Linear Regression Final Project Write-up

Sophie Wang, Jordan Uyeki

Description of the dataset

Data source: https://www.kaggle.com/mirichoi0218/insurance

The medical cost dataset is from Kaggle and consists of four numeric variables:

1. Age
2. BMI
3. Number of children
4. Medical charges

And three categorical variables:

1. Sex
2. Smoker
3. Region

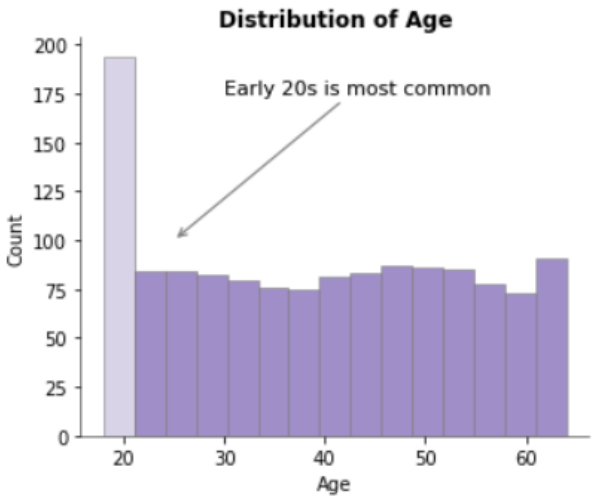
None of them are encoded in the dummy variables. The dataset contains 1338 rows in total.

Statement of research problem & Summary of methods

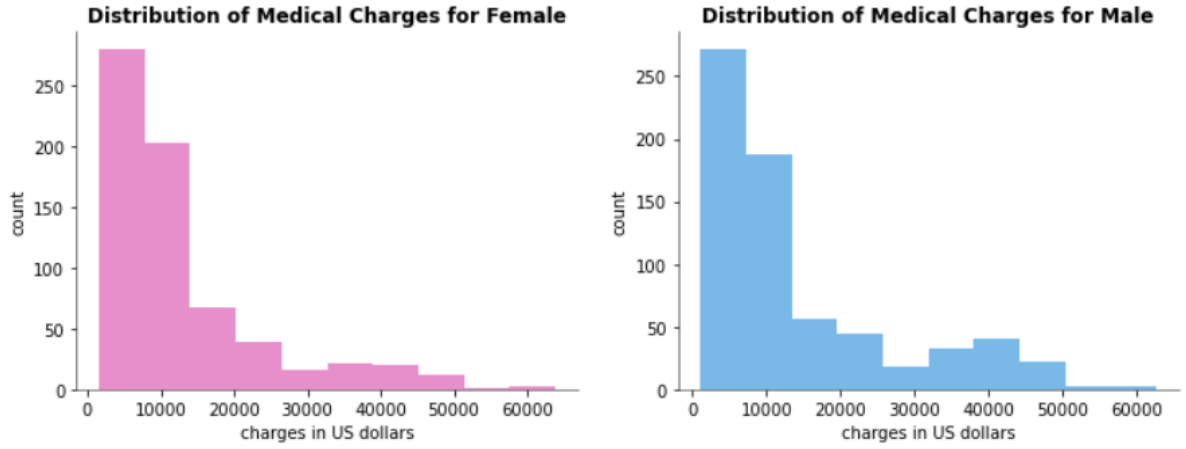
Our research problem is to build a model to predict post-insurance medical charges from the age, BMI, number of children, sex, smoker status, and geographic region of residence of individuals in the dataset. To address the problem, leveraged the best subset algorithm to identify the most significant multiple linear regression model.

EDA

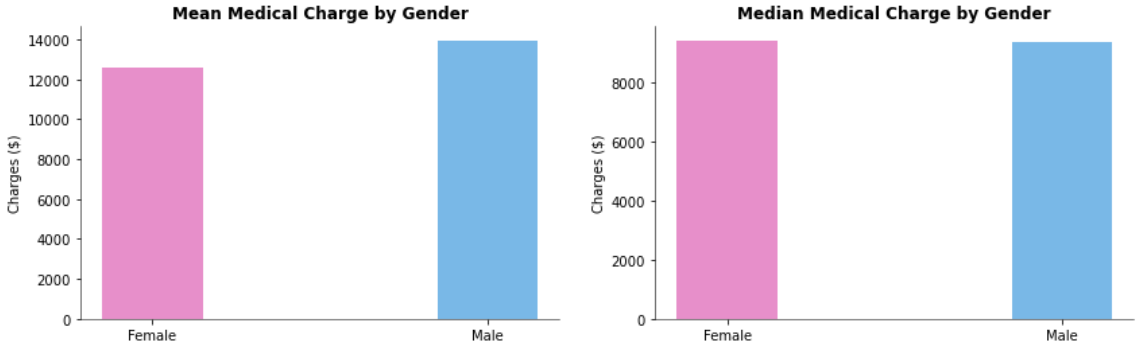
First and foremost, we plotted several graphs to explore the relationships between different variables. We first plotted **the histogram of age** as shown below. From this plot, the dataset shows most people (around 200) are in their early 20s. Besides this young age group, each of the other age groups has about 80 to 90 people.



Secondly, we explored whether gender causes any disparity in medical costs. The dataset has roughly 50% females and 50% males records. We first plotted the general **distribution of medical charges by gender**. The shapes of the distribution for female and male look roughly the same. Most people of both gender have medical charges below $10K. The count of people decreases as the medical cost goes up in the plot. However, the histogram shows a slight cluster (increase) for counts when medical cost is around $40K, and males have a larger cluster than females in that region.

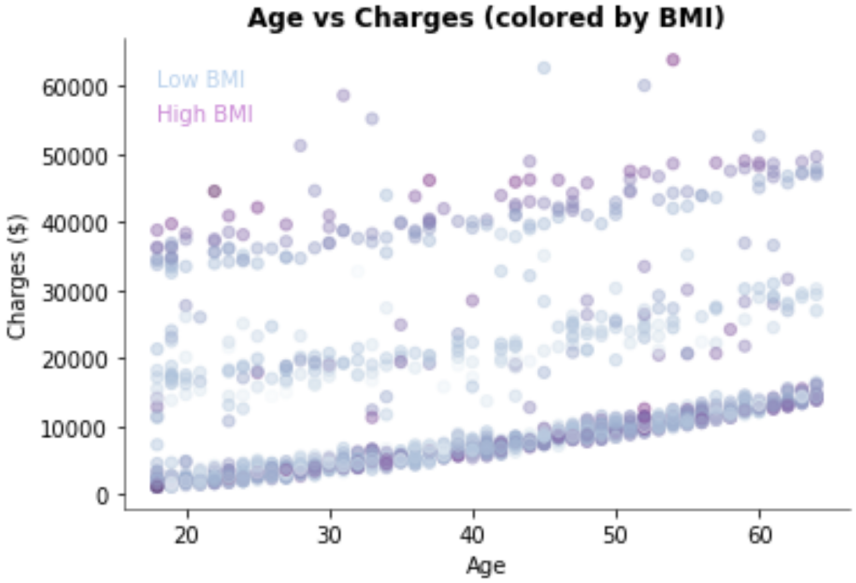


Thirdly, we explored the **average and median of medical charges by gender** and plotted the graph below:



From the plot, we can see males have higher average charges than females, although the median charges for both genders are roughly the same. This can also be confirmed by the exact number extracted from the data set: average insurance charge is $12569.58 for females and $13956.75 for males. Median insurance charge is $9412.96 for females and $9369.62 for males. The disparity between the means implies that males are more likely to incur very high (>$30k) medical costs than females.

Next, we plotted the relationship between **three variables: age, BMI, and charges**. This scatterplot below shows the relationship between age and medical charges, with the color gradient of each data point determined by BMI value. Light blue indicates less BMI and dark purple indicates higher BMI.

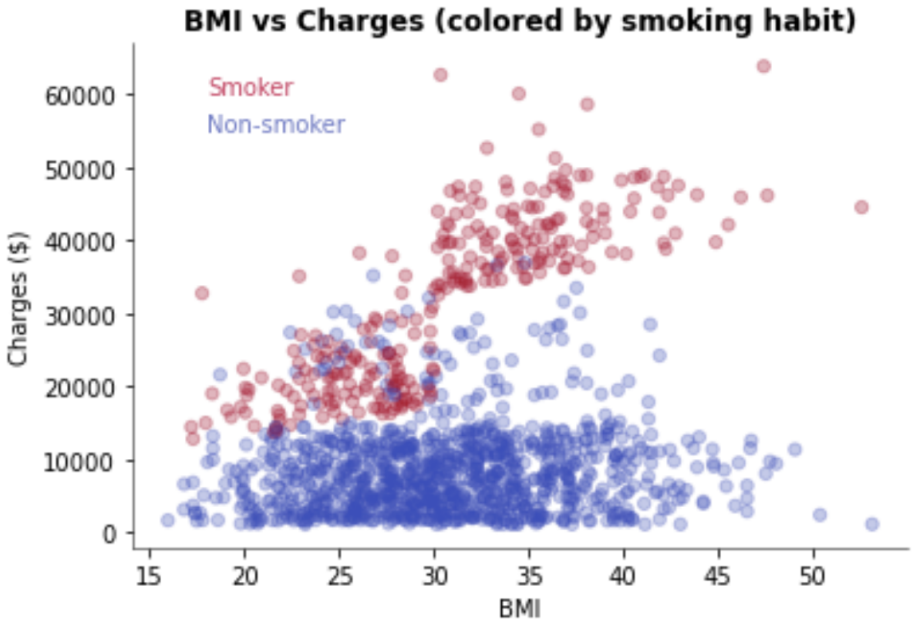


From the plot, we can see the age and charges is not a simple linear relationship. There are three obvious clusters in the plot. The top cluster (higher medical costs) shows more purple points than the two lower clusters, which indicates that people with higher BMI tend to have higher medical costs. In addition, all three clusters show a

slightly upward trend. Holding all other factors constant, an increase in age also increases medical cost.

Now that we know BMI plays a role in influencing medical charges, we are also curious about how smoking habit affects medical charges since smoking is commonly known to affect health. Out of 1338 records in the dataset, 1065 records are from non-smokers and only 274 records are from smokers. Thus, most people in this dataset do not smoke. The plot below illustrates **how medical charges can be affected by both BMI and smoking**.

By using color to differentiate smoker and non-smoker, we can see two obvious clusters in the scatterplot above. For smokers (indicated by red), the charge increases significantly as BMI increases. By contrast, the medical charge for non-smokers (indicated by blue) does not significantly change as BMI increases. This implies that smoking puts a significant health risk on people. Even for people with lower BMI, smoking increases medical costs.



Regression analysis and model diagnosis

Before deploying the linear regression model, we need to check if there is any problem in the data structure and also if the model assumptions are met.

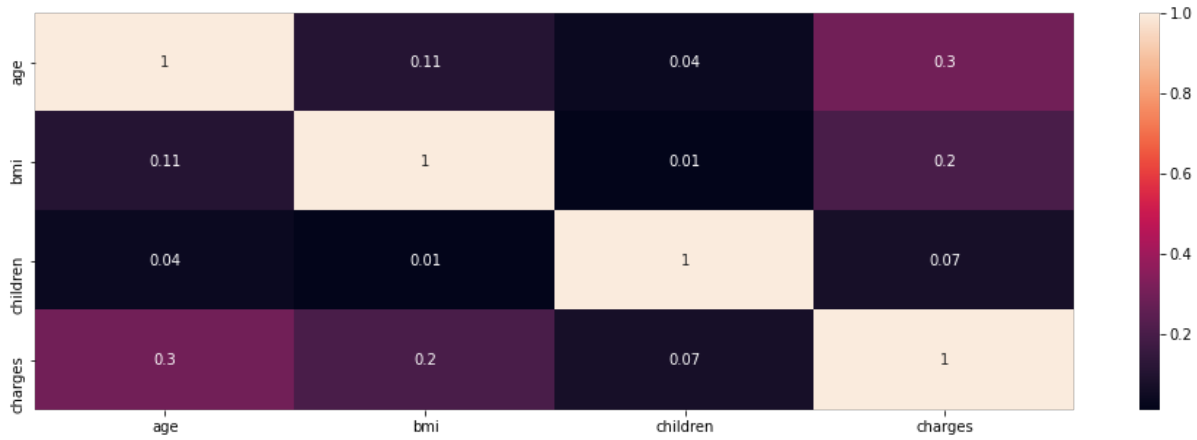
Data structure problems can be multicollinearity and influential points.

Model assumption violations can be: heteroscedasticity , non-normality residuals, y and x are not linearly related. Having either data structure problems or model assumption violation can make linear regression models not reliable.

1: Multicollinearity

Before fitting any model, we checked whether multicollinearity exists among predictors by using heatmap as well as VIF scores.

From the **correlation heatmap** below, we can see that most of the variables are not strongly correlated, which is good for modeling using linear regression models.



Using the numeric method **VIF score** can confirm what the correlation heatmap indicates. From the VIF table below, we can conclude that none of the predictor variables is causing multicollinearity. As we can see from the VIF table, all of the predictors have a VIF score less than 10. VIF score for a predictor variable measures how much the variance is inflated in the coefficient estimates caused by that predictor variable. When the VIF score for the predictor variable is between 1 and 10, then that variable is not causing serious multicollinearity. but when the VIF score is larger than 10, then that variable is causing serious multicollinearity problems.

	VIF Factor	features
0	35.527488	Intercept
1	1.008900	C(sex)[T.male]
2	1.012074	C(smoker)[T.yes]
3	1.518823	C(region)[T.northwest]
4	1.652230	C(region)[T.southeast]
5	1.529411	C(region)[T.southwest]
6	1.016822	age
7	1.106630	bmi
8	1.004011	children

After checking multicollinearity for all predictors, we **fitted our initial model** which includes all the predictors. From the summary table of the initial model below, we can see that the Adj. R-squared is 0.749. Sex is not significant with t-test p-value of 0.693. Other predictors all seem to be significant from the results of t-test and p-value.

OLS Regression Results

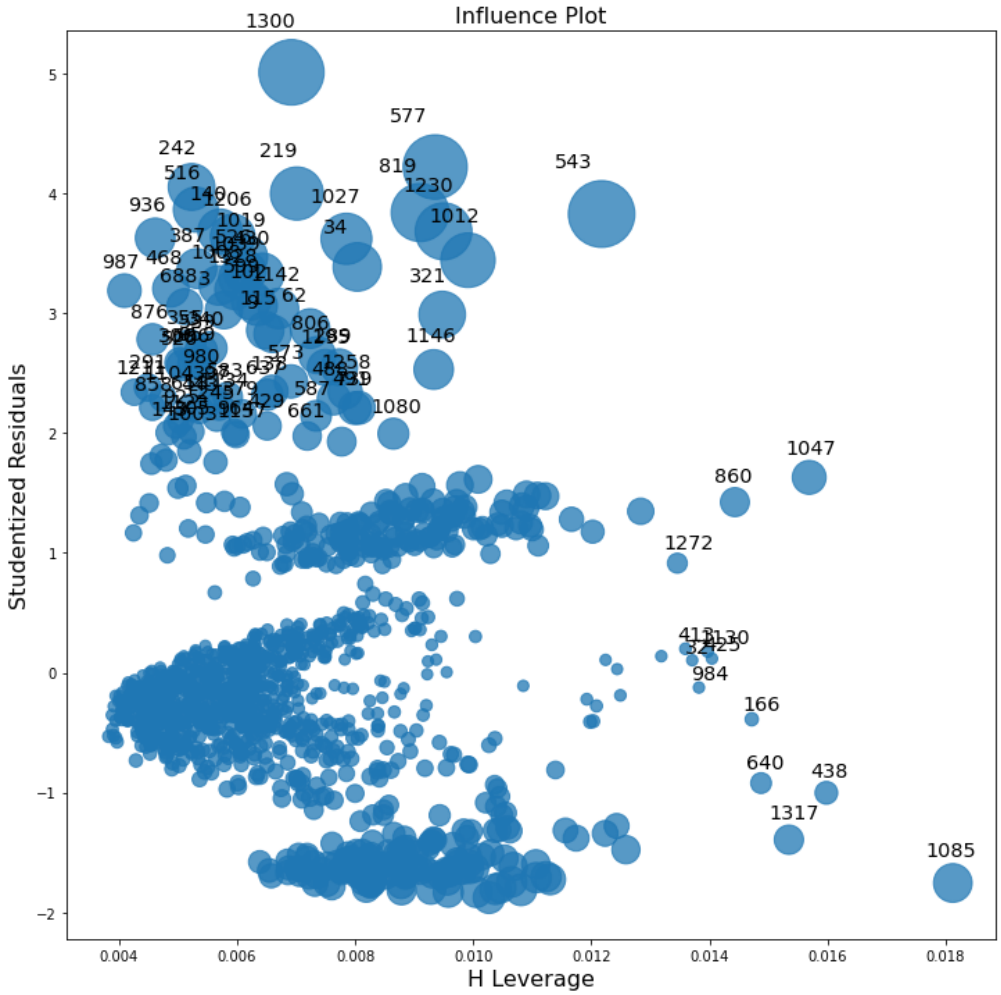
Dep. Variable:	charges	**R-squared:**	0.751
Model:	OLS	**Adj. R-squared:**	0.749
Method:	Least Squares	**F-statistic:**	500.8
Date:	Sun, 29 Nov 2020	**Prob (F-statistic):**	0.00
Time:	01:13:32	**Log-Likelihood:**	-13548.
No. Observations:	1338	**AIC:**	2.711e+04
Df Residuals:	1329	**BIC:**	2.716e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.194e+04	987.819	-12.086	0.000	-1.39e+04	-1e+04
C(sex)[T.male]	-131.3144	332.945	-0.394	0.693	-784.470	521.842
C(smoker)[T.yes]	2.385e+04	413.153	57.723	0.000	2.3e+04	2.47e+04
C(region)[T.northwest]	-352.9639	476.276	-0.741	0.459	-1287.298	581.370
C(region)[T.southeast]	-1035.0220	478.692	-2.162	0.031	-1974.097	-95.947
C(region)[T.southwest]	-960.0510	477.933	-2.009	0.045	-1897.636	-22.466
age	256.8564	11.899	21.587	0.000	233.514	280.199
bmi	339.1935	28.599	11.860	0.000	283.088	395.298
children	475.5005	137.804	3.451	0.001	205.163	745.838

Omnibus:	300.366	**Durbin-Watson:**	2.088
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	718.887
Skew:	1.211	**Prob(JB):**	7.86e-157
Kurtosis:	5.651	**Cond. No.**	311.

2: Influential points

Next we detected influential points using **leverage vs. external studentized residuals plots**. The plot as shown below shows that observation numbers 1047, 438, 1317 and 1085 are high leverage points, the datapoint with index 543 has the largest Cook's distance, observation number 1300 has the largest studentized residual.



Out of 1338 observations, there are 74 observations out of 1338 observations identified as influential points by **externally studentized residual** method and 87 observations identified as influential points using **Cook's distance** method. 60 observations are identified as influential points by both methods so we dropped those 60 observations from the original data set and fitted the model again. The summary result of the new model below shows that the adjusted R-squared increased to from the original 74.9% to 83.4%. However, there is a caveat of dropping influential points: we should report the both models with and without influential points included. Without consulting with subject-matter experts, we cannot be certain if dropping influential points will cause serious information loss.

OLS Regression Results

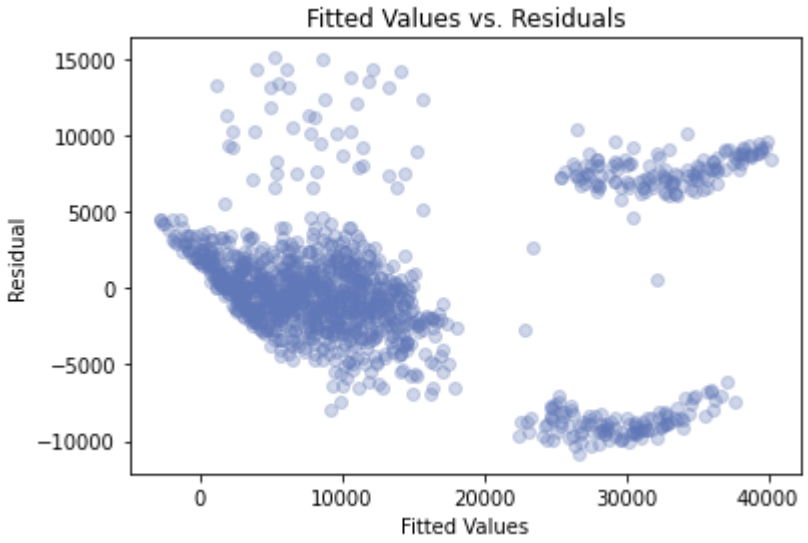
Dep. Variable:	charges	**R-squared:**	0.835
Model:	OLS	**Adj. R-squared:**	0.834
Method:	Least Squares	**F-statistic:**	801.8
Date:	Sun, 29 Nov 2020	**Prob (F-statistic):**	0.00
Time:	01:13:36	**Log-Likelihood:**	-12613.
No. Observations:	1278	**AIC:**	2.524e+04
Df Residuals:	1269	**BIC:**	2.529e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.299e+04	783.185	-16.582	0.000	-1.45e+04	-1.15e+04
C(sex)[T.male]	97.4030	263.872	0.369	0.712	-420.271	615.076
C(smoker)[T.yes]	2.393e+04	326.516	73.299	0.000	2.33e+04	2.46e+04
C(region)[T.northwest]	-622.9066	378.497	-1.646	0.100	-1365.455	119.642
C(region)[T.southeast]	-1055.7324	380.073	-2.778	0.006	-1801.372	-310.093
C(region)[T.southwest]	-770.7206	377.542	-2.041	0.041	-1511.395	-30.046
age	253.6273	9.459	26.812	0.000	235.069	272.185
bmi	348.9254	22.682	15.383	0.000	304.426	393.424
children	410.8806	108.843	3.775	0.000	197.349	624.412

Omnibus:	24.910	**Durbin-Watson:**	2.051
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	29.470
Skew:	0.268	**Prob(JB):**	3.99e-07
Kurtosis:	3.515	**Cond. No.**	311.

3: Heteroscedasticity

One of the most important assumptions of linear regression is homoscedasticity (residuals have constant variance). **Residual vs. fitted value plot** below shows a strong pattern existing. To put it in other words, the bandwidth of the plot around the 0 horizontal line (residual =0) varies a lot for different data points. The inconsistent bandwidth suggests serious heteroscedasticity, meaning the residuals do not have constant variance.



Additionally, by applying the numeric method **Breussch-Pagan test**, we rejected the null hypothesis of the test due to p-value less than 0.05. Thus we confirmed our findings from the Residual vs. fitted value plot above: heteroscedasticity exists.

Trying to alleviate the heteroscedasticity issue, we applied **log transformation** on response variable medical charges and ran the **Breusch-Pagan test** again. The p-value (3.1104422551941013e-24) from the test is still less than 0.05 and heteroscedasticity persists. We choose to proceed as it is for now and we will discuss the impact of heteroscedasticity on the model inference later.

As for now, our updated mode summary is as below (log transformed Y):

OLS Regression Results

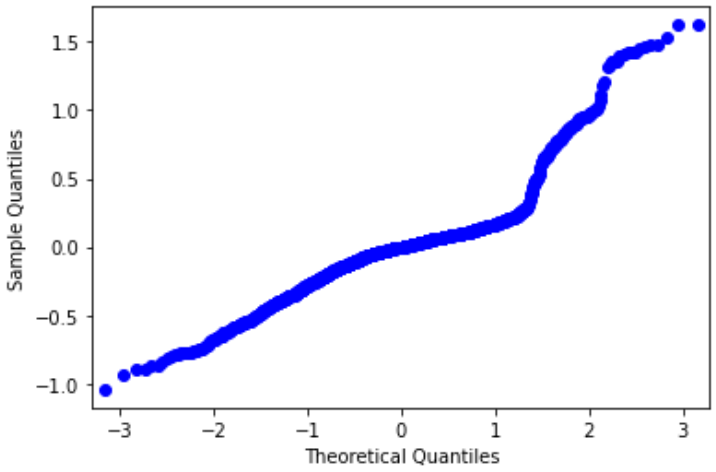
Dep. Variable:	log_charges	**R-squared:**	0.846
Model:	OLS	**Adj. R-squared:**	0.845
Method:	Least Squares	**F-statistic:**	874.2
Date:	Sun, 29 Nov 2020	**Prob (F-statistic):**	0.00
Time:	01:13:37	**Log-Likelihood:**	-483.90
No. Observations:	1278	**AIC:**	985.8
Df Residuals:	1269	**BIC:**	1032.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.8629	0.059	115.967	0.000	6.747	6.979
C(sex)[T.male]	-0.0653	0.020	-3.277	0.001	-0.104	-0.026
C(smoker)[T.yes]	1.6018	0.025	64.924	0.000	1.553	1.650
C(region)[T.northwest]	-0.0811	0.029	-2.835	0.005	-0.137	-0.025
C(region)[T.southeast]	-0.1612	0.029	-5.614	0.000	-0.218	-0.105
C(region)[T.southwest]	-0.1144	0.029	-4.011	0.000	-0.170	-0.058
age	0.0356	0.001	49.786	0.000	0.034	0.037
bmi	0.0155	0.002	9.042	0.000	0.012	0.019
children	0.0974	0.008	11.846	0.000	0.081	0.114

Omnibus:	322.059	**Durbin-Watson:**	2.052
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1152.124
Skew:	1.194	**Prob(JB):**	6.60e-251
Kurtosis:	6.992	**Cond. No.**	311.

4: Normality assumption

We used the **QQ plot** shown below to check the distribution of the residuals of the model. QQ plot shows the relationship between expected residual values versus the sample residual values. If the residuals of the model are normally distributed, the points on the plot should roughly form a straight diagonal line. The plot below does not show a straight diagonal line, meaning residuals might violate the normality assumption.



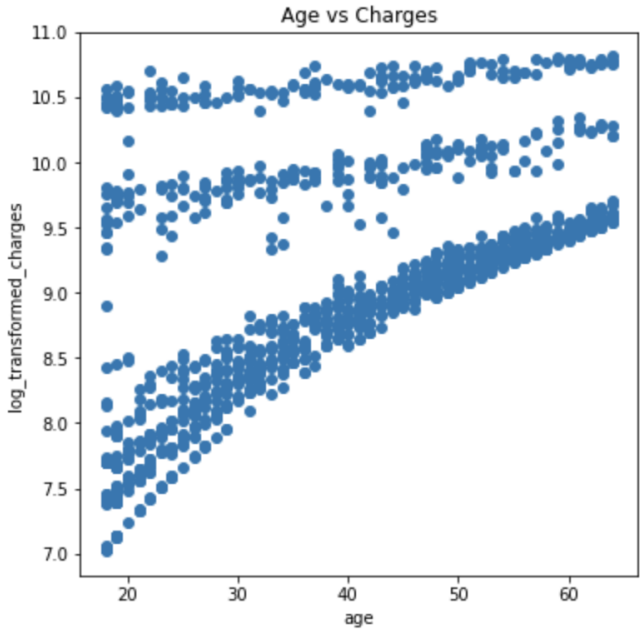
From the latest model summary table in the “heteroscedasticity” section above(with log transformed Y), we can see that the **JB test** statistics is 1152.124 and the p-value for JB test is 6.60e-251. We reject the null hypothesis of JB test (residuals of the model follow normal distribution). So the conclusion is that the residuals of the model do not follow normal distribution, which matches what QQ plot indicates.

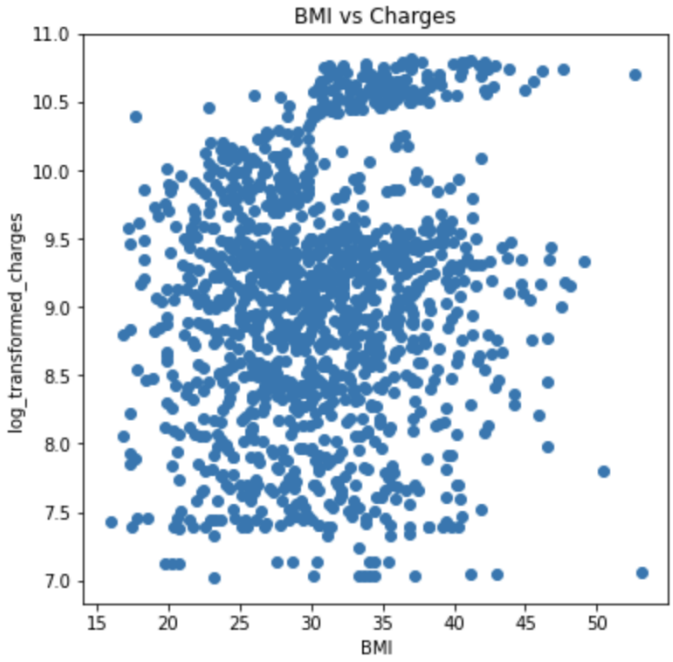
However, non-normality does not cause serious problems in our model fitting. Since our sample size is very large (definitely larger than 30), Central Limit Theorem makes sure that the coefficient estimator follows approximately normal distribution (so that we can still make inference on the coefficient estimator).

5: Linearity between response and predictor variables

We checked linearity relationship between numeric predictors and response variables medical charges (after natural log transformation).

Below are the scatterplots showing the relationship between age and log transformed medical charges and the relationship between BMI and log transformed medical charges. Both plots show non-linear relationships between individual predictors and the response variable charges. Having non-linearity problems could cause large prediction bias in linear regression models. The solution could be 1) natural log transform variables 2) use the non-linear approach and add polynomial terms. We should be very cautious when adding polynomial terms since it adds more parameters in the model and may introduce multicollinearity. Thus the model could lose inference power. We choose to not introduce polynomial terms to the model.





Model selection

To identify the best regression model for the prediction of an individual's medical cost owed after insurance, we used the Best Subset method. We chose this method because it is the most thorough and was computationally feasible since we only had six potential predictor variables. This means that the Best Subset method algorithm only

had to compare 63 candidate models ($2^k - 1$models, where there are k possible predictors).

To select the subset of best candidate models, we used the Mallow's Cp and Adjusted R^2criterion. While being cautious of overfitting, we aimed to minimize the Mallow's Cp as it is a measure of how much error is left unexplained by the partial model in comparison to the full model. We also looked to maximize the Adjusted $R^2$to find the model(s) that best capture the variance in post-insurance medical cost (our response variable). Our top 5 candidate models based on the Mallow's Cp and Adjusted R^2criterion can be summarized as follows.

# of Features	Mallow's Cp	Adj. R^2	AIC	BIC	Features
6	9.000	0.845	985.791	1032.168	age, sex, bmi, children, smoker, region
5	17.739	0.844	994.560	1035.785	age, bmi, children, smoker, region
5	36.335	0.842	1012.929	1043.847	age, sex, bmi, children, smoker
4	44.756	0.841	1021.114	1046.880	age, bmi, children, smoker
5	88.759	0.836	1063.586	1104.811	age, sex, children, smoker, region

To choose our final model from our best candidate models, we used the Aikake's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both of these criterions measure the amount of information that is lost by the model. Based on this, the full model is the best model as it has the lowest AIC and BIC values out of all the top candidate models.

Final model of choice and interpretation

The final model that we chose, based on the Mallow's Cp, Adjusted R^2, AIC, and BIC, is the full model. This model predicts the log transformed post-insurance medical costs from the age, sex, BMI, number of children, smoking status, and geographic region of the individual patient.

The Adjusted $R^2$and F statistic values are indicative of the overall significance and predictive power of our final model. The final model has an Adjusted R^2value of 0.845, which means that this model is able to capture 84.5% of the variation in our response variable. Additionally, the final model has a F statistic of 874.2 with a p-value of 0.000, which means that the overall model is highly significant. Besides looking at the overall model, we can also look at the effect of the individual model predictors' on post-insurance medical cost.

The first predictor, age, represents the age of the primary beneficiary. The coefficient of age is 0.0356 which means that for every one increase in age, we would expect there to be a $0.04 increase in the post-insurance medical cost. The 95% confidence interval for

age ranges from 0.034 to 0.037. Since this entire interval is positive, we can be 95% certain that there is a positive relationship between age and post-insurance medical cost. Because age has a t-statistic of 49.786 with a p-value of 0.000, we can confirm that age is a significant predictor in this model. The older an individual gets, the higher their post-insurance medical cost would be expected to be.

The next predictor, sex, represents the gender of the primary beneficiary. Using female as the reference level, the coefficient of the male level for the sex predictor is -0.0653. This means that if all of the other predictors besides sex were held constant, we would expect the post-insurance medical cost for males to be $0.07 less than the post-insurance medical cost for females. The 95% confidence interval for sex ranges from -0.104 to -0.026. Since this entire interval is negative, we can be 95% certain that there is a difference in post-insurance medical costs between males and females. Becaue sex has a t-statistic of -3.277 with a p-value of 0.001, we can confirm that sex is a significant predictor in this model.

The next predictor, BMI, represents the body mass index of an individual. The coefficient of BMI is 0.0155 which means that for every one unit increase in age, we would expect there to be a $0.02 increase in post-insurance medical cost. The 95% confidence interval for BMI ranges from 0.012 to 0.019. Since this entire interval is positive, we can be 95% certain that there is a positive relationship between BMI and post-insurance medical cost. Because BMI has a t-statistic of 9.042 with a p-value of 0.0000, we can confirm that BMI is a significant predictor in this model. The higher an individual's BMI is, the higher their post-insurance cost would be expected to be.

The next predictor, children, represents the number of children or dependents that are covered under an individual's insurance plan. The coefficient of children is 0.0974 which means that for every one unit increase in children, we would expect there to be a $0.10 increase in post-insurance medical cost. The 95% confidence interval for children ranges from 0.081 to 0.114. Since this entire interval is positive, we can be 95% certain that there is a positive relationship between children and post-insurance medical cost. Because children has a t-statistic of 11.846 with a p-value of 0.0000, we can confirm that children is a significant predictor in this model. The more children or dependents an individual has on their insurance plan, the higher their post-insurance cost would be expected to be.

The next predictor, smoker, represents whether or not the individual smokes. Using non-smokers as the reference level, the coefficient for the smokers level of the smoker variable is 1.6018. This means that if all of the other predictors besides smoker were held constant, we would expect the post-insurance medical cost for smokers to be

$1.60 more than the post-insurance medical cost for non-smokers. The 95% confidence interval for smoker ranges from 1.553 to 1.650. Since this entire interval is positive, we can be 95% certain that there is a difference in post-insurance medical costs between smokers and non-smokers. Because smoker has a t-statistic of 64.924 with a p-value of 0.0000, we can confirm that smoker is a significant predictor in the model.

The last predictor, region, represents the beneficiary's geographical region of residence in the United States. Using the northeast region as the reference level, the coefficient for the northwest region is -0.0811. This means that if all other predictors besides region were held constant, we would expect the post-insurance medical cost to be $0.08 more for individuals living in the northeast region in comparison to the post-insurance medical cost for individuals living in the northwest region. The coefficient for the southeast region is -0.1612. This means that if all other predictors besides region were held constant, we would expect the post-insurance medical cost to be $0.16 more for individuals living in the northeast region in comparison to the post-insurance medical cost for individuals living in the northwest region. The coefficient for the southwest region is -0.1144. This means that if all other predictors besides region were held constant, we would expect the post-insurance medical cost to be $0.11 more for individuals living in the northeast region in comparison to the post-insurance medical costs for individuals living in the southwest region. The entire confidence intervals for all three of the region levels are negative, which indicates that we can be 95% certain that there is a difference in post-insurance medical costs between the northeast region and each of the other regions. This can further be confirmed by considering the t-statistic and p-values. The northwest region has a t-statistic of -2.835 with a p-value of 0.005, the southeast region has a t-statistic of -5.614 with a p-value of 0.000, and the southwest region has a t-statistic of -4.011 with a p-value of 0.000. This confirms the significance of the region predictor in the model. Individuals living in the northeast region would be expected to have the highest post-insurance medical cost while individuals living in the southeast region would be expected to have the lowest.

Summary of our findings and results

By using the best subset method and comparing the Mallow's Cp, Adjusted R^2, Aikake's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) of the candidate, models, the final model that we chose to go with was the full model. To help alleviate some of the model discrepancies caused by heteroscedasticity, we fitted the model based on the log transformation of the medical charges variable. Even though the issue of heteroscedasticity still remained after the transformation, we decided to proceed using the transformed data due to the improvement in the model's Adjusted R^2.

In conclusion, our final model was,

log(charges) ~ age + sex + bmi + children + smoker + region

The F-test and high Adjusted R^2alluded to the model's significance and high predictive power. The model was significant at the individual predictor level as all of the variables in the model were significant according to the individual t-tests. According to our final model, a relatively young, non-smoking male with a low BMI and minimal children living in the southeast region of the United States, would be predicted to have the lowest post-insurance medical cost. On the contrary, a relatively old, smoking female with a high BMI and many children living in the northeast region of the United States, would be predicted to have the highest post-insurance medical cost.

Because many of the critical model assumptions were violated and we were unable to remedy them, we lost the reliability in many of our findings and interpretations. However we still proceeded and found some interesting insights regarding the factors that may have an impact on post-insurance medical costs.