UNIVERSITY OF SAN FRANCISCO

Using Linear Regression to Predict Health Care Cost

Group Members:
Cassidy Newberry, Shubham Thakur, Haotian Gong

Description of Dataset

This report uses the dataset "Medical Cost Personal Dataset: Insurance Forecast by using Linear Regression" published on the website Kaggle. This dataset, insurance.csv, can be found at this link: https://www.kaggle.com/mirichoi0218/insurance. The dataset includes seven columns and 1338 rows. The seven variables are age, sex, bmi, children, smoker, region, and charges. Here is a description of each variable included in the dataset:

- age - the age of the insurance holder
- sex - the sex of the insurance holder: male or female
- bmi - body mass index of the insurance holder
- children - number of children on the insurance plan
- smoker - the smoking status of the insurance holder: yes or no
- region - the insurance holder's location in the U.S: northeast, southeast, southwest, northwest
- charges - medical cost billed by health insurance

We can see that this dataset includes three numerical variables: age, bmi, and charges. We can also see that the dataset includes four categorical variables: sex, children, smoker, and region. The dataset contains 1338 observations, meaning it has information on 1338 individual people. This is not a very large dataset, but it is large enough to find general trends. The response variable will be charges, and we will attempt to find which of the given variables are most important in predicting the charges, which is the medical cost billed by the health insurance company.

Statement of Research Questions

When beginning this project, we wanted to see which variables had the greatest impact on the response variable, charges. Within this, we had four main research questions:

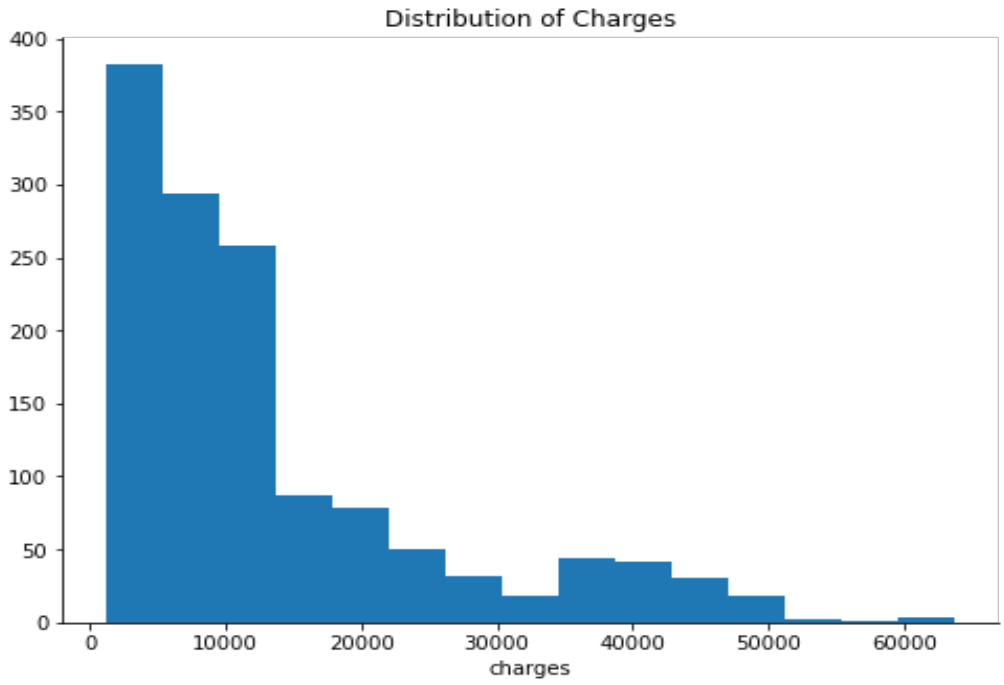
1. Which variables do not have a significant impact on the charges?
2. Which variables do have a significant impact on the charges?
3. How accurate are the predictions of charges using these significant variables?
4. What assumptions are violated when modeling on this data set?

In order to answer these questions, we conducted an in-depth analysis on this dataset. The first step we took was conducting an exploratory data analysis, where we looked at the relationship between variables through visualization, made sure all variables had the correct data type, and cleaned the data. Then we fit a model including all predictor variables available in the dataset. From there we performed various methods of model selection, like t tests and ANOVA. We also checked all linear regression model assumptions, as well as controlled for violated assumptions that occur in the model. After conducting various forms of model selection, we came to a conclusion about which model best predicts the response variables, charges.

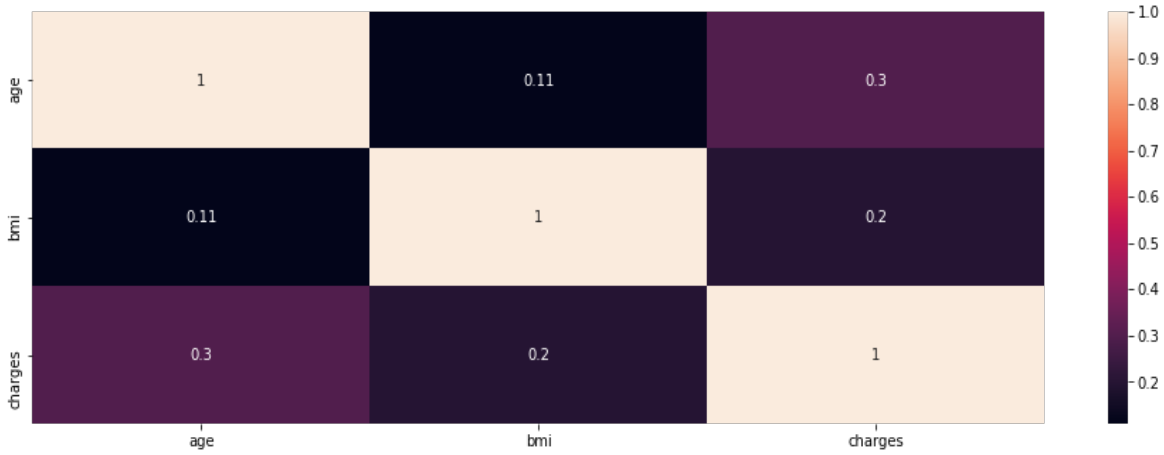
Exploratory Data Analysis

Our first step was to conduct an exploratory data analysis. First, we checked the data types of all columns to make sure they looked correct. The only one that concerned us was the column "children" which was currently represented as a numerical variable. We believed it should be a categorical variable because it only included values from 0-5. We changed the column "children" to be a categorical variable. We then checked to make sure our dataset did not contain any null

values, luckily the data did not so no further cleaning was necessary. We then wanted to check the distribution of our response variable, charges. Here is what it looked like:

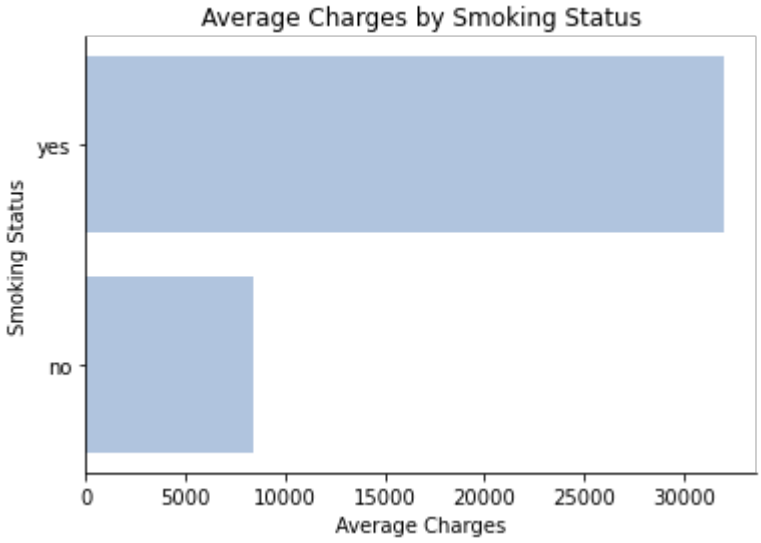


Here we can see that the distribution of charges is skewed right, and may have a few influential points that need to be examined. We will address these outliers later in the analysis, for right now it is important to note that they exist. We also looked at the correlation between all numerical variables to make sure that none of the columns were dependent on another. Here is the correlation plot we created:



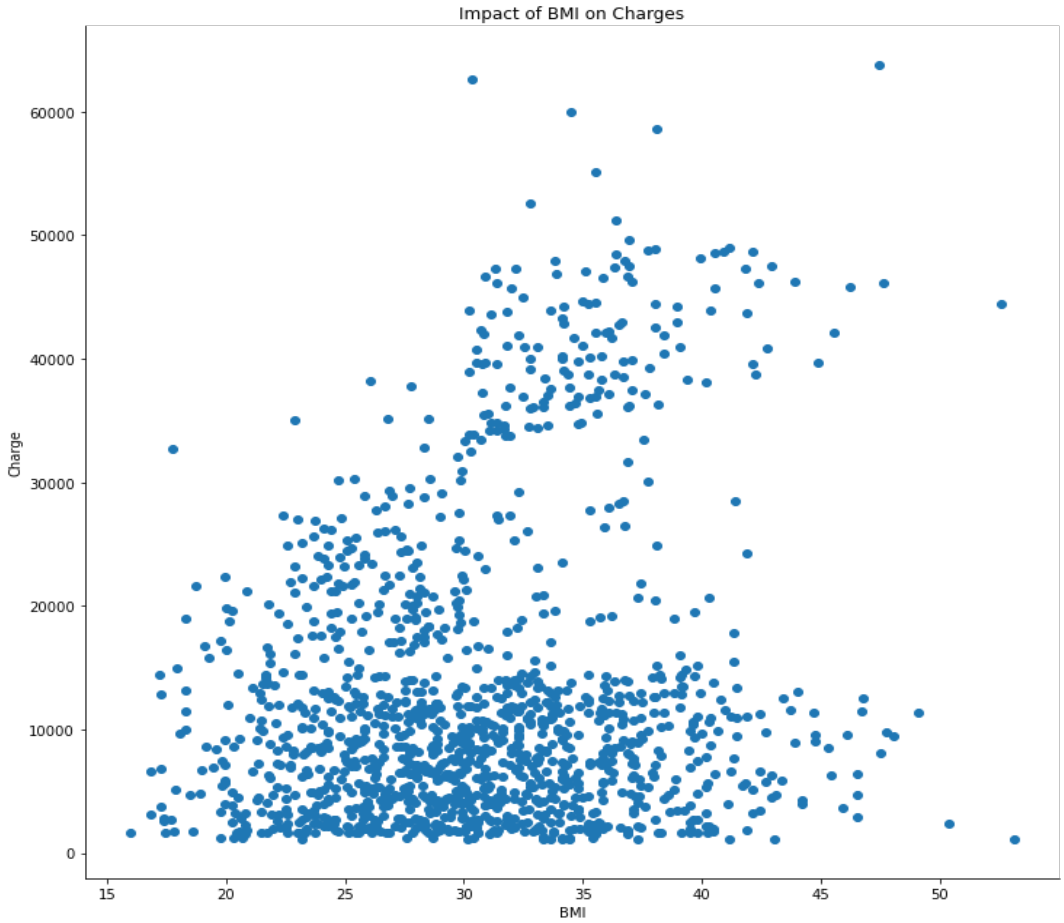
This plot made it clear that none of the numerical variables were highly correlated, meaning that they are likely all independent of each other and multicollinearity will most likely not be a problem in the future.

Next, we wanted to see what the relationship was like between each predictor variable and the response variable. First, we looked at the predictor variable smoking. For this, we looked at the average charges for the group of people that smoked versus the average charges for the group that did not smoke. Here are the results:



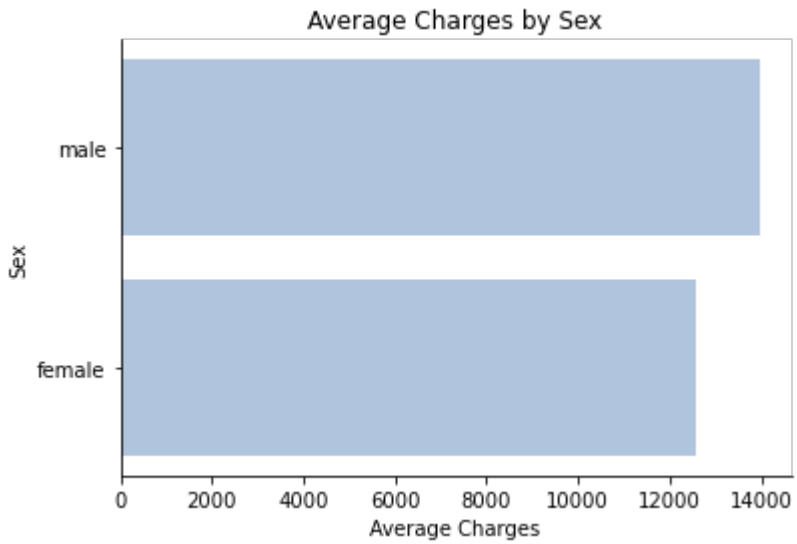
From this we can see that the variable smoking likely has a large impact on the charges, and will need to be included in the model. The difference between the average charges for these two groups is obvious and indicates that there is a relationship between smoking status and charges.

Next, we looked at the relationship between bmi and charges. We plotted a scatter plot of bmi versus charges for each observation:



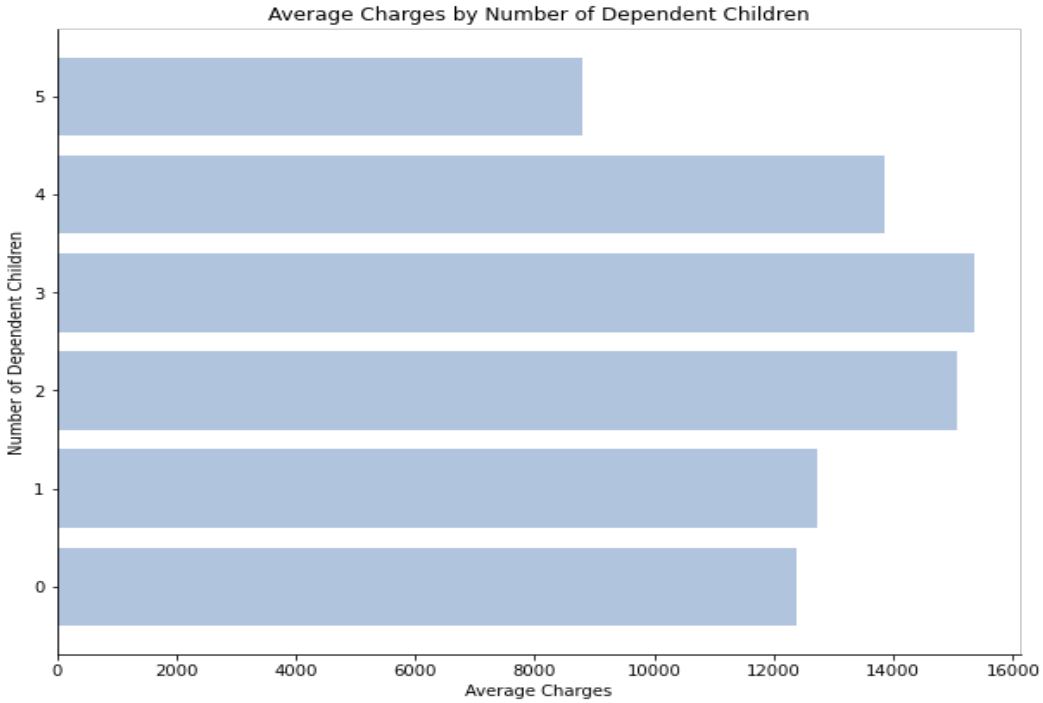
This plot shows that the relationship between bmi and charge is weak and linear. We can see that as the bmi increases the charge also increases but the relationship is not obvious due to the several points on the low end of charges that do not follow a linear relationship. From this, we noted that increasing bmi in general causes charges to increase, but this relationship is weak. We will need to look at hypothesis tests when we are doing model selection to see if this relationship is significant.

We then looked at the relationship between sex and charges. Since sex is a categorical variable, we again compared the average charges for each group:



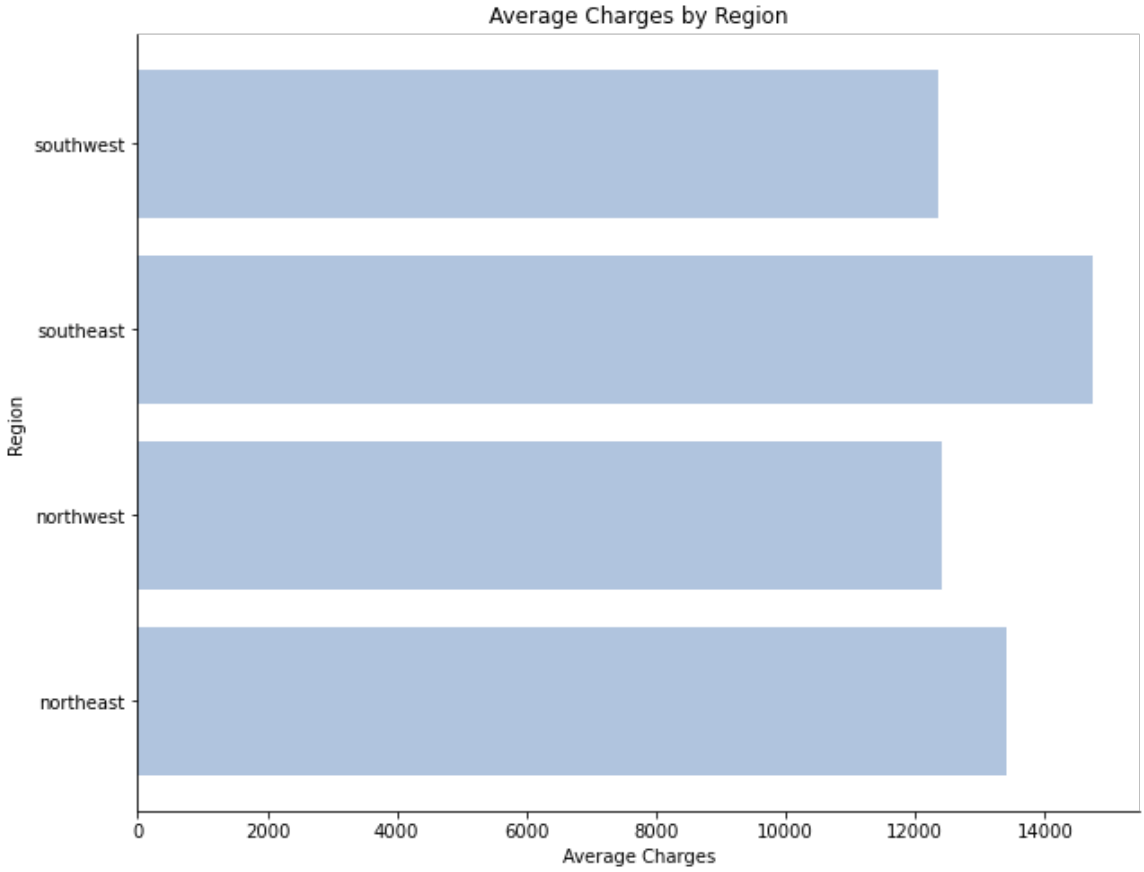
This graph does not suggest that there is a strong relationship between sex and charges. This led us to conclude that sex will most likely not be included in the model, because the average charges for both groups are similar.

Next, we looked at the relationship between children and charges. Once again, since children is a categorical variable, we plotted the average charges per number of children:



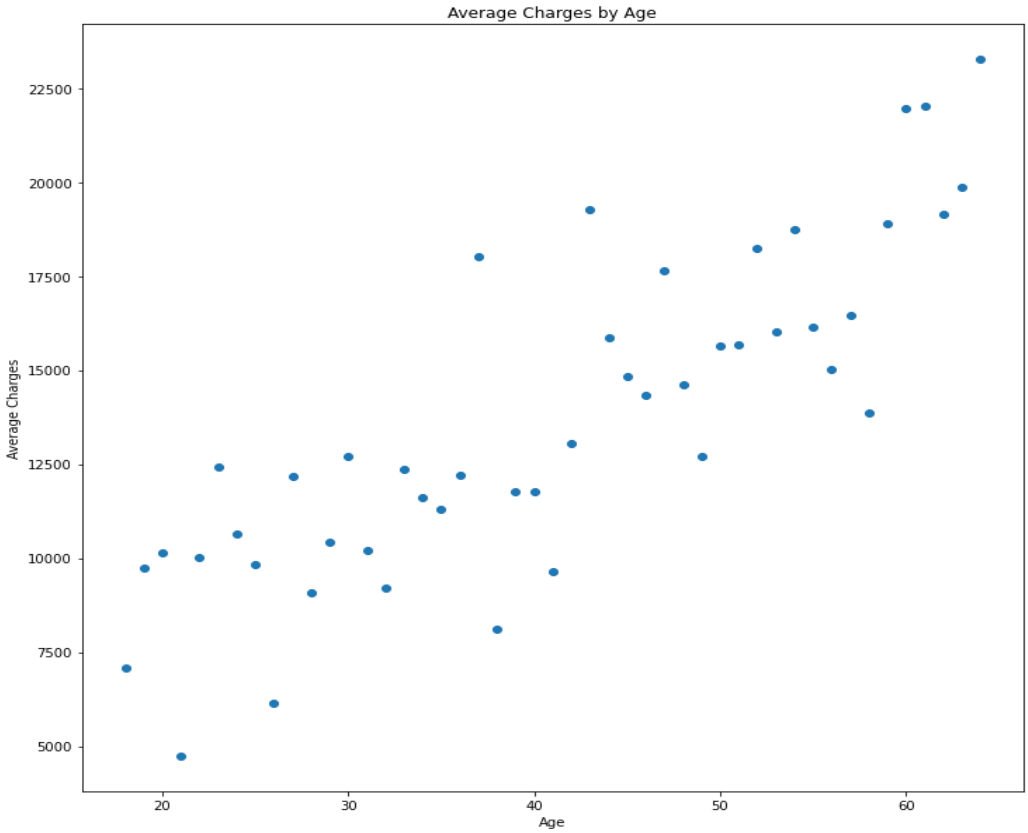
When looking at this bar plot, we concluded that the relationship between number of children and charges is most likely not significant. The bars are not very different from each, and no relationship is notable. Number of children will most likely not be included in the model.

Next, we checked the relationship between region and charges. We plotted this categorical variable the same way we plotted all other categorical variables:



Once again, we did not see a significant difference between the bars in this plot and concluded that there was likely not a significant relationship between region and charges. Region would most likely not be included in the model.

Lastly, we checked the relationship between age and charges. We plotted the average charge per age:



From this scatter plot we concluded that a linear relationship between age and charges exists, and it is a strong relationship. After seeing this plot, we believe that age will be an important, significant predictor of charges and will be included in the model.

In summary, after conducting our EDA we noted that age and sex have a large impact on charges, we are unsure of the impact of children and bmi on charges, and sex and region do not seem to have an impact on charges.

Regression Analysis

To start with, we chose to regress charges with all the available predictors. Based on the inferences coming from the model we will decide the most relevant predictors significant for the prediction of medical charges billed by health insurance.

Model Equation-1: *Charges ~ age+sex+bmi+children+smoker+region*

OLS Regression Results

Dep. Variable:	charges	**R-squared:**	0.752
Model:	OLS	**Adj. R-squared:**	0.750
Method:	Least Squares	**F-statistic:**	334.7
Date:	Fri, 15 Oct 2021	**Prob (F-statistic):**	0.00
Time:	11:30:50	**Log-Likelihood:**	-13545.
No. Observations:	1338	**AIC:**	2.712e+04
Df Residuals:	1325	**BIC:**	2.718e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.193e+04	993.657	-12.003	0.000	-1.39e+04	-9977.861
sex[T.male]	-128.1616	332.834	-0.385	0.700	-781.101	524.778
children[T.1]	390.9782	421.350	0.928	0.354	-435.608	1217.565
children[T.2]	1635.7772	466.670	3.505	0.000	720.284	2551.270
children[T.3]	964.3403	548.097	1.759	0.079	-110.893	2039.574
children[T.4]	2947.3680	1239.163	2.379	0.018	516.432	5378.304
children[T.5]	1116.0395	1456.015	0.767	0.444	-1740.307	3972.386
smoker[T.yes]	2.384e+04	414.139	57.557	0.000	2.3e+04	2.46e+04
region[T.northwest]	-380.0439	476.559	-0.797	0.425	-1314.936	554.848
region[T.southeast]	-1033.1375	479.139	-2.156	0.031	-1973.091	-93.184
region[T.southwest]	-952.8878	478.153	-1.993	0.046	-1890.908	-14.867
age	257.1933	11.914	21.587	0.000	233.820	280.567
bmi	336.9088	28.612	11.775	0.000	280.779	393.039

Omnibus:	293.990	**Durbin-Watson:**	2.084
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	695.789
Skew:	1.191	**Prob(JB):**	8.15e-152
Kurtosis:	5.609	**Cond. No.**	453.

The $R^2 \ -Adjusted$ for the baseline model with all the predictors is coming to be 0.75. The p-value for the global F-test is coming close to zero suggests that at least one of the predictors is significant for the prediction of the response variable. The summary table shows that the p-value for the predictor's age, bmi, children are less than 0.05. Also, at least one category of children and region is coming out significant, making the entire predictor significant for predicting response variable charges. The summary chart shows that predictor sex has a p-value less than 0.05, thus making it insignificant for the response variable charges. This is supported by our previous exploratory analysis for predictor sex, which suggests no strong relation between sex and charges.

Next, we checked results from the partial ANOVA to test if a predictor is adding any value to the model (with all the predictors) without it.

	sum_sq	df	F	PR(>F)
sex	5.442885e+06	1.0	0.148272	7.002538e-01
children	6.379964e+08	5.0	3.475992	3.987342e-03
smoker	1.216074e+11	1.0	3312.764072	0.000000e+00
region	2.264060e+08	3.0	2.055878	1.042786e-01
age	1.710575e+10	1.0	465.985849	8.446004e-89
bmi	5.089679e+09	1.0	138.650315	1.641433e-30
Residual	4.863908e+10	1325.0	NaN	NaN

Anova type-2 suggests that the predictor sex and region are coming out to be insignificant. Even the initial exploratory analysis suggests weak relation between region and charges.

As a result, we decided to regress charges against all the predictors except sex and region.

Model Equation-2: Charges ~ age+bmi+children+smoker

This model has all the significant predictors coming out of the first model. So, there is a good chance of the model showing all the predictors significant for the response variable.

OLS Regression Results

Dep. Variable:	charges	**R-squared:**	0.751
Model:	OLS	**Adj. R-squared:**	0.749
Method:	Least Squares	**F-statistic:**	500.4
Date:	Tue, 12 Oct 2021	**Prob (F-statistic):**	0.00
Time:	13:39:28	**Log-Likelihood:**	-13548.
No. Observations:	1338	**AIC:**	2.711e+04
Df Residuals:	1329	**BIC:**	2.716e+04
Df Model:	8		
Covariance Type:	nonrobust		

We see there is not a significant improvement in $R^2 \quad - Adjusted$ value compared to the last model. However, there is some improvement in log-likelihood.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.209e+04	947.781	-12.760	0.000	-1.4e+04	-1.02e+04
children[T.1]	368.7710	421.573	0.875	0.382	-458.250	1195.792
children[T.2]	1626.5095	466.561	3.486	0.001	711.233	2541.786
children[T.3]	996.9511	547.801	1.820	0.069	-77.697	2071.599
children[T.4]	2984.3586	1239.595	2.408	0.016	552.582	5416.135
children[T.5]	899.1294	1453.361	0.619	0.536	-1952.003	3750.262
smoker[T.yes]	2.38e+04	412.053	57.752	0.000	2.3e+04	2.46e+04
age	258.0760	11.912	21.665	0.000	234.707	281.445
bmi	319.8047	27.375	11.682	0.000	266.101	373.508

Omnibus:	294.840	**Durbin-Watson:**	2.084
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	697.185
Skew:	1.195	**Prob(JB):**	4.06e-152
Kurtosis:	5.607	**Cond. No.**	452.

Partial Anova:

	sum_sq	df	F	PR(>F)
children	6.425648e+08	5.0	3.494812	3.833167e-03
smoker	1.226453e+11	1.0	3335.245907	0.000000e+00
age	1.725954e+10	1.0	469.360028	2.226815e-89
bmi	5.018576e+09	1.0	136.476326	4.383973e-30
Residual	4.887065e+10	1329.0	NaN	NaN

From the t-test and Partial Anova, all the p-values of the predictors are coming out to be significant. The extremely small p-value suggests the strong relationship between the predictors and response variables. This supports the initial exploratory analysis for these predictors. Now that the predictors are significant, we will check if the estimated statistics and tests are reliable.

Model Diagnostics

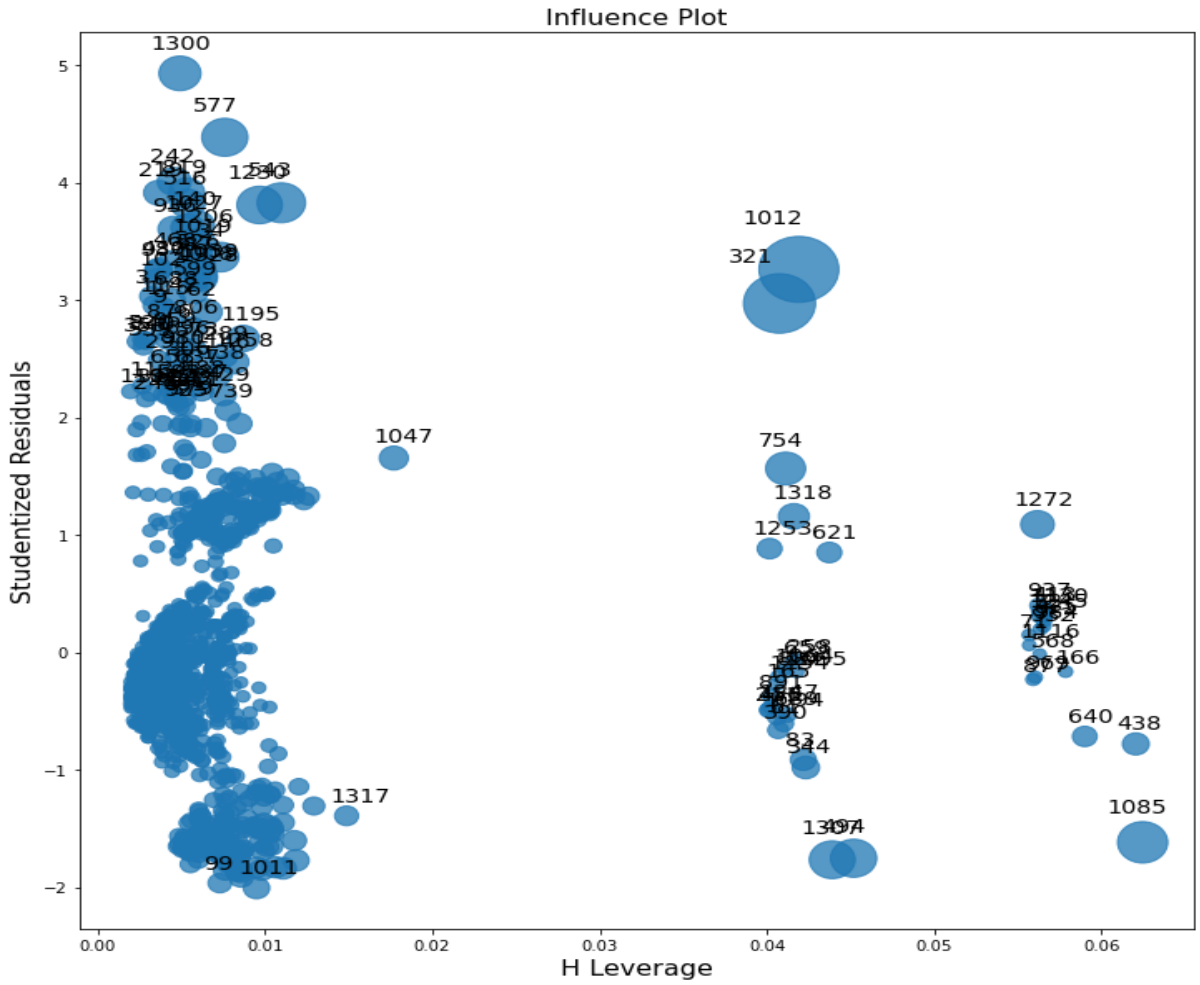
As a model diagnostics test, we will check if the model's assumptions are violated or some data structure issues.

Data Structure Problems:

Influential Points

We will first check if there are influential points in the data. If there are a high number of influential points, they might significantly impact the fitted line. We used Cook's distance to

detect influential points.



It's a plot between Studentized residuals and Leverage with the size of the bubble as the cook's distance. We can see that there are some points with high influence. We found 80 points with Cook's distance greater than 4/n, where n is the number of observations in the data. These points can significantly change the parameter estimate of coefficients of the predictors.

Multicollinearity

Multicollinearity is one of the data structure problems that significantly impact the coefficient of the predictors. It makes the t-test and ANOVA test unreliable. As we saw in the heat

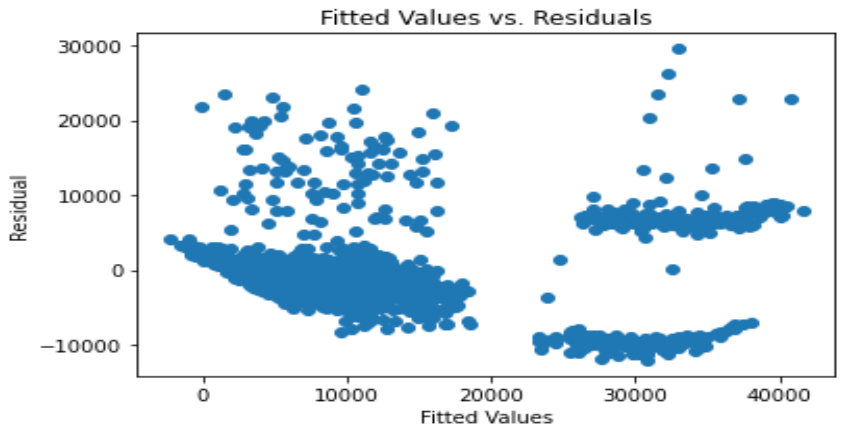
map of the predictors, there is no significant correlation among predictors. Also, we checked the Variance Inflation Factor (VIF) to detect any signs of multicollinearity among the predictors.

VIF Factor	features
32.685011	Intercept
1.186722	children[T.1]
1.165873	children[T.2]
1.130876	children[T.3]
1.025145	children[T.4]
1.020033	children[T.5]
1.006048	smoker[T.yes]
1.018469	age
1.013263	bmi

Ideally, VIF factor less than 4 suggests minimal signs of multicollinearity. From the above figure with the VIF of all the predictors, we can say no multicollinearity among the predictors used in this model.

Model Assumption Violation:

We will use Fitted Values plot vs Residual to evaluate if any model assumptions are getting violated.

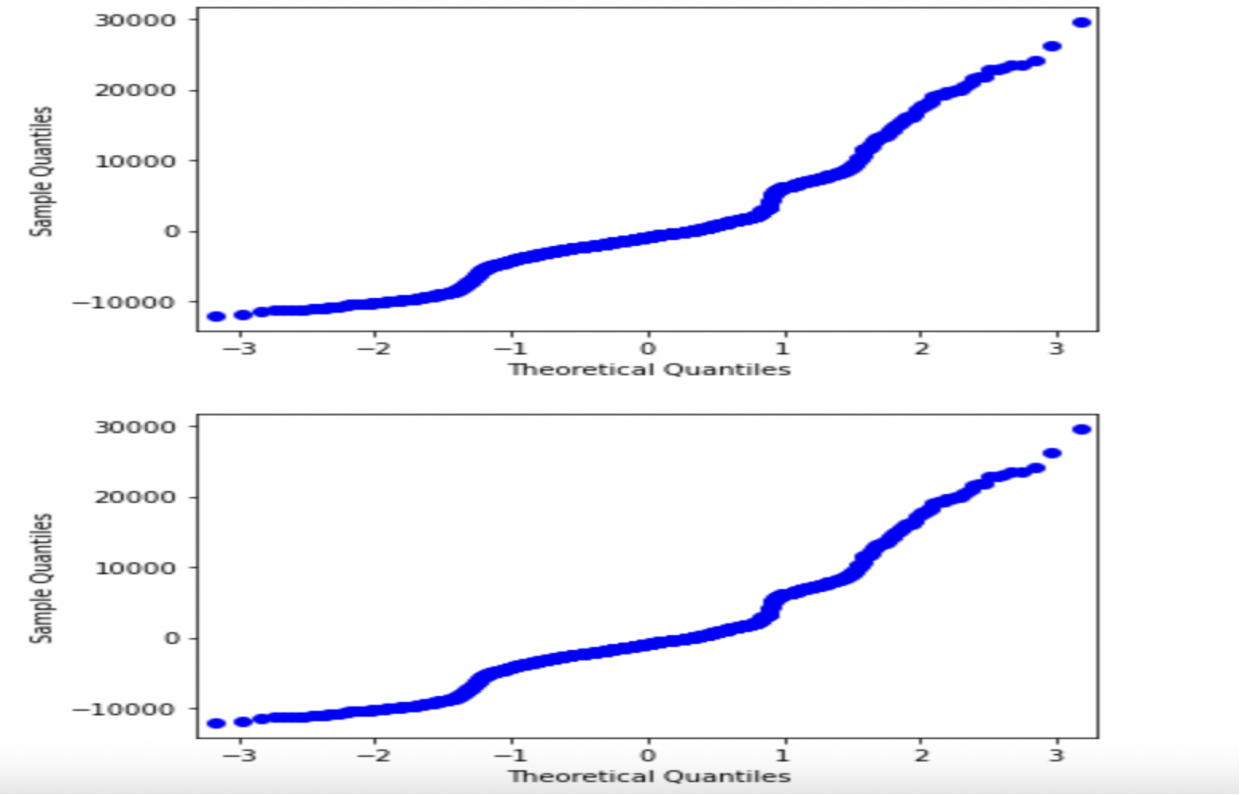


From the plot, we can see that

1. The E(ε) not equals zero as there are some points scattered far from the mean.
2. There is significant heteroskedasticity in the model as the variance of errors is not constant.
3. There is a significant number of influential points which we discussed in the above section.

Non Normality Residuals:

Normality of error term is required to perform t-test, Anova, and Confidence interval. We used the QQ plot to check if the normality assumption is violated.



As we can see from the QQ plot the normality assumption is violated. To support this test we checked the results of JB test. We can see from the model summary the JB p-value is less than the significance level of 0.05 which suggests that the normality assumption violated.

Heteroscedasticity

Similar to the problems we face in the case of multicollinearity, the t-test results might not be reliable. We also face challenges in getting the prediction and confidence intervals for y.

We used BP test to detect Heteroskedasticity. The p-value coming out was approximately close to zero which suggests that predictors used in the model are significant in predicting the errors and thus it violated the assumption of constant variance of error irrespective of predictor values. This suggests the model has high heteroskedasticity.

Final Model Choice

For the current model few assumptions are violated, thus we used the weighted least square regression to reduce the problem. While fitting the model this gives relatively less weightage to large error variance and high weightage to low error variance. Thus, we expect this model to have better performance compared to last model.

Weighted Least Square Regression

Model Equation: WLS(Charges ~ age+bmi+children+smoker)

We regressed errors with predictors and then used the estimated values of errors to get inverse of weights for each observation. We used the same predictors from the last model. However, as we don't know the significance of influential points, we fitted two model: One with influential points and other without influential points.

WLS Regression Results

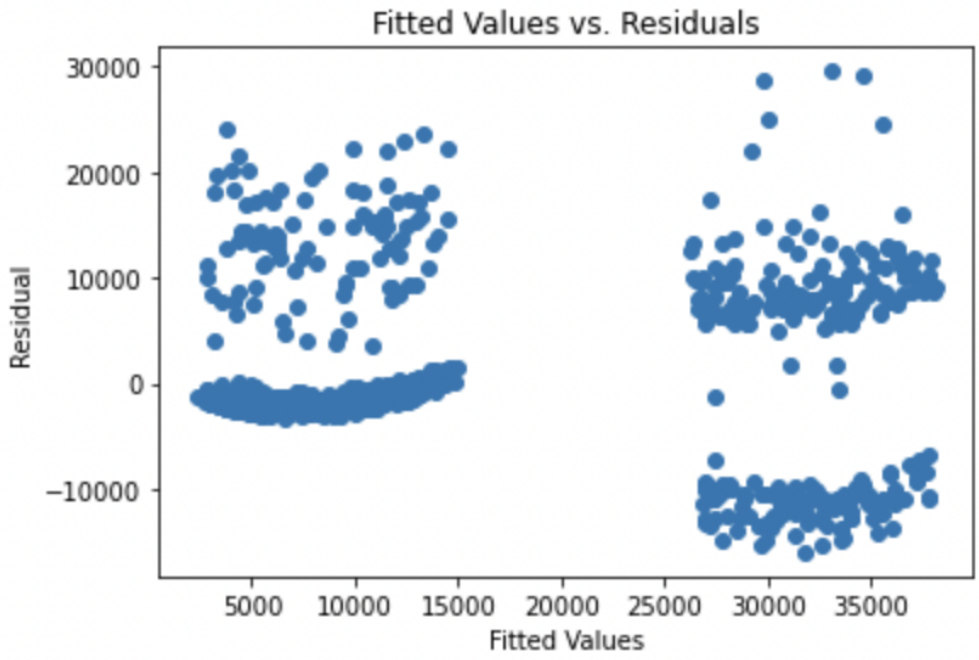
Dep. Variable:	y	**R-squared (uncentered):**	0.850
Model:	WLS	**Adj. R-squared (uncentered):**	0.850
Method:	Least Squares	**F-statistic:**	1896.
Date:	Tue, 12 Oct 2021	**Prob (F-statistic):**	0.00
Time:	13:39:39	**Log-Likelihood:**	-13411.
No. Observations:	1338	**AIC:**	2.683e+04
Df Residuals:	1334	**BIC:**	2.685e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	235.8653	9.082	25.969	0.000	218.048	253.683
x2	-36.2661	12.625	-2.873	0.004	-61.033	-11.500
x3	316.1644	98.638	3.205	0.001	122.661	509.668
x4	2.354e+04	558.415	42.147	0.000	2.24e+04	2.46e+04

Omnibus:	553.116	**Durbin-Watson:**	2.081
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2140.732
Skew:	2.020	**Prob(JB):**	0.00
Kurtosis:	7.698	**Cond. No.**	205.

From the summary we can see the R-squared value and Adjusted squared value improved a lot by using weighted least square. This is because we used gave less weightage to the points that have high variance and thus there were less influencing the coefficient of the predictors. From the t-test results from the model, all the predictors used for the model are coming out to be significant.

While the model evaluation metric like R-square improved a lot, we need to check if the assumption are still violated.



There is a small improvement from the last residual plot however, the assumptions of linear regression are still violated.

1. It has heteroskedasticity i.e non constant variance of error.
2. The E(ε) not equals to zero

Since there are still some influential points as see from the graph, we rebuilt the weighted least square without the influential points. We removed the 80 influential points that we found from the model.

WLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.918
Model:	WLS	**Adj. R-squared (uncentered):**	0.917
Method:	Least Squares	**F-statistic:**	3489.
Date:	Tue, 12 Oct 2021	**Prob (F-statistic):**	0.00
Time:	13:47:31	**Log-Likelihood:**	-12103.
No. Observations:	1258	**AIC:**	2.421e+04
Df Residuals:	1254	**BIC:**	2.424e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	238.2296	5.930	40.171	0.000	226.595	249.864
x2	-56.7657	8.279	-6.857	0.000	-73.007	-40.524
x3	180.1596	45.571	3.953	0.000	90.755	269.564
x4	2.441e+04	515.437	47.350	0.000	2.34e+04	2.54e+04

Omnibus:	617.036	**Durbin-Watson:**	1.970
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3753.298
Skew:	2.238	**Prob(JB):**	0.00
Kurtosis:	10.181	**Cond. No.**	307.

After removing the influential data points, we got the Adjusted R-square of 0.917 which is a big improvement from the baseline model we started with. Also, the all the predictors used for this model have approximately zero p-value which suggests that the model predictors are significant in predicting the response variable. This suggests that influential points were significantly impacting our model. Depending upon the business requirement we may thus decide to remove them completely from the model.

So, the final model is charges ~ *age+bmi+children+smoker*, fitted by Weighted Least Squares regression. And the dataset excludes influential points which are identified by the cook's distance.

Model Selection

We used Partial Anova and t-test to come up with the final model. As a model selection step, we first fitted the model with all the predictors and then based on the results from Anova and t-test we started removing all the predictors that are not significant in the prediction of the response variable. In our final model we used *age, bmi, children and smoker* to get the best model.

Model Summary

In summary, we first fitted the OLS model with all predictors (model 1), and the F Anova test shows that region and sex are insignificant predictors. Removing both predictors, the new OLS model has R^2 score of 0.75 (model 2). For multicollinearity checking, we did the VIF and no violation for selected predictors. We also ran the BP test, which shows significant heteroskedasticity. To construct a better model, we fitted the WLS regression and R^2 is 0.85 (model 3), a great improvement compared with the OLS model. With Cook's Distance, we identified 80 influential data points. After removing them, the final WLS model (model 4) achieves R^2 0.92, a good performance!

Potential Problems and Further Discussion

In the final model, we still observe non-normality. The JB test score didn't improve much, even with WLS regression, though R^2 is better. We believe there are two reasons. The first is that people with lower charge and higher charge follow different patterns. Like shown in the residual plot of our final model, even after WLS adjustment, residuals with higher charges are still scattered compared with lower-charged data points. The second reason is that there are not enough data in the \$15K~\$25K range, causing inconsistency in our dataset.

In the book *Machine Learning with R*, the author fitted a moderated model which takes the product of some original predictors. We think that might be worth of trying, but it would require a deeper understanding of the dataset.