MSDS 601

Linear Regression Analysis

FINAL PROJECT

"PREDICTIVE ANALYTICS FOR MEDICAL INSURANCE COST"

Group Members: Eren Bardak, Ho Nam (Felix) Tong, Mark Lam

TABLE OF CONTENTS

1.1 System Description

The source of the dataset is Kaggle. To access the data, please refer to this link. The dataset contains a total of 1,338 rows and includes both numerical and categorical predictor variables. The columns present are: age, sex, bmi, children, smoker, region, and charges. Below is an initial overview of the data:

```
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

The data set provides a comprehensive view that might affect an individual's medical insurance charges. Each column represents a specific attribute:

- age: This signifies the age of the primary beneficiary.
- sex: It denotes the gender of the insurance contractor, which can either be 'female' or 'male'.
- bmi: Stands for Body Mass Index, a measure that provides an understanding of an individual's body in terms of weight relative to their height. It's an objective index calculated using the formula weight (kg) divided by height squared (m^2). A BMI range of 18.5 to 24.9 is considered ideal.
- children: This column reflects the number of children or dependents covered by the health insurance of the primary beneficiary.
- smoker: A straightforward indication if the individual is a smoker or not.

- region: Points to the residential area of the beneficiary in the US, with categories being 'northeast', 'southeast', 'southwest', and 'northwest'.
- charges: Represents the individual medical costs that are billed by the health insurance.

2.1 Objective

In this paper, the group will particularly focus on the 'charges' column as the response variable to determine its relationship with the predictor variables provided. First, a data cleaning and preprocessing phase will be introduced to ensure the integrity of the 'charges' column and its related data. Following this, emphasis will be placed on findings from exploratory data analysis (EDA), where individual t-tests will be used to determine significant relationships andto investigate potential biases within the data. Proceeding further, the focus will shift to model diagnosis, aiming to establish the relationship of the 'charges' column with predictor variables provided. After finding and evaluating the best-suited model, attention will be given to possible data structural issues or model-related problems, ensuring a well-rounded analysis.

3.1 Research Problems & Summary of Methods

Throughout the analysis, a noteworthy observation surfaced regarding a slight gender bias on BMI. This posed the question of whether the dataset may have structural imbalances that could potentially impact the modeling process. It is crucial to address such biases as they can introduce a skew in predictions, leading to incorrect inferences. Moreover, as the analysis deepened, various data structural problems emerged. The challenges ranged from heteroscedasticity to violations of the normality assumption, despite attempts to rectify them through methods like natural log transformations and outlier removals. Furthermore, the model assumption problems presented themselves. While linear regression models come with certain assumptions, ensuring their satisfaction is important for the model's validity. In this dataset, despite many efforts of the group, the violation of the normality assumption couldn’t be handled, however considering the data set sample size being bigger than one thousand seems to solve the issue for normality thanks to the Central Limit Theorem. There were attempts to apply transformations, especially on predictors like BMI, but these did not yield substantial improvements. Moreover, while the model displayed a bigger than 0.80 adjusted R-square indicating a good fit, tests like the Jarque-Bera and lack of fit F-test underlined the violations of non normality and non linearity. These findings underscore the importance of not just looking at traditional fit metrics but holistically evaluating a model by verifying assumptions and addressing inherent biases and structural challenges.

4.1 Explanatory Data Analysis

4.11 Preprocessing & Cleaning the Data

None of the columns in the dataset include null values, so we will just drop the duplicates and make data type conversions for the categorical columns: sex, smoker, region. Convert them into categorical values from object type.

- **Dropping duplicates:**

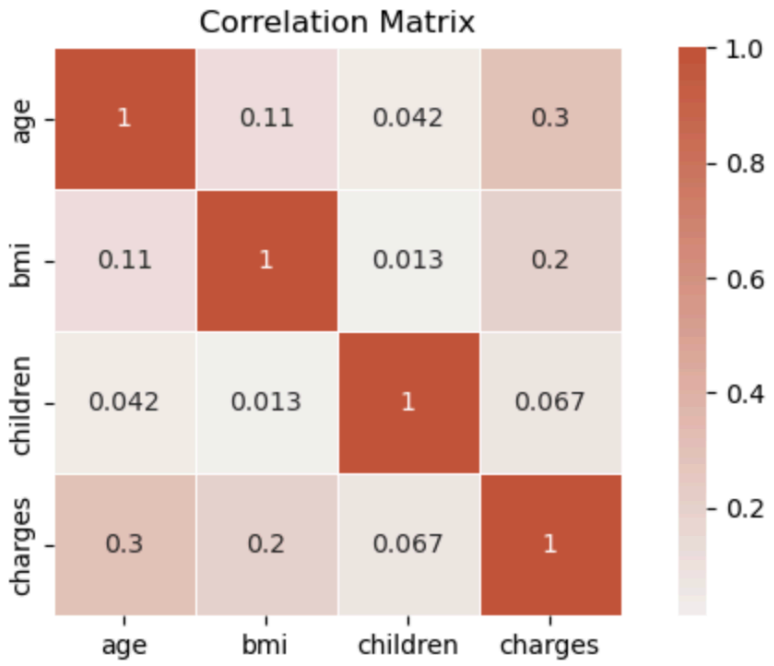
As there are no null values in any column, we can check for duplicate rows to prevent redundancy. Given the size of the dataset and the number of variables, the presence of a duplicate row is likely an error. After dropping it we will have 1337 rows. The dropped row is as follows:

	age	sex	bmi	children	smoker	region	charges
581	19	male	30.59	0	no	northwest	1639.5631

4.12 Summary Statistics

- **Correlation Matrix for Numerical Predictors:**

The correlation matrix for the numerical predictors indicates that there are no extremely strong linear relationships among the variables. The highest correlation is observed between 'age' and 'charges', with a value of 0.298308, but this is still considered a relatively weak correlation. The absence of high correlation values suggests that multicollinearity might not be a concern for this dataset.

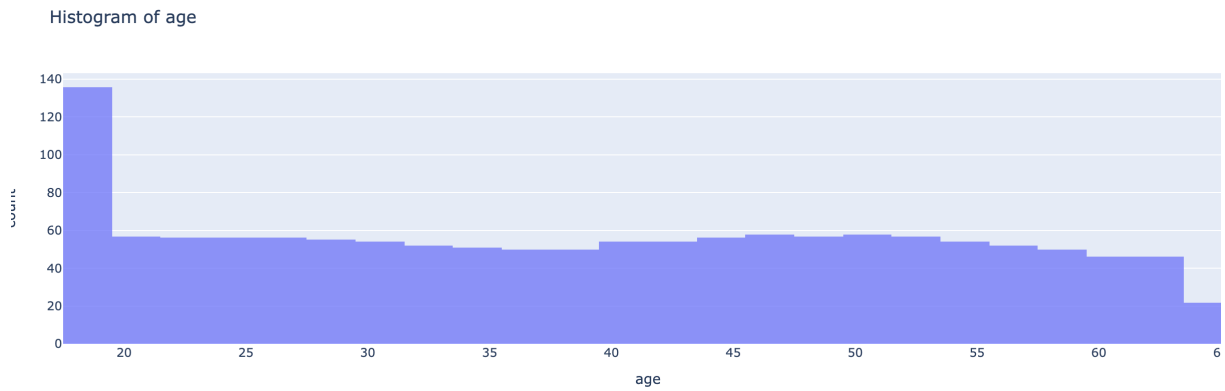


- **Demographics for Numerical Predictors:**

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

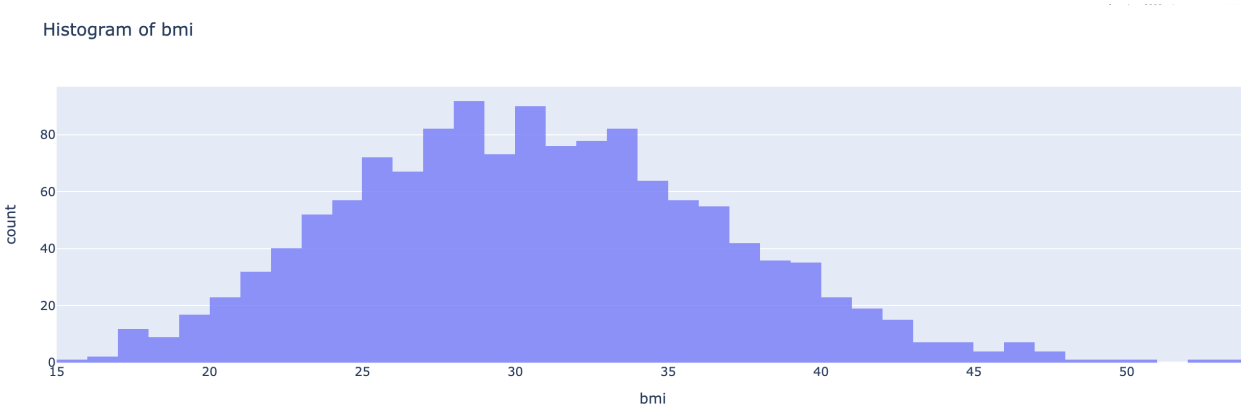
- Age:

The minimum age in the dataset is 22, the maximum age is 69, and the mean age is 39. According to the histogram graph, the age distribution seems to resemble a uniform distribution, except for ages close to 20. This wide age range suggests a diverse dataset in terms of age groups, which indicates that it may be representative of a broader population.



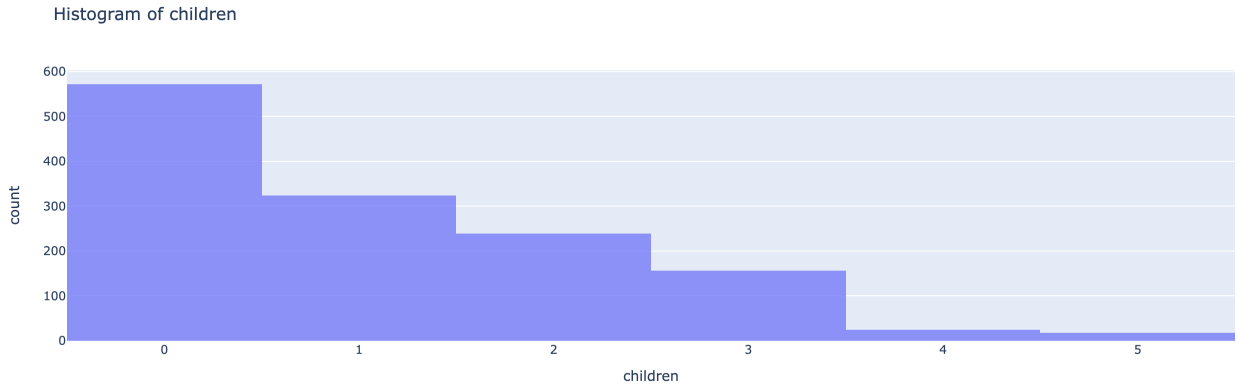
- Bmi:

The distribution of BMI appears to be close to a normal distribution, exhibiting a shape close to the bell-shaped curve. The mean value for BMI is 30.66, with the lowest value being 15.96 and the highest at 53.13. BMI seems to be a standard metric in this dataset, highlighting its importance, especially due to the well-known association between BMI and various health indicators.



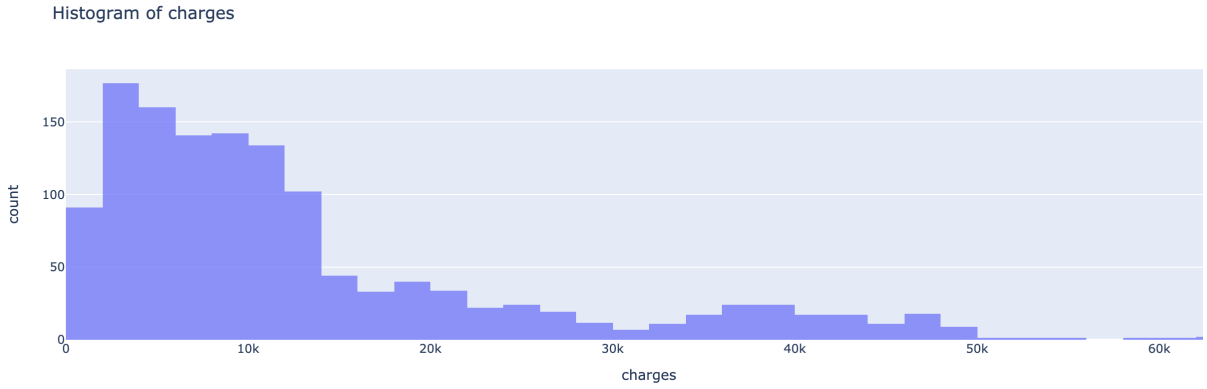
- Children:

In the children column, the maximum number of children is 5, while the minimum is 0. About 43 percent of individuals in the dataset have no children, and the next most common count is one child, represented by 24.23% of the dataset.



- Charges:

For the charges column, the average charge is 13,270, while the range spans from a minimum of 1,122 to a maximum of 63,770. The median charge stands at 9,382, and the third quartile (or the 75th percentile) is 16,640. Only 12.12% of the data has charges above 30,000. It's worth noting that a relatively small proportion of individuals in the dataset bear extremely high charges, suggesting that there could be significant skewness, which may impact the predictive modeling of the 'charges' column.



- **Cross Table for Categorical Predictors:**

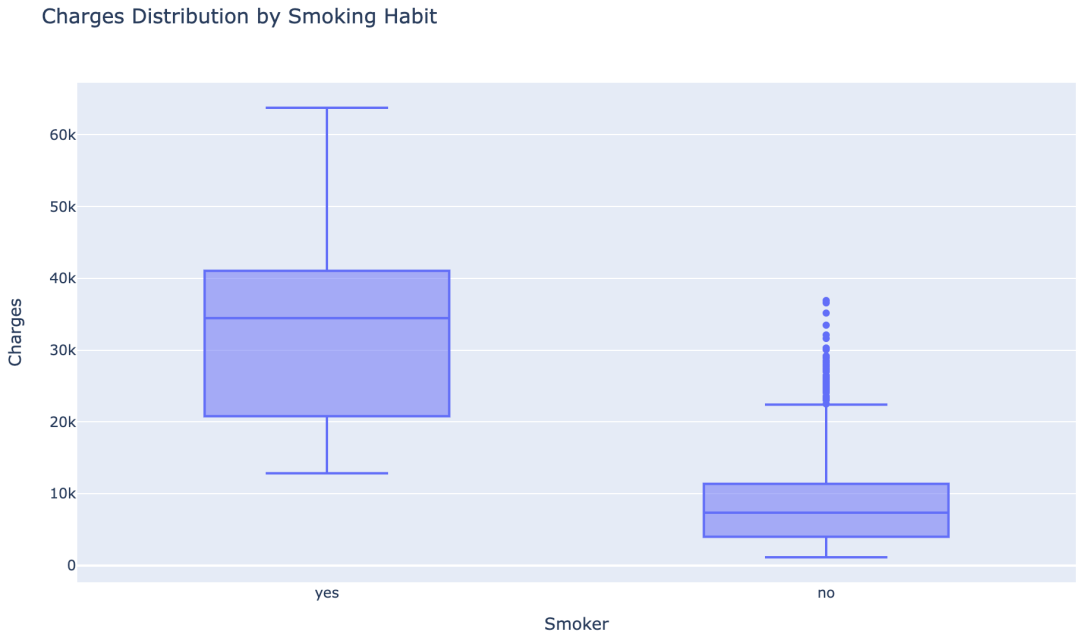
It seems there is no deliberate imbalance for the region and gender columns. However, people who don't smoke make up approximately 80 percent of the data. Since we have a sample size larger than 1000, it's not going to be a problem unless we have extreme imbalances. This distribution can potentially have a significant impact on the charges and should be kept in mind during the model diagnosis and evaluation stages. Ensuring that such categorical predictors are balanced or their imbalances are acknowledged will enhance the validity of our findings and recommendations.

	northeast	northwest	southeast	southwest
('female', 'no')	132	135	139	141
('female', 'yes')	29	29	36	21
('male', 'no')	125	131	134	126
('male', 'yes')	38	29	55	37

4.13 Individual Tests

- **Smoker vs Non-Smoker:**

The below graph already hints that people who smoke pay more in general.



$$H_0: \mu_{smokers} = \mu_{non\text{-}mokers}$$

Null Hypothesis: There is no difference in charges between smokers and non-smokers.

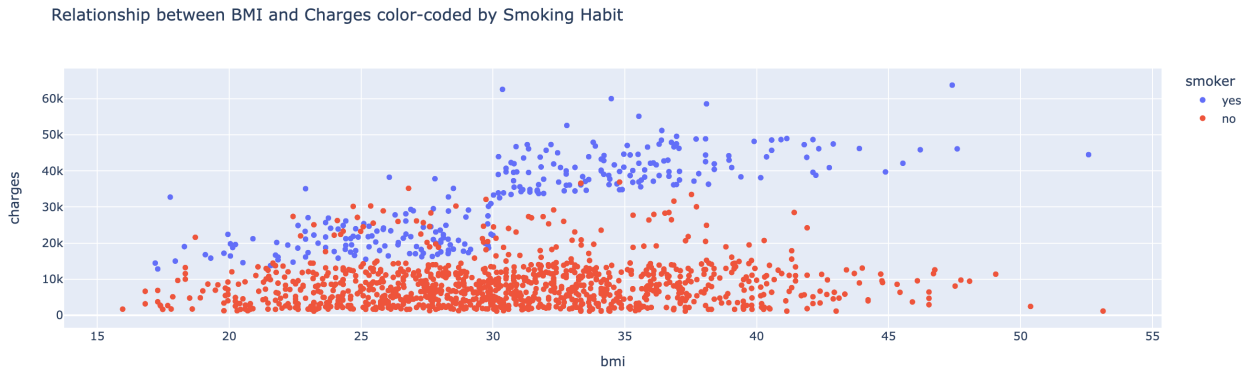
$$H_1: \mu_{smokers} \neq \mu_{non\text{-}mokers}$$

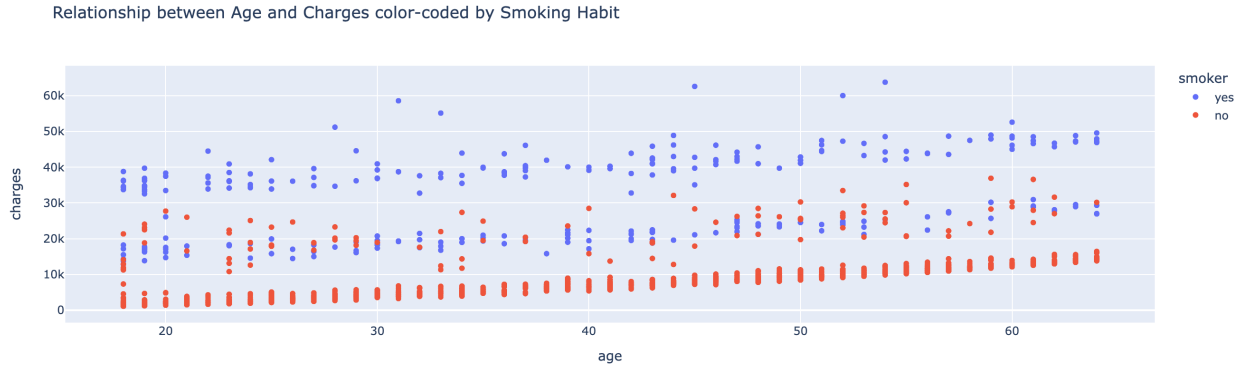
Alternative Hypothesis: There is a difference in charges between smokers and non-smokers.

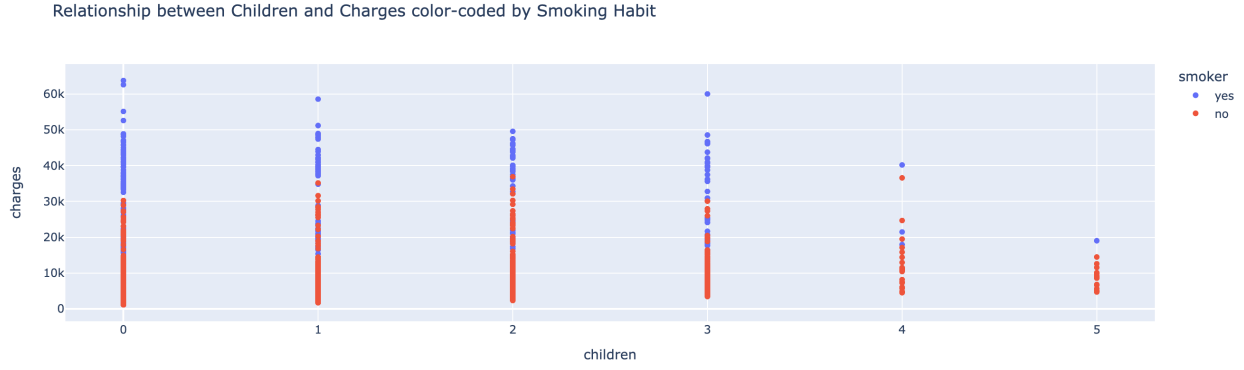
Test result: T-statistic: 46.64, P-value: 0.0000

Conclusion: We reject the null hypothesis at significance level 0.05, and conclude that there is a statistically significant difference in charges between smokers and non-smokers.

The below graphs support that the hypothesis is valid across all levels of bmi, age, and children :







- **BMI_{female} vs BMI_{male}:**

The expected result is BMI being slightly higher for females compared to males, based on real-world data. We are testing if the BMI between males and females is statistically different in an attempt to identify any sampling bias in our dataset. The difference is barely visible in the violin plot.

$$H_0: \mu_{Bmi(female)} = \mu_{Bmi(male)}$$

Null Hypothesis: There is no difference in BMI between females and males.

$$H_1: \mu_{fBmi(female)} \neq \mu_{Bmi(male)}$$

Alternative Hypothesis: There is a difference in BMI between females and males.

Test result: T-statistic for BMI: 1.70, P-value: 0.0899

Conclusion: We fail to reject the null hypothesis at the significance level of 0.05. Considering that there is no difference in BMI, there might be potential gender bias in the sampling of our dataset.

- **BMI Distribution Across Regions:**

$$H_0: \mu_{Northeast} = \mu_{Southeast} = \mu_{Northeast} = \mu_{Southwest} = \mu_{Northwest}$$

Null Hypothesis: There is no difference in BMI accross regions.

$$H_1: \mu_{region(i)} \neq \mu_{region(j)}$$

Alternative Hypothesis: There is a difference in BMI across regions.

Test result: F-Value: 39.48593864487439, P-Value: 1.9087293927440606e-24

Conclusion: We reject the null hypothesis at the significance level of 0.05. This means there is a statistically significant difference in BMI among at least two of the regions.

5.1 Regression Analysis & Model Selection (MLR)

As previously mentioned, our response variable is "charges." Here's a summary of the initial full model:

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.751
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	500.0
Date:	Wed, 11 Oct 2023	Prob (F-statistic):	0.00
Time:	21:01:57	Log-Likelihood:	-13538.
No. Observations:	1337	AIC:	2.709e+04
Df Residuals:	1328	BIC:	2.714e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.194e+04	988.227	-12.079	0.000	-1.39e+04	-9997.900
sex[T.male]	-129.4815	333.195	-0.389	0.698	-783.128	524.165
smoker[T.yes]	2.385e+04	413.348	57.693	0.000	2.3e+04	2.47e+04
region[T.northwest]	-349.2265	476.824	-0.732	0.464	-1284.637	586.183
region[T.southeast]	-1035.2656	478.867	-2.162	0.031	-1974.684	-95.847
region[T.southwest]	-960.0814	478.106	-2.008	0.045	-1898.007	-22.156
age	256.7646	11.912	21.555	0.000	233.396	280.133
bmi	339.2504	28.611	11.857	0.000	283.122	395.379
children	474.8205	137.897	3.443	0.001	204.301	745.340

Omnibus:	299.816	Durbin-Watson:	2.089
Prob(Omnibus):	0.000	Jarque-Bera (JB):	716.552
Skew:	1.211	Prob(JB):	2.53e-156
Kurtosis:	5.646	Cond. No.	311.

Skewness exceeding 1 suggests a potential violation of the normality assumption. Upon backward testing and omitting "sex" from the model, our refined model appears as follows:

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.751
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	571.8
Date:	Wed, 11 Oct 2023	Prob (F-statistic):	0.00
Time:	21:29:39	Log-Likelihood:	-13538.
No. Observations:	1337	AIC:	2.709e+04
Df Residuals:	1329	BIC:	2.713e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.199e+04	979.209	-12.242	0.000	-1.39e+04	-1.01e+04
smoker[T.yes]	2.384e+04	412.038	57.847	0.000	2.3e+04	2.46e+04
region[T.northwest]	-348.2514	476.665	-0.731	0.465	-1283.349	586.846
region[T.southeast]	-1034.6268	478.711	-2.161	0.031	-1973.739	-95.515
region[T.southwest]	-959.4167	477.950	-2.007	0.045	-1897.036	-21.798
age	256.8751	11.905	21.577	0.000	233.520	280.230
bmi	338.7325	28.571	11.856	0.000	282.683	394.782
children	473.8629	137.831	3.438	0.001	203.473	744.253

Omnibus:	300.175	Durbin-Watson:	2.090
Prob(Omnibus):	0.000	Jarque-Bera (JB):	718.139
Skew:	1.211	Prob(JB):	1.14e-156
Kurtosis:	5.650	Cond. No.	308.

The p-values from the table indicate that every predictor is statistically significant when accounting for the other variables in the model at a significance level of 0.05. Skewness of the table(1.211) being bigger than 1 indicates that there is a skewness in the model which might hint at possible violation of the normality in error residuals. Curtosis value (5.650) of the table indicates that there's a higher probability of extreme values or outliers in the data.

The group evaluated all possible subsets of predictors after dropping the predictor ‘sex’. Based on adjusted R-square values, here are the top 5 models:

Top 5 models by Adjusted R2:

Number of predictors	Adj-R2	AIC	BIC	Predictors in the model
7	0.749407	27092.4	27134	age + bmi + children + smoker_yes + region_northwest + region_southeast + region_southwest
4	0.748775	27092.8	27118.7	age + bmi + children + smoker_yes
1	0.619453	27645	27655.4	smoker_yes
4	0.619132	27649.1	27675.1	smoker_yes + region_northwest + region_southeast + region_southwest
6	0.119101	28772.1	28808.5	age + bmi + children + region_northwest + region_southeast + region_southwest

According to the table, the full model—with predictors age, bmi, children, smoker, and region—exhibits the highest adjusted R-square. This implies it's the model best equipped to explain variance in the residuals. Switching focus to the AIC and BIC values, here are the top 5 models prioritized by the smallest values:

Top 5 models by BIC:

Number of predictors	Adj-R2	AIC	BIC	Predictors in the model
4	0.748775	27092.8	27118.7	age + bmi + children + smoker_yes
7	0.749407	27092.4	27134	age + bmi + children + smoker_yes + region_northwest + region_southeast + region_southwest
1	0.619453	27645	27655.4	smoker_yes
4	0.619132	27649.1	27675.1	smoker_yes + region_northwest + region_southeast + region_southwest
3	0.117703	28771.3	28792.1	age + bmi + children

Although the full model offers the lowest AIC value, the team has chosen to proceed with the model boasting the smallest BIC value. This decision is rooted in the fact that while both AIC and BIC penalize model complexity, BIC does so more strictly than AIC. Hence, choosing based on BIC emphasizes a simpler model, especially pertinent given our large sample size of 1337. Thus, our selected model based on the BIC criterion is: ‘**charges ~ age + bmi + children + smoker**’.

Moving forward, let's delve deeper into our chosen model. Here is the summary:

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.661
Model:	OLS	Adj. R-squared:	0.661
Method:	Least Squares	F-statistic:	868.1
Date:	Wed, 11 Oct 2023	Prob (F-statistic):	7.58e-313
Time:	23:29:34	Log-Likelihood:	-13743.
No. Observations:	1337	AIC:	2.749e+04
Df Residuals:	1333	BIC:	2.751e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4054.9905	1006.772	-4.028	0.000	-6030.022	-2079.959
smoker[T.yes]	2.357e+04	477.973	49.322	0.000	2.26e+04	2.45e+04
bmi	386.5125	31.640	12.216	0.000	324.443	448.582
children	594.1121	160.109	3.711	0.000	280.019	908.205

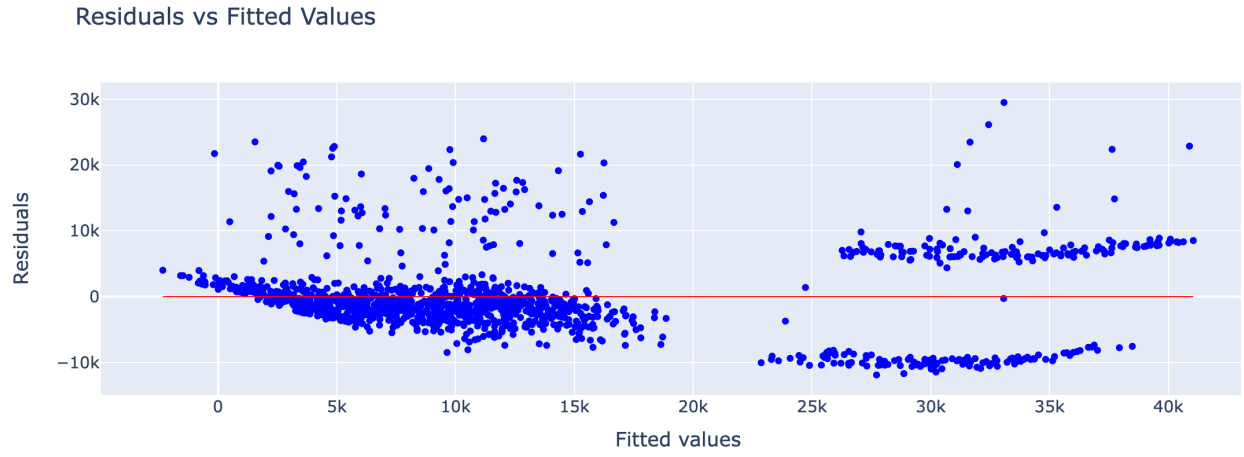
Omnibus:	159.291	Durbin-Watson:	2.043
Prob(Omnibus):	0.000	Jarque-Bera (JB):	249.272
Skew:	0.828	Prob(JB):	7.43e-55
Kurtosis:	4.317	Cond. No.	164.

The p-values from the table indicate that every predictor is statistically significant when accounting for the other variables in the model at a significance level of 0.05. Skewness of the table indicates that there is a moderate skewness in the model which might hint at possible violation of the normality in error residuals. Curtosis value (4.317) of the table indicates that there's a higher probability of extreme values or outliers in the data.

VIF score is calculated by fitting a regression model, using the the predictor variable that we are going to measure its VIF as the response variable and other variables as independent variables. According to the below VIF table all of the VIF scores are ignorable indicating that there does not exist multicollinearity in the model:

	VIF Factor	features
0	32.243043	Intercept
1	1.000768	smoker[T.yes]
2	1.014465	age
3	1.012213	bmi
4	1.001867	children

According to the residual plot, the variance in the errors seems not to be constant across all levels of the fitted values. Let's run a breusch pagan test and make sure:




```
{'LM Statistic (White Test)': 128.25331368027636, 'LM-Test p-value (White Test)': 4.654930198178968e-21}
```

```
{'LM Statistic(Breusch-Pagan)': 116.9799134263226, 'LM-Test p-value(Breusch-Pagan)': 2.3581868085083413e-24}
```

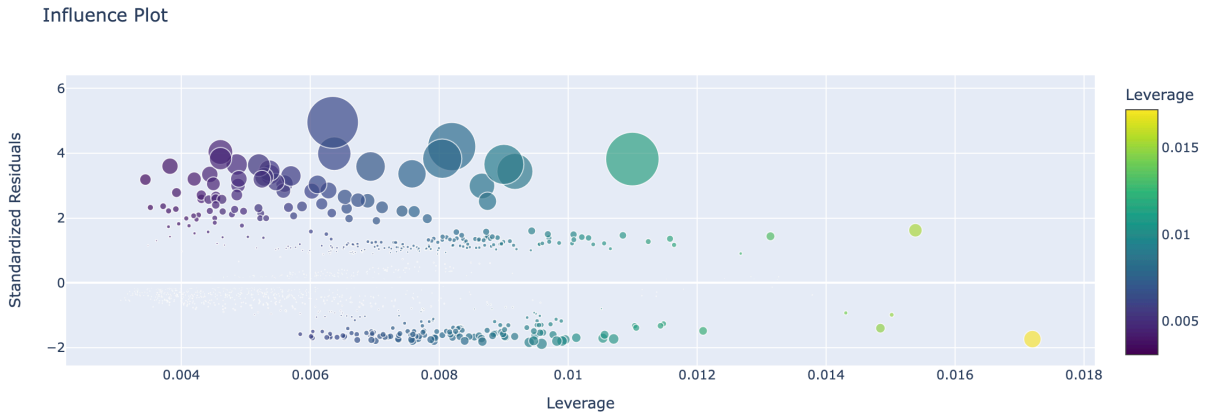
The p-value of Breusch-Pagan and White Test validates that that the variance in error is not constant across all levels of the predictors.

Given the detection of heteroscedasticity, possible solutions are :

- Deleting potential influential points
- Transformation on y
- Robust standard erros to correct the OLS
- Weighted least squares

The group first will try the natural log transformation and dropping some of the influential and outlier data points. Points are dropped from the dataset if they meet both of the following criteria:

- Their studentized residuals exceed the 97.5th percentile of a t-distribution, indicating they are unusually large and differ significantly from the expected residuals.
- Their Cook's distance values surpass a threshold of 4/n (n= 1337) , suggesting they have influence on the regression model's predictions.

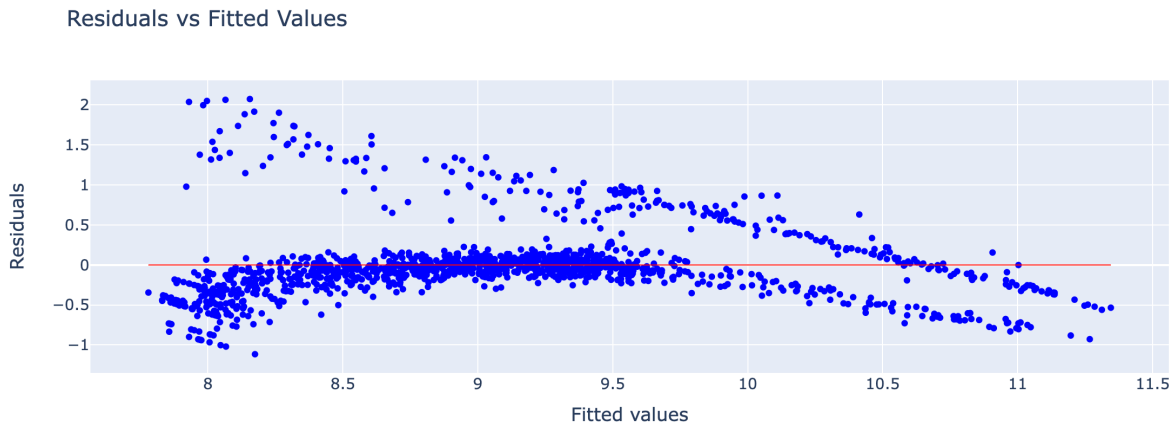


After dropping the common points in outliers and influentials the dropped model has 1288 data points. The p-values in the result of the breusch pagan and white test as also indicates that the variance in error is not constant across all levels of the predictors:

```
{'LM Statistic (Breusch-Pagan)': 395.1206991839326, 'LM-Test p-value (Breusch-Pagan)': 3.151540250123594e-84}
```

```
{'LM Statistic (White Test)': 424.52078373360456, 'LM-Test p-value (White Test)': 1.4668618366069762e-82}
```

Since it didn’t change again, being close to 0, we can deduce that the heteroscedasticity issue hasn't been resolved. Turning back to our original model and applying a logarithm to the 'charge' response variable, we have an updated residual plot:



Below tests are the breusch pagan and white test results after the log transformation of the response variable and exclusion of the potential influential points with the same procedure mentioned on the 16th page:

```
{'LM Statistic (White Test)': 140.86060857103698, 'LM-Test p-value (White Test)': 1.4149123668933737e-23}
```

```
{'LM Statistic (Breusch-Pagan)': 81.22999375614125, 'LM-Test p-value (Breusch-Pagan)': 9.558304014715811e-17}
```

Lastly, the Box-Cox method has been tried, however the test results indicate the same outcome:

```
{'LM Statistic (Breusch-Pagan)': 79.1568228725152, 'LM-Test p-value (Breusch-Pagan)': 7.225230277480339e-14}
```

The p-values of the tests indicates that heterosedasticity still exists and since it will cause inefficient estimates, invalid standard errors, biased R-squared, and compromised confidence intervals, the group will continue with the robust standard errors to make the variance in the errors constant:

```
{'LM Statistic (White Test)': 142.45829246865682, 'LM-Test p-value (White Test)': 2.7654493706241687e-17}
```

5.11 Regression Analysis & Model Selection (Robust Standard Errors)

For the robust model the group choose the HC0 model considering the sample size is bigger than a thousand. Here is the top 5 models considering the BIC:

Top 5 models by BIC:

Number of predictors	Adj-R2	AIC	BIC	Predictors in the model
4	0.748775	27092.8	27118.7	age + bmi + children + smoker_yes
7	0.749407	27092.4	27134	age + bmi + children + smoker_yes + region_northwest + region_southeast + region_southwest
1	0.619453	27645	27655.4	smoker_yes
4	0.619132	27649.1	27675.1	smoker_yes + region_northwest + region_southeast + region_southwest
3	0.117703	28771.3	28792.1	age + bmi + children

The group chooses to continue with the same model ‘charges ~ age + bmi + children + smoker’ eventough the full model’s adjusted r-square value is slightly higher than because of the same reason in the model selection of mlr regression without robust standard errors.

Here is the summary of the robust model:

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.750
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	601.1
Date:	Thu, 12 Oct 2023	Prob (F-statistic):	1.96e-296
Time:	10:35:30	Log-Likelihood:	-13541.
No. Observations:	1337	AIC:	2.709e+04
Df Residuals:	1332	BIC:	2.712e+04
Df Model:	4		
Covariance Type:	HC0		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.21e+04	986.186	-12.268	0.000	-1.4e+04	-1.02e+04
smoker[T.yes]	2.381e+04	570.536	41.733	0.000	2.27e+04	2.49e+04
age	257.7728	11.819	21.811	0.000	234.609	280.937
bmi	321.8708	29.887	10.770	0.000	263.294	380.448
children	472.9751	130.508	3.624	0.000	217.183	728.767

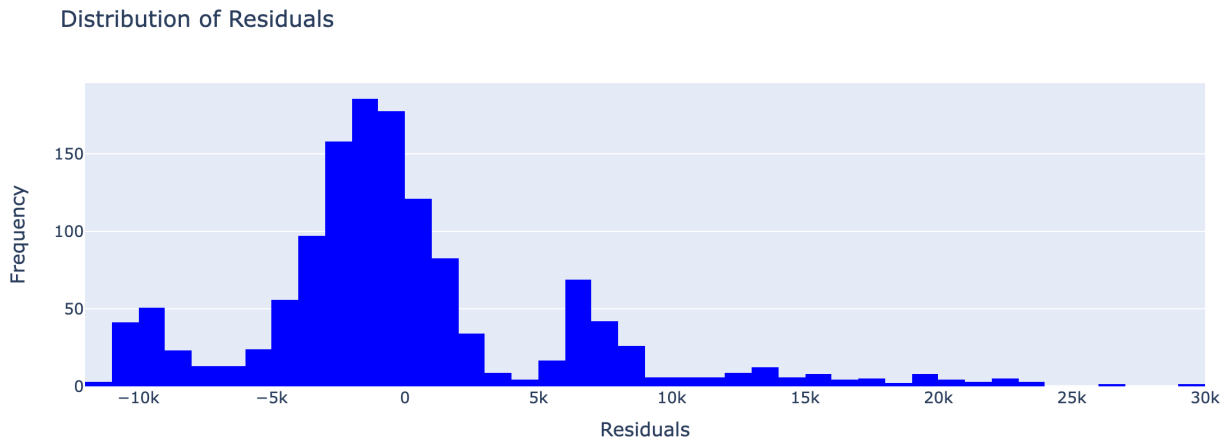
Omnibus:	300.944	Durbin-Watson:	2.088
Prob(Omnibus):	0.000	Jarque-Bera (JB):	719.880
Skew:	1.215	Prob(JB):	4.79e-157
Kurtosis:	5.650	Cond. No.	292.

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)

The p-values from the table indicate that every predictor is statistically significant when accounting for the other variables in the model at a significance level of 0.05. Skewness (1.215) of the table indicates that there is an issue of non-symetrical shape in the model which might hint at possible

violation of the normality in error residuals. Curtosis value (5.650) of the table indicates that there's a higher probability of extreme values or outliers in the data.

Examining the “residual distribution graph” below also indicates the non-normality of the residuals:



After removal of the potential influential points and outliers as discussed on the page 16, the JB Statistic score decreased significantly, however the test result still suggests that there is a violation of the assumtpion of normality:

{'JB Statistic': 100.3399814204597, 'JB p-value': 1.6272335032978037e-22, 'Skewness': 0.5055671893784343, 'Kurtosis': 3.920487024833667}

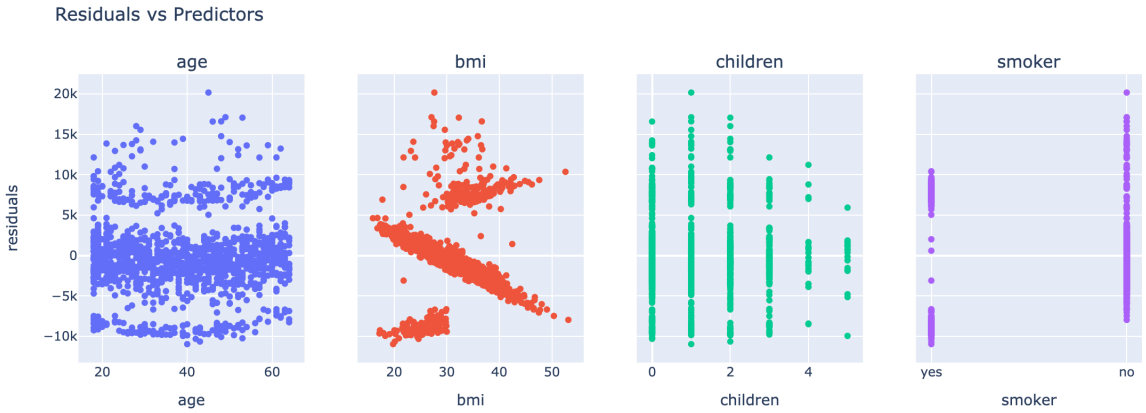
Comparing the JB-stat value above with Chi-square distribution value with degrees of freedom 2 (100.34 > 5.991) and the p-value in JB Statistic (p-val <0.05) suggests that the data does not meet the normality assumption at a significance level of 0.05.

After the log transformation of the response variable(**’ log_charges ~ age + bmi + children + smoker’).** The JB statistic score increases indicating no improvement for the violation of the normality:

{'JB Statistic': 1170.8161083155946, 'JB p-value': 5.761192909934833e-255, 'Skewness': 1.2639003699759117, 'Kurtosis': 6.92768031315826}

After the lack-of-fit F-test result indicating that there is also possible non linearity at a significance level 0.05 and examining the residual vs fitted graph of the predictor variables, the group tried the log transformation on bmi predictor aiming to solve the issue: however the normality issue got worse, and there has not been any significant improvement in the lack of fit F-test score observed so the group decides to turn back to the former model:

```
F-statistic: 502580.0490
p-value: 0.0000
The model seems to suffer from a lack of fit.
```



- The lack-of-fit F-test result after the log transformation on bmi:

```
F-statistic: 501656.1456
p-value: 0.0000
The model seems to suffer from a lack of fit.
```

The summary of the model after the log transformation on bmi predictor:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            log_charges   R-squared:                       0.824
Model:                            OLS   Adj. R-squared:                  0.823
Method:                 Least Squares   F-statistic:                     1132.
Date:                Thu, 12 Oct 2023   Prob (F-statistic):               0.00
Time:                        11:21:48   Log-Likelihood:                -575.48
No. Observations:                1288   AIC:                             1161.
Df Residuals:                    1283   BIC:                             1187.
Df Model:                           4
Covariance Type:                  HC0
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.9482      0.166     35.882      0.000       5.623       6.273
smoker[T.yes]  1.5778      0.031     50.199      0.000       1.516       1.639
age            0.0356      0.001     39.796      0.000       0.034       0.037
log_bmi        0.3772      0.054      7.031      0.000       0.272       0.482
children       0.0950      0.008     12.214      0.000       0.080       0.110
==============================================================================
Omnibus:                      337.561   Durbin-Watson:                   2.029
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1154.358
Skew:                           1.260   Prob(JB):                    2.16e-251
Kurtosis:                       6.893   Cond. No.                         734.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
```

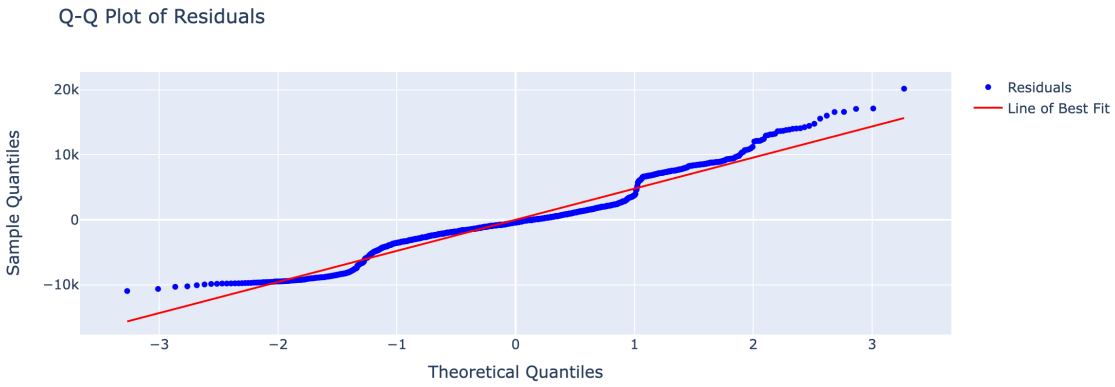
6.1 Final Model & Diagnosis

The group chooses the model with the predictors ‘charges ~ age + bmi + children + smoker’ with robust standard errors. The summary of the model:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.819
Model:                            OLS   Adj. R-squared:                  0.818
Method:                 Least Squares   F-statistic:                     786.9
Date:                Thu, 12 Oct 2023   Prob (F-statistic):               0.00
Time:                        11:15:04   Log-Likelihood:                -12771.
No. Observations:                1288   AIC:                         2.555e+04
Df Residuals:                    1283   BIC:                         2.558e+04
Df Model:                           4
Covariance Type:                  HC0
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -1.27e+04    809.453    -15.694      0.000   -1.43e+04   -1.11e+04
smoker[T.yes]  2.378e+04    521.986     45.559      0.000    2.28e+04    2.48e+04
age            256.3522      9.587     26.739      0.000     237.562     275.143
bmi            324.0605     25.519     12.699      0.000     274.043     374.078
children       392.4921    105.452      3.722      0.000     185.810     599.174
==============================================================================
Omnibus:                       73.359   Durbin-Watson:                   2.041
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              100.340
Skew:                           0.506   Prob(JB):                     1.63e-22
Kurtosis:                       3.920   Cond. No.                         292.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
```

Even tough the violation of the non linearity assumption(can be seen from the inflated Jarque-Bera test stat value) decreasing the reliability of the hypothesis tests and confidence intervals, considering the large sample size and the influential points have not been deducted (meaning that the double exclusion of the influential points) from the model to not to lose information this model is choosen. As also the diagonal line in QQ plot is not so clear, indicating the violation of the assumption of normality in the residuals:



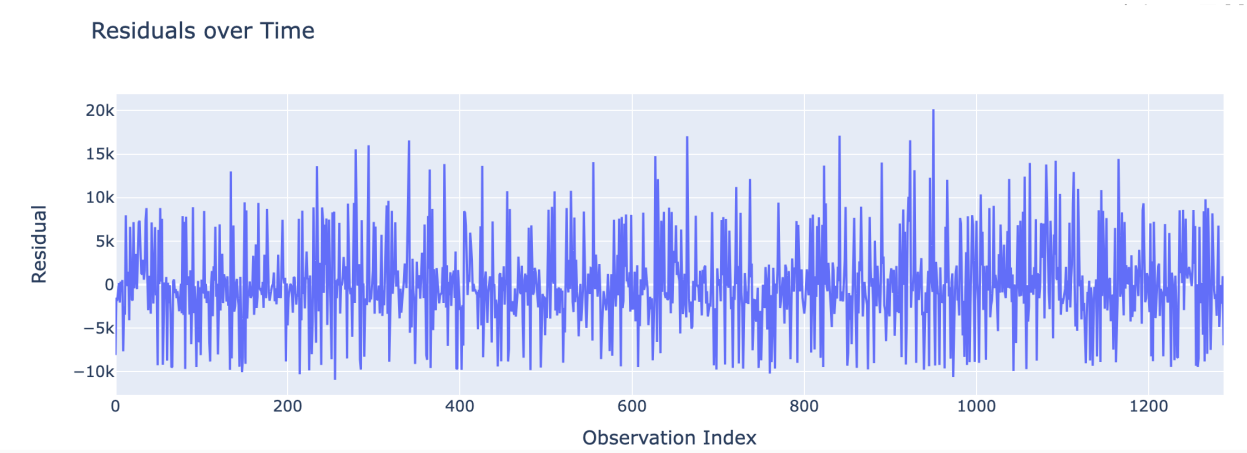
The model has non linearity issue causing Gaus Markov theorem to fail and making the BLUE invalid (not the best unbiased estimators anymore). The result of the lack-of-fit F-test indicating the non-linearity in the model:

```
F-statistic: 501656.1456
p-value: 0.0000
The model seems to suffer from a lack of fit.
```

The VIF scores of the potential influential points excluded model indicating no multicollinearity in the model can be seen below:

	VIF Factor	features
0	32.335757	Intercept
1	1.001134	smoker[T.yes]
2	1.014755	age
3	1.012084	bmi
4	1.002354	children

Autocorelation plot suggest that there may be a possible autocorrelation, this is because there are repeating irregular cycles up and down with different time cycles. This could hint at error residuals being not independent. Their order or the index numbers might have effect on the response variable, however since the pattern is not so obvious it is still hard to state that there is autocorrelation for sure.



7.1 Summary of the Results

The primary objective was to identify the best fitting model for predicting "charges". Initially, all potential predictors were considered, but after backward testing, "sex" was excluded for not being statistically significant. Despite the full model showcasing a high adjusted R-square, the group considered the model complexity using AIC and BIC values, leaning more towards BIC due to its stricter penalization of model complexity. This meticulous approach led to the selection of the model 'charges ~ age + bmi + children + smoker'. However, further diagnostic tests highlighted issues like heteroscedasticity and violations of the normality assumption. Various attempts, including natural log transformations and outlier removals, were made to rectify these, but they remained persistent.

Considering the heteroscedasticity detected, the team pivoted to using robust standard errors. The HC0 model was selected given the large sample size, and though the full model's adjusted R-square was marginally higher, the decision was to continue with the simpler model, 'charges ~ age + bmi + children + smoker', due to its lower BIC. Notably, even after data transformations, the normality assumption remained a concern. Additionally, tests indicated

potential non-linearity in the model, with attempts to log-transform predictors, like bmi, not delivering any substantial improvement.

Ultimately, the chosen model was 'charges ~ age + bmi + children + smoker' with robust standard errors. Though this model still displayed potential violations of some linear regression assumptions, it was deemed the best fit given the large sample size and intent to retain all data points. Indicators like the Jarque-Bera test, lack of fit F test and QQ plots reiterated the concern about non-normality and non-linearity. However, with the VIF scores being acceptable, multicollinearity was not an issue. The presence of potential autocorrelation was acknowledged but deemed not definitive. In essence, while the model may not be perfect, it's the best compromise given the available data and constraints.