskal Wallis test on separation of charges across regions

The NPAR1WAY Procedure

Analysis of Variance for Variable charges_log Classified by Variable region		
region	N	Mean
southwest	325	3.922156
southeast	364	3.961809
northwest	325	3.938935
northeast	324	3.981946

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	3	0.670162	0.223387	1.4020	0.2406
Within	1334	212.550412	0.159333		
Average scores were used for ties.					

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable charges_log Classified by Variable region					
region	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
southwest	325	208201.0	217587.50	6061.02490	640.618462
southeast	364	247671.0	243698.00	6289.69861	680.414835
northwest	325	212793.0	217587.50	6061.02490	654.747692
northeast	324	227126.0	216918.00	6054.67934	701.006173
Average scores were used for ties.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
4.7342	3	0.1923

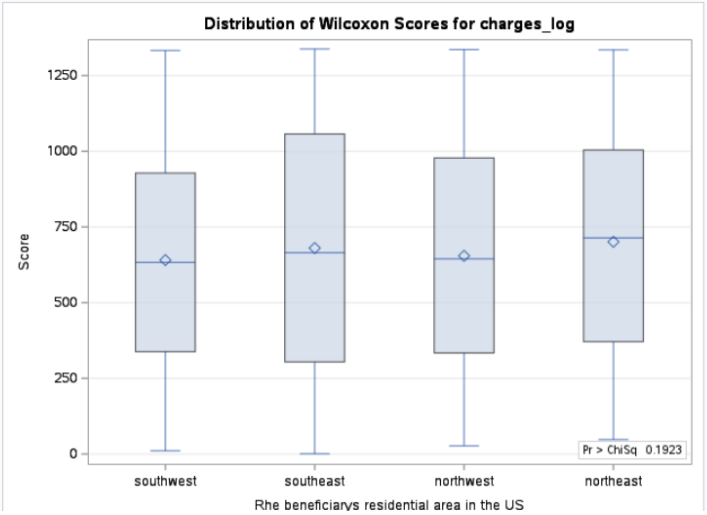


Table 18: Distribution of charges after transforming to check for the assumption of normality

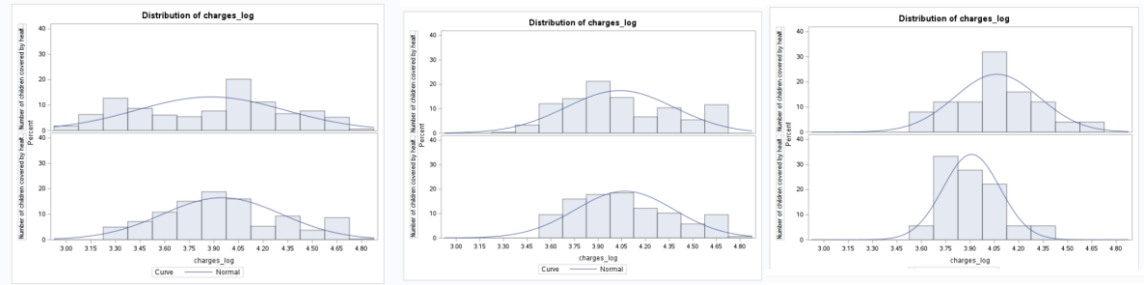


Table 19: Test for Homogeneity

The GLM Procedure

Levene's Test for Homogeneity of charges_log Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
region	3	0.9441	0.3147	9.29	<.0001
Error	1334	45.1842	0.0339		

Table 20: Kruskal Wallis Test

The NPAR1WAY Procedure

Analysis of Variance for Variable charges_log Classified by Variable region		
region	N	Mean
southwest	325	3.922156
southeast	364	3.961809
northwest	325	3.938935
northeast	324	3.981946

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	3	0.670162	0.223387	1.4020	0.2406
Within	1334	212.550412	0.159333		
Average scores were used for ties.					

The NPAR1WAY Procedure

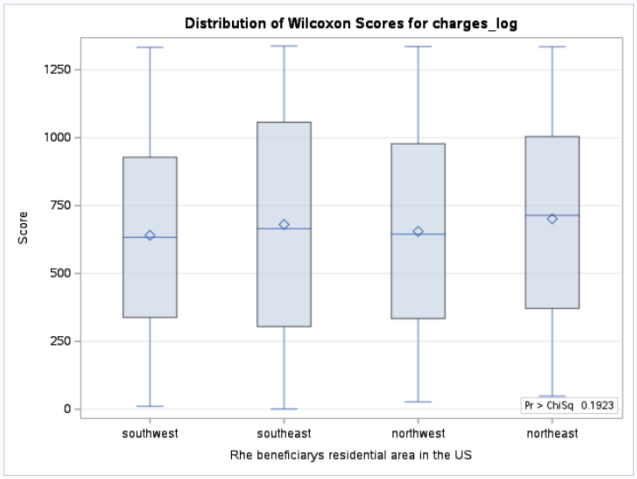
Wilcoxon Scores (Rank Sums) for Variable charges_log
Classified by Variable region

region	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
southwest	325	208201.0	217587.50	6061.02490	640.618462
southeast	364	247671.0	243698.00	6289.69861	680.414835
northwest	325	212793.0	217587.50	6061.02490	654.747692
northeast	324	227126.0	216918.00	6054.67934	701.006173

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
4.7342	3	0.1923



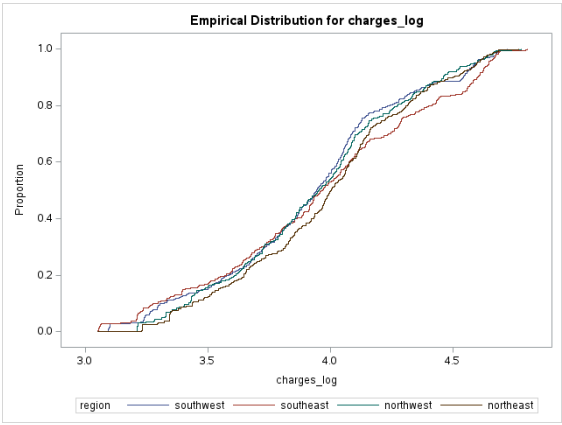


Table 21 Distribution of Charges (SmokervsNonSmoker)

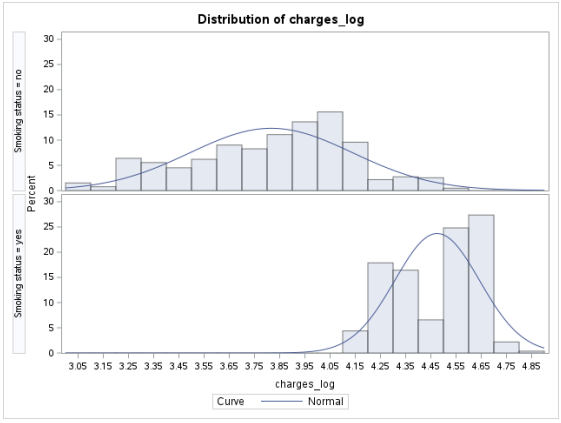


Table 22: Testing the assumption of variance using Levene's test

The GLM Procedure

Levene's Test for Homogeneity of charges_log Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
smoker	1	1.2636	1.2636	96.87	<.0001
Error	1336	17.4276	0.0130		

Table 23: Wilcoxon Test for Smoker vs Non-Smokers

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable charges_log Classified by Variable smoker					
smoker	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
yes	274	321808.0	183443.0	5703.55666	1174.48175
no	1064	573983.0	712348.0	5703.55666	539.45771
Average scores were used for ties.					

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|
321808.0	24.2593	<.0001	<.0001	<.0001	<.0001
Z includes a continuity correction of 0.5.					

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
588.5197	1	<.0001

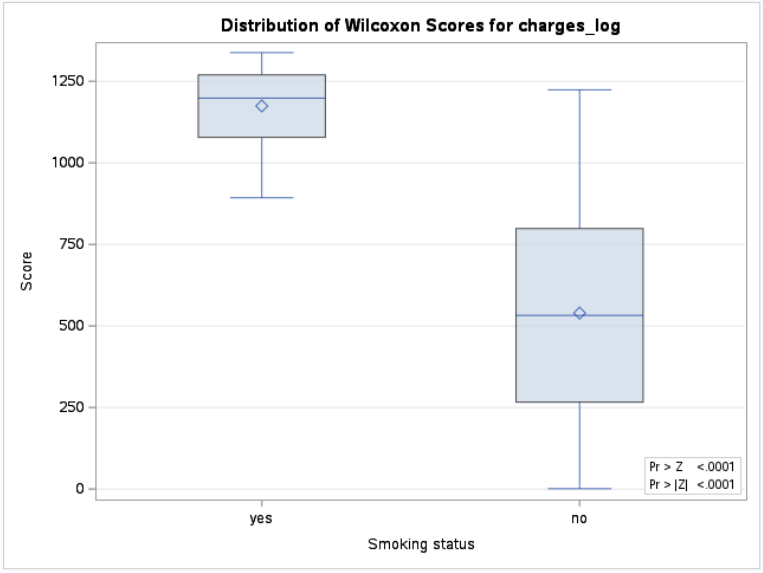


Table 24 Distribution of Charges (Gender)

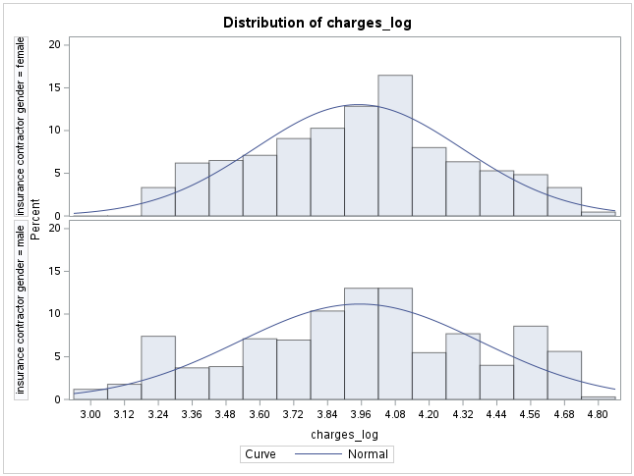


Table 25: Testing the assumption of variance using Levene's test

The GLM Procedure

Levene's Test for Homogeneity of charges_log Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
sex	1	0.7979	0.7979	23.42	<.0001
Error	1336	45.5223	0.0341		

Table 26: Wilcoxon Test for Gender

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable charges_log
Classified by Variable sex

sex	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
female	662	440757.0	443209.0	7066.46167	665.799073
male	676	455034.0	452582.0	7066.46167	673.127219

Average scores were used for ties.

Wilcoxon Two-Sample Test

				t Approximation	
Statistic	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
440757.0	-0.3469	0.3643	0.7287	0.3644	0.7287

Z includes a continuity correction of 0.5.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
0.1204	1	0.7286

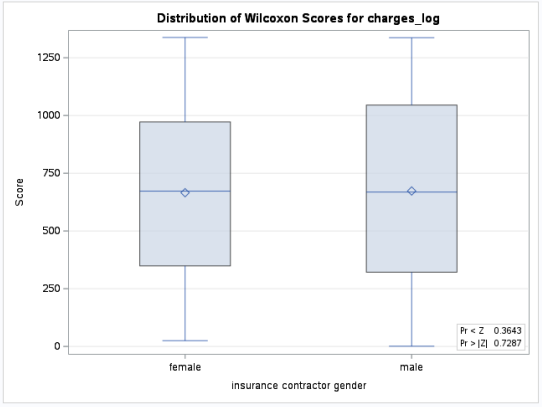


Table 27: Distribution of Charges (Age Group)

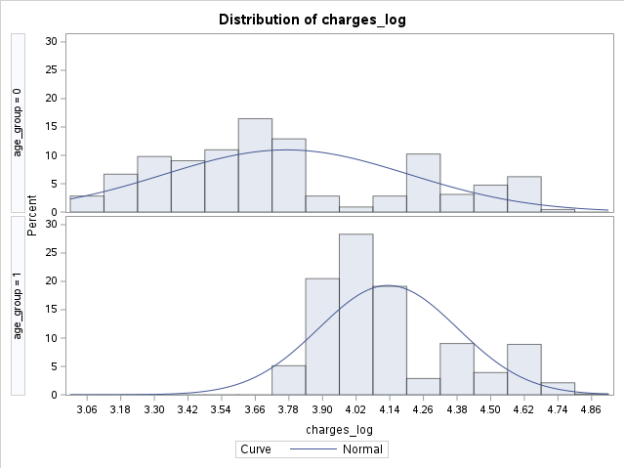


Table 28: Testing the assumption of variance using Levene's test

The GLM Procedure

Levene's Test for Homogeneity of charges_log Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
age_group	1	5.5602	5.5602	220.81	<.0001
Error	1336	33.6422	0.0252		

Table 29: Wilcoxon test for Age groups.

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable charges_log
Classified by Variable age_group

age_group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	674	331101.0	451243.0	7066.65115	491.247774
1	664	564690.0	444548.0	7066.65115	850.436747

Average scores were used for ties.

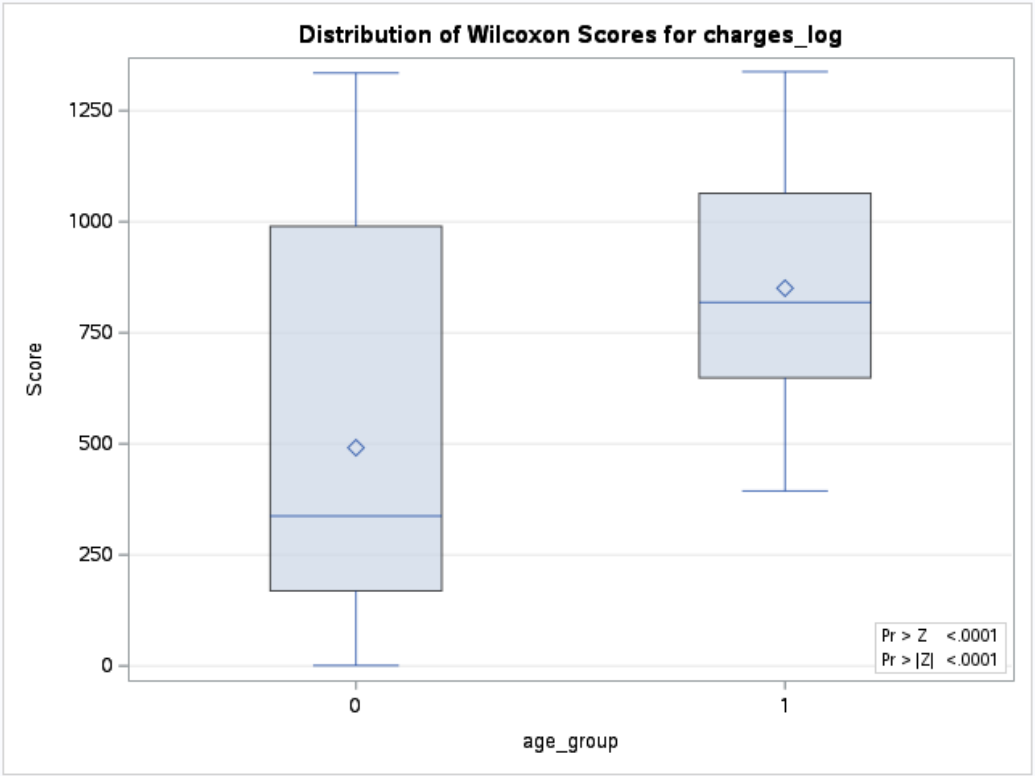
Wilcoxon Two-Sample Test

Statistic	Z	Pr > Z	Pr > |Z|	t Approximation Pr > Z	t Approximation Pr > |Z|
564690.0	17.0012	<.0001	<.0001	<.0001	<.0001

Z includes a continuity correction of 0.5.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
289.0430	1	<.0001



----------END OF REPORT----------