Regression Analysis on Health Insurance Cost

Christabelle Pabalan, Dashiell Brookhart, and Sicheng Zhou

1. Introduction

Regression analysis is a technique in statistics to investigate and model the relationship between variables (Douglas Montgomery, Peck, & Vinning, 2012). Multiple linear regression depends on modeling a relationship between the dependent (or response) variable and the independent (or predictor) variables. The purpose of this research is to analyze health insurance data to verify whether or not regression analysis models work effectively to predict health insurance charges. This study was also used to discover relevant factors that affect the cost of health insurance and investigate the extent in which these predictor variables are successful at prediction.

2. Problem Statement

We examined 5 main questions:

1. What are the differences in significant predictors of health insurance cost for individuals who smoke and those who do not?
2. For a 45 year old individual with a BMI of 35 in the subset population of smokers, what is the expected cost of their health insurance?
3. For a 45 year old individual living in the northwest with 3 children in the subset population of nonsmokers, what is the expected cost of their health insurance?
4. How does the cost of health insurance change for an individual in the nonsmoker population when they have a child?
5. How does health insurance price change in respect to an increase in body mass index within the smoker population?

3. Methodology

3.1 Dataset Description The dataset we analyzed is titled, "Medical Cost" from Kaggle. A link to the dataset is here. This dataset has 1,338 observations with seven different variables. Three of these variables are numerical (Age, BMI, and Charges) and the rest are categorical (Sex, Children, Smoker, and Region).

Feature Description	
Age	The age of the primary beneficiary.
Sex	The gender of the insurance beneficiary (male or female)
BMI	The body mass index of insurance beneficiary; ratio of mass to height $BMI = \frac{m}{h^2}$ where m is mass in kilograms and height is in meters.
Children	The number of children covered by the health insurance (number of dependents). *CHANGE*
Smoker	Whether the primary beneficiary smokes or not (yes or no)
Region	The beneficiary's residential area in the U.S. (northeast, southeast, southwest, or northwest).
Charges	The individual medical costs billed by the health insurance

Our initial explanatory analysis included creating a **histogram** of the variables in the dataset as well as **cross tables**:

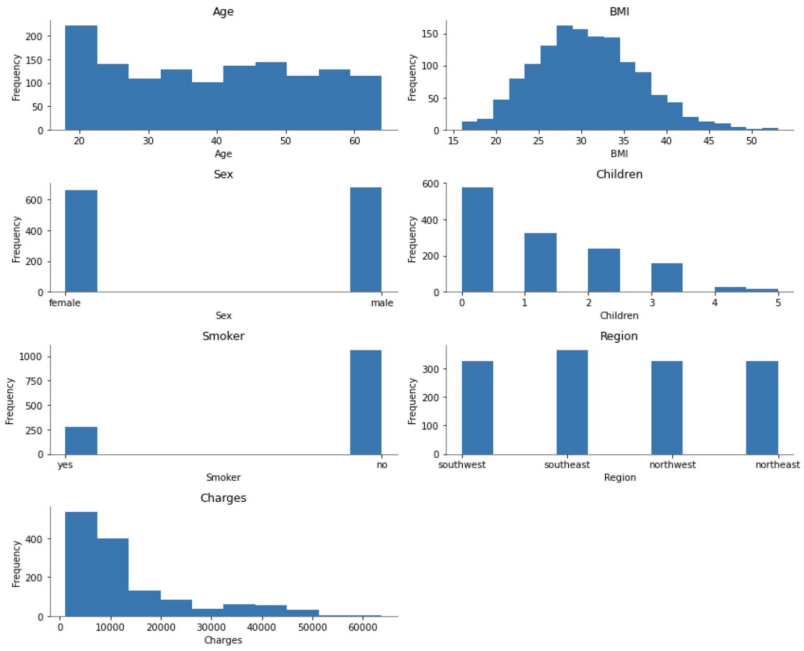


Figure 1 *Histograms of Variables*

From **Figure 1** we can immediately make some inferences about the dataset. We have a wide variety of ages, but there seems to be a large amount of ~20 year olds compared to other age groups. The histogram of BMI gives the impression that the distribution is roughly normal, with a long right tail obviously from some large BMI values in the observations. We seem to have a balanced data set for sex and regions as there seem to be an equal number of male and female data points and an equal number of data points between regions (with Southeast being only a little larger). For this particular dataset, it seems that it is more common to have fewer children (0 - 2 children) than it is to have more. Finally, we see from the histogram for charges that in our dataset that there is a large amount of charges in the $0 - $10,000 range with a long right tail of higher insurance charges extending to $60,000.

From **Figure 2** we are able to state that the number of children is relatively similar across all regions. Again, it seems that most observations include no children and that it is very uncommon to have an observation that has four or five children.

Regions / Number of children	northeast	northwest	southeast	southwest	Total
0	147	132	157	138	574
1	77	74	95	78	324
2	51	66	66	57	240
3	39	46	35	37	157
4	7	6	5	7	25
5	3	1	6	8	18
Total	324	325	364	325	1338

Figure 2 *Number of children per region*

Figure 3 indicates that it is more common to have a smoker observation where the sex is male rather than female. Overall, from what we viewed in the histograms of smokers vs. non-smokers, the majority of our observations came from non-smokers.

Figure 3 *Number of smokers per sex*

Sex / Smoker	female	male	Total
no	547	517	1064
yes	115	159	274
Total	662	676	1338

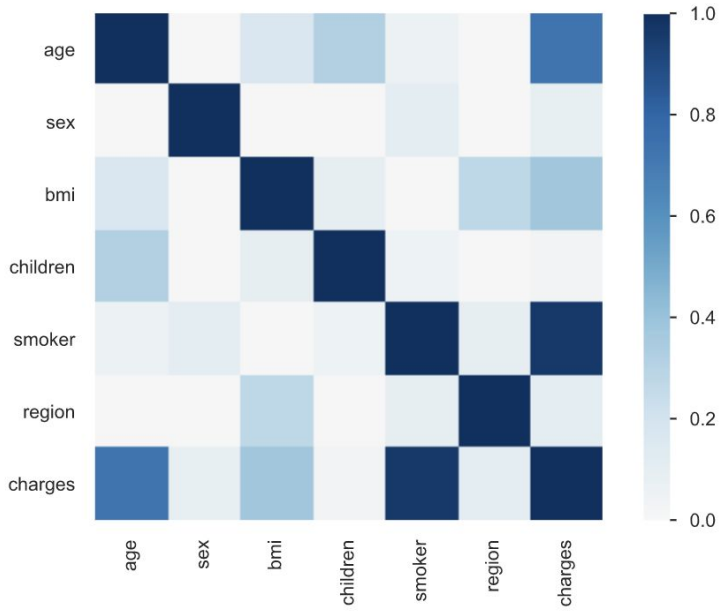
Regions Smoker	northeast	northwest	southeast	southwest	Total
no	257	267	273	267	1064
yes	67	58	91	58	274
Total	324	325	364	325	1338

Figure 4 *Number of smokers per region*

From **Figure 4,** we see that the northwest and southwest variables have the exact same number of smokers to non-smokers. The southeast region also has the highest ratio of smokers to non-smokers.

After receiving some summary details about the dataset from the initial explanatory analysis, we moved onto a more in-depth explanatory analysis of the variables in order to explore data assumptions. This included a **phiK correlation coefficients heatmap,** a **scatterplot matrix, interaction plots,** and **partitioned regression plots**:

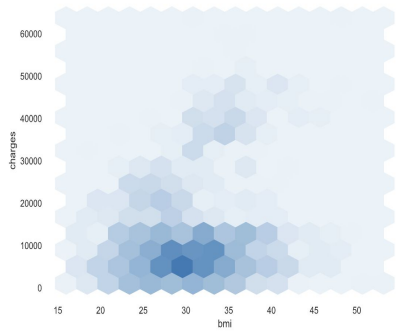
Figure 5 *PhiK Correlation Heatmap*



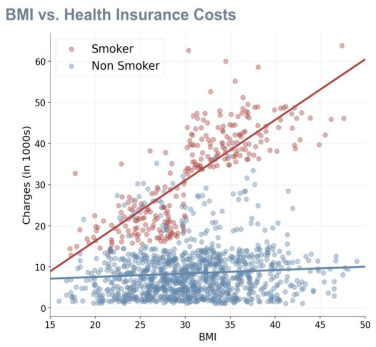
In **Figure 5,** we see that the strongest correlations with our response variable (charges) are with smoker, age, and BMI. One data assumption includes independence of predictor variables. This means that there are no

signs of correlation between predictor variables. From the figure above, there does not seem to be the presence of multicollinearity (strong correlations between predictor variables).

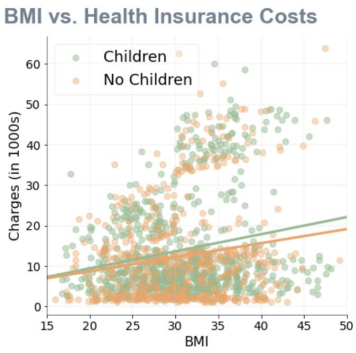
Figure 6 *Interaction Plots Response Variable vs. Individual Numerical Predictors*



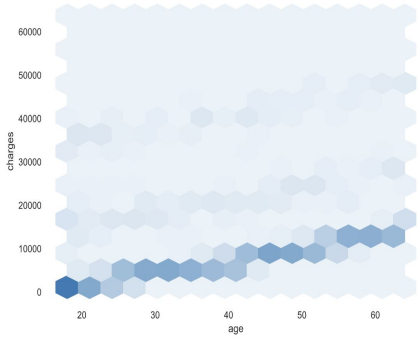
6.1 *Hexbin Plot: BMI vs.Charges*



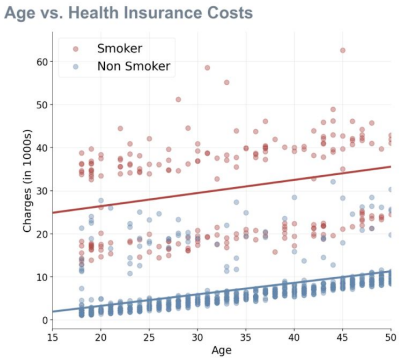
6.3 *Partitioned by Smoker Scatter / Regression Plot for BMI vs. Charges*



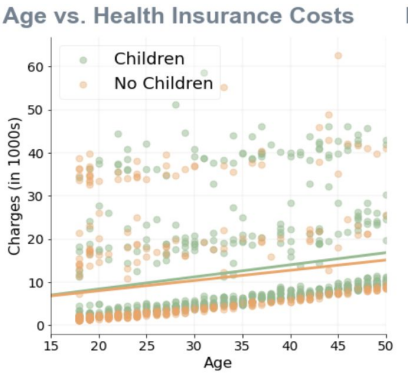
6.5 *Partitioned by Children Scatter / Regression Plot for BMI vs. Charges*



6.2 *Hexbin Plot: Age vs.Charges*



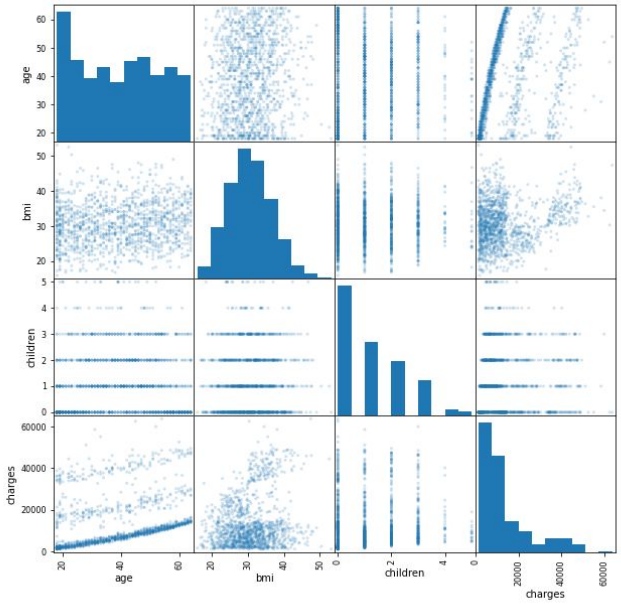
6.4 *Partitioned by Smoker Scatter / Regression Plot for Age vs. Charges*



6.6 *Partitioned by Children Scatter / Regression Plot for Age vs. Charges*

From **Figures 6.1, 6.2 and 7,** we observe that there is a clear linear relationship between age and charges. However, there are faintly visible partitions of linearly distributed data. This could mean that the data is best split into various subpopulations. We can also see this split in populations from the BMI vs. Charges plot. In **Figure 6.3 and 6.4,** we detail the graphs by color-coding based on whether or not the individual comes from the smoker or non-smoker population. Here we observe that the population of those who smoke consistently pay more than those who do not throughout all ages and BMI.

Figure 7 *Scatterplot Matrix of Features*



In **Figure 6.3,** there should be two different regression models for those who fall under the category of smoker and those who fall under nonsmoker. There are no smokers in the lower cluster, however, there are non-smokers in the upper cluster. This could cause an issue in modeling (for the nonsmoker population) if we decide to split the dataset. **Figure 6.4** affords us the ability to visualize an evident tier of charges clustered into three groups which is consistently demonstrated over age. Another interesting observation from this plot is that we can see a definite differentiation in prices between smokers and nonsmokers; nonsmokers are mainly clustered in the lowest tier while smokers are distributed among the top two tiers. From **Figure 6.5 and 6.6,** there does not seem to be a distinct relationship between BMI and health insurance costs, even when the dataset is split between observations with or without children. In the Age vs. Health Insurance Costs plot, for the lowest tier (composed of non-smokers as previously shown), those who do not have children pay less on insurance.

Since the total number of observations is 1,338 we will most likely not have to worry about normality issues. We only have 6 predictor variables to choose from and the sample size is larger than 30 * p, where p is the number of parameters. However, it is evident that there are clear partitions in data that may require different linear models if we cannot fix it with robust regression techniques. In this case, we may need to recheck this if we decide to split the dataset into subsets. Additionally, the phiK correlation coefficients heatmap shows that we most likely won't have to deal with a collinearity issue if we use charges as the response variable.

3.2 Methods and Measures

There are several fundamental model and data assumptions necessary to justify the use of linear regression models for the purpose of inference or prediction. These include: independence of predictor values, constant variance of errors, normality of the error distribution, independence of errors, and linearity between the response variable and any predictor variable. This section displays the methods and measures that were selected to optimize our regression model and correct issues with underlying model assumptions.

Sequential Analysis of Variance (ANOVA): The Sequential form of ANOVA is the calculation of the reduction in the error sum of squares (SSE) when one or more predictor variables are added to the model. It helps us by indicating the significance of a predictor variable given the predictors listed before it are already in the model.

Partial Analysis of Variance (ANOVA): The Partial form of ANOVA differs from the Sequential version as the order of the predictors does not matter. The Partial form calculates the significance of a predictor given all the rest of the predictors are already in the model.

Variance Inflation Factor (VIF): VIF measures how much the variance is inflated in the coefficient estimate caused by a predictor variable. We calculate the VIF score for each predictor variable and say that if the value is greater than or equal to 10, then that particular predictor variable is causing serious multicollinearity.

PhiK Correlation Heatmap: The PhiK measure is similar to the Pearson correlation coefficient in its interpretation but it works consistently between categorical, ordinal and interval variables. The correlation heatmap allows us at a glance to distinguish which variables in our dataset are correlated to each other. This helps us determine whether or not we may have a serious multicollinearity problem. In our report, variables that are highly correlated are shaded in a different color in the heatmap.

Externally Studentized Residuals: Externally studentized residuals help identify points that negatively affect a regression model. For a specific point it is the residuals of the model without that observation included over the estimated standard deviation without that observation included. The rule of thumb to calculate possible influential points using this method is to collect all points whose absolute value of their externally studentized residual is greater than or equal to $t_{/2},\ df = n - p - 1$. We use the intersection of points highlighted by both Cook's Distance and Externally Studentized Residuals as our influential points

Cook's distances: Cook's distance helps identify points that negatively affect a regression model. Cook's distance is a combination of an observation's leverage and residual value. The rule of thumb is that if the value of the Cook's distance for an observation is greater than $\frac{4}{n}$ (where n is the number of observations in our sample) then that observation may be an influential point. We use the intersection of points highlighted by both Cook's Distance and Externally Studentized Residuals as our influential points.

Log Transformation on Response Variable: Logarithmic transformation is a common method to transform a highly skewed variable into an approximately normal distribution. For regression purposes, it is also used to stabilize variance if the residuals consistently increases and to linearize a regression model with an increasing slope.

Quantile-quantile plots (Q-Q): Q-Q plots are a graphical technique that helps determine whether a model has a normal distribution. If the resulting plot is approximately linear on the diagonal of the plot, then it suggests that the error term of the model follows a normal distribution.

Breusch-Pagan test: The Breusch-Pagan test follows the idea that the variance of errors should not change given different predictor values. If the assumption of constant variance is violated, then the variance would change with the predictor values. This test helps us identify if our model has a significant heteroscedasticity problem. If the p-value of the Breush-Pagan test is less than or equal to 0.05, then we state that we have a significant heteroscedasticity problem.

Omnibus and Jarque-Bera tests: The Omnibus and Jarque-Bera tests are used to check for a violation of normality. They both test for this by calculating the skewness and kurtosis levels of a model. If the p-value is less than or equal to 0.05 for either test, then we say that the residuals of the model may not follow normality.

Residuals vs. Predictor plot: The residuals vs. predictors plots are used to test the assumption of linearity between the response variable and the predictor variables. If the plot shows a clear linear relationship then we say that the assumption of linearity holds for that particular predictor variable.

Robust Regression: Robust regression was used as compromise between excluding the high influential points entirely and including them as an equal data point. This model is essentially a replacement for the least squares regression and gives a different weight to the outliers found in our dataset. We ran through a parameter selection process by comparing various robust linear model functions (which defines the weight given to outliers). The parameters used included Huber, Least Squares, Hampel, Andrew Wave, Trimmed Mean and RamsayE.

4. Model Selection Results

4.1 Initial Model Fitting We began by regressing health insurance charges against age, sex, body mass index, children, smoker and region utilizing ordinary least squares through the statsmodels module.

Figure 8 *Individual Independent T-Test from OLS model*

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.2e+04	993.991	-12.074	0.000	-1.4e+04	-1.01e+04
sex[T.male]	-126.4104	333.313	-0.379	0.705	-780.287	527.466
children[T.1]	999.5804	335.880	2.976	0.003	340.668	1658.493
smoker[T.yes]	2.385e+04	413.627	57.660	0.000	2.3e+04	2.47e+04
region[T.northwest]	-352.2153	476.882	-0.739	0.460	-1287.738	583.308
region[T.southeast]	-1057.3312	479.275	-2.206	0.028	-1997.549	-117.114
region[T.southwest]	-944.2598	478.401	-1.974	0.049	-1882.764	-5.755
age	256.9072	11.915	21.562	0.000	233.533	280.281
bmi	339.5097	28.631	11.858	0.000	283.342	395.678

4.1.1 Individual Independent T-Test. As shown on **Figure 8,** all predictors except for sex are significant predictors of charges. Since there is at least one level that is significantly different in "charges" from the baseline level for both of the categories "children" and "region", then both predictors have a significant role in predicting "charges".

4.1.2 Sequential and Partial ANOVA Tests Based on the sequential ANOVA test (**Figure 9.1)**, we should drop 'region' from the model. This is divergent from the conclusion in the individual t-test. According to the partial ANOVA test (**Figure 9.2**), we should drop both 'region' and 'sex' from the model. Again, while the individual t-test has a different conclusion for region, it has the same conclusion for "sex" .

Figure 9 *ANOVA results*

	df	sum_sq	mean_sq	F	PR(>F)
sex	1.0	6.435902e+08	6.435902e+08	17.473263	3.104289e-05
children	1.0	8.006638e+08	8.006638e+08	21.737760	3.440406e-06
smoker	1.0	1.207210e+11	1.207210e+11	3277.536727	0.000000e+00
region	3.0	1.180499e+08	3.934996e+07	1.068339	3.614650e-01
age	1.0	1.966094e+10	1.966094e+10	533.788053	1.477155e-99
bmi	1.0	5.179078e+09	5.179078e+09	140.610268	6.655185e-31
Residual	1329.0	4.895087e+10	3.683286e+07	NaN	NaN

Figure 9.1 *Sequential ANOVA Test Results*

	sum_sq	df	F	PR(>F)
sex	5.297821e+06	1.0	0.143834	7.045593e-01
children	3.262144e+08	1.0	8.856615	2.973079e-03
smoker	1.224565e+11	1.0	3324.653125	0.000000e+00
region	2.363591e+08	3.0	2.139024	9.352488e-02
age	1.712372e+10	1.0	464.903499	1.162116e-88
bmi	5.179078e+09	1.0	140.610268	6.655185e-31
Residual	4.895087e+10	1329.0	NaN	NaN

Figure 9.2 *Partial ANOVA Test Results*

The resulting R-squared and adjusted R-squared values from our model are 0.75 and 0.749, respectively.

OLS Regression Results

Dep. Variable:	charges	**R-squared:**	0.750
Model:	OLS	**Adj. R-squared:**	0.749
Method:	Least Squares	**F-statistic:**	499.3
Date:	Tue, 01 Dec 2020	**Prob (F-statistic):**	0.00
Time:	20:30:37	**Log-Likelihood:**	-13549.
No. Observations:	1338	**AIC:**	2.712e+04
Df Residuals:	1329	**BIC:**	2.716e+04
Df Model:	8		
Covariance Type:	nonrobust		

4.2 Checking Underlying Data and Model Assumptions

4.2.1 Multicollinearity Detection.

Figure 10 *VIF Scores*

	VIF Factor	features
0	35.891002	Intercept
1	1.008828	C(sex)[T.male]
2	1.003879	C(children)[T.1]
3	1.012088	C(smoker)[T.yes]
4	1.519226	C(region)[T.northwest]
5	1.652486	C(region)[T.southeast]
6	1.528925	C(region)[T.southwest]
7	1.017269	age
8	1.106585	bmi

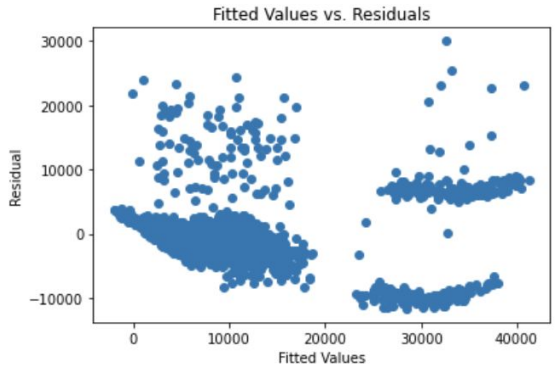
Recall from **Figure 5**, we know that the coefficients correlation heatmap only shows strong correlations between the response variable (charges) and the predictor variables: smoker, age, and BMI. There does not seem to be strong correlations between any other variables. Thus we are not that worried about there being a significant multicollinearity

problem. However, we also calculated the variance inflation factors to ensure that the variance is not being inflated by any individual predictor variable. Based on our calculations shown in **Figure 10,** the VIF scores for each predictor variable is ~1 indicating that there is no significant multicollinearity in this data.

4.2.2 Heteroscedasticity Detection.

We began our detection process by examining the fitted vs. residual plot. If there is not a significant heteroscedasticity issue, we would expect to see the points form a horizontal band around the 0 line to suggest that the variances of the error terms are equal. However, we can evidently see in **Figure 11** that there appears to be a major heteroscedasticity issue since the variances do not form a horizontal band around the 0 line.

Figure 11 *Fitted Values vs. Residuals*



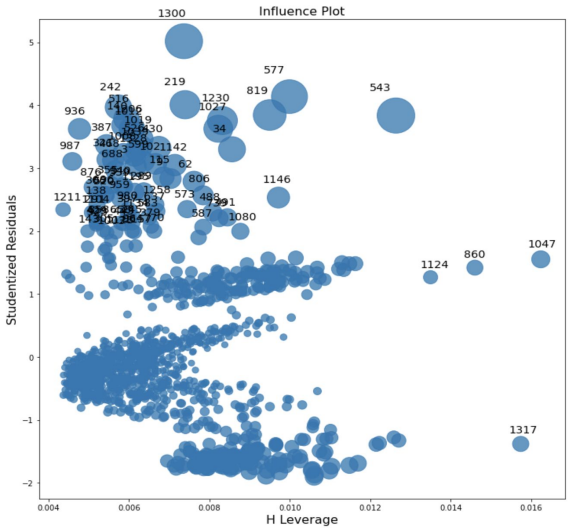
To confirm our suspicions, we ran a Breusch-Pagan test. This resulted in a LM statistic of 124.40 and a p-value of 4.09e-23 which is less than an alpha level of 0.05. This means we rejected the null hypothesis and concluded that there is a significant model assumption violation. This result was expected since our exploratory analysis revealed that it may be better to partition the data into subgroups with the most striking differentiation being smokers and non-smokers. Before we partition the model, we decided to attempt to use heteroscedasticity methods to correct the test results which included logarithmic transformation on the response variable and robust linear regression methods.

4.2.2 Heteroscedasticity Solutions.

4.2.2.1 Dropping Influential Points

We began by examining the effect of influential points on our model assumptions. **Figure 12** is the influence plot where x-axis is the leverage, y-axis is the external studentized residuals, and the size of the points represent Cook's distance. We dropped the intersection of the points identified by the externalized studentized residual test and the Cook's distance test. After dropping the influential points, we re-ran the Breusch-Pagan test. This resulted in a LM statistic of 603.12 and a p-value of 4.979e-125.

Figure 12 *Influence plot where x-axis is the leverage, y-axis is the external studentized residuals, and the size of the points represent Cook's distance.*



Therefore, we still reject the null hypothesis and conclude there is a significant heteroscedasticity issue.

4.2.2.1 Log Transformation on Response Variable

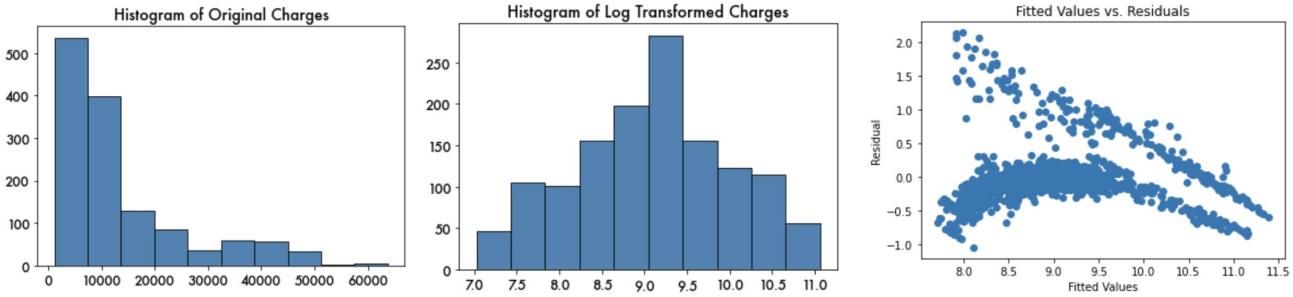


Figure 13 *Results of the logarithmic transformation. Shows the histogram of the original response values (left), histogram of the log transformed response values (middle) and the fitted vs. residual plot for the log transformed values (right).*

We had hypothesized that since the residual plot doesn't change proportionally with the fitted values, the logarithmic transformation will likely not work well for the stabilization of variance. However, since the distribution of Y is positively skewed, the logarithmic transformation helps to normalize the distribution. After the transformation, we re-ran the Breusch-Pagan test and received a LM statistic of 79.07 and a p-value of 7.5e-14. The logarithmic transformation evidently improved the model; however, there remains a significant heteroscedasticity problem.

4.2.2.1 Robust Linear Models

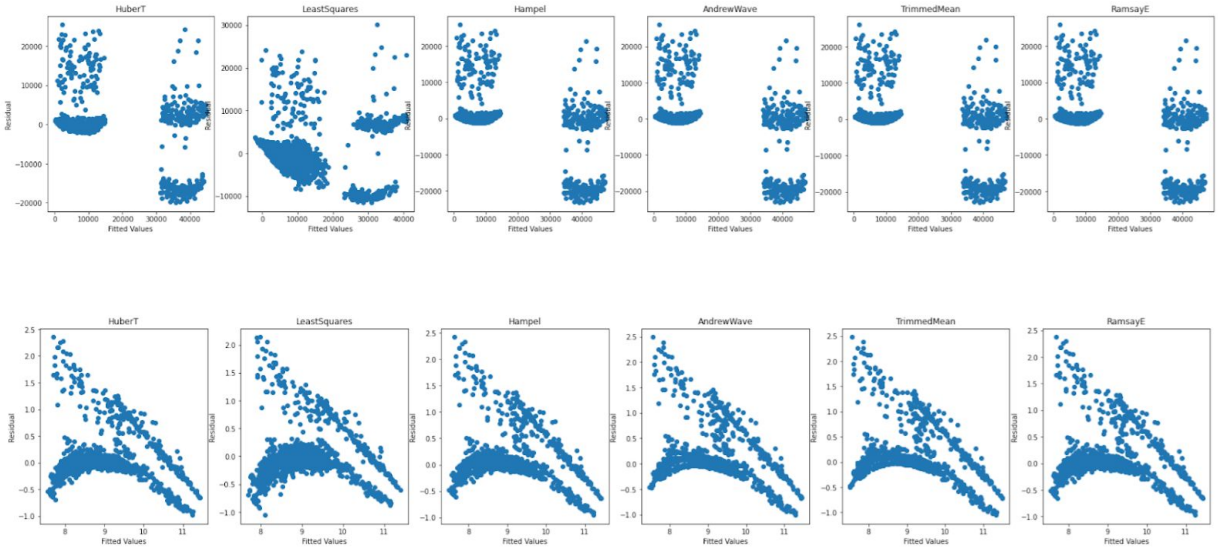
For our final attempt at correcting the inconsistencies in the variance of error terms, we used robust regression as a compromise between excluding the influential points entirely and treating them as equal to all other data points. Below is the resulting table for our parameter selection process for both the original data and the log transformed response variable data. We compared various robust linear model functions for downweighting the outliers.

Figure 14 *Results of the Breush-Pagan test for the various robust regression models*

	Norm	LM Statistic	LM-Test p-value
0	HuberT	309.642378	3.645811e-62
1	LeastSquares	124.395810	4.093354e-23
2	Hampel	405.376752	1.325366e-82
3	AndrewWave	397.864641	5.361933e-81
4	TrimmedMean	389.789306	2.859421e-79
5	RamsayE	400.149307	1.740344e-81

	Norm	LM Statistic	LM-Test p-value
0	HuberT	64.046041	7.448954e-11
1	LeastSquares	79.070325	7.520502e-14
2	Hampel	73.762244	8.725142e-13
3	AndrewWave	78.557263	9.536643e-14
4	TrimmedMean	70.067692	4.764260e-12
5	RamsayE	66.998288	1.940415e-11

Figure 15 *Fitted vs. Residual Plot with respect to the various downweighting functions in the robust regression model*



As seen from **Figure 14,** the robust linear model function that resulted in the best model was Huber; however, the residual vs. fitted plots (**Figure 15**) do not reveal a horizontal band around the 0 mean and the Breusch-Pagan test still failed. Therefore, none of the robust linear models fixed the heteroscedasticity issue. To address this, we created two different models for the population of smokers and nonsmokers based on our initial exploratory analysis.

5. Model Selection Results for Partitioned Data

5.1 Initial Model Fitting

For both datasets (smoker and non-smoker), we regressed health insurance charges against age, sex, body mass index, children and region utilizing ordinary least squares.

5.1.1 Individual Independent T-Test. As indicated in **Figure 16,** for the smoker population, only age and bmi come out as significant predictors of health insurance charges. As for the nonsmoker population, the best model according to the individual t-test is health insurance cost regressed against age, region, children.

Figure 16 *Individual Independent T-Tests for smokers (left) and non-smokers (right) from an OLS model*

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.239e+04	2076.619	-10.780	0.000	-2.65e+04	-1.83e+04
sex[T.male]	-378.5339	740.967	-0.511	0.610	-1837.701	1080.633
children[T.1]	-144.2168	738.795	-0.195	0.845	-1599.107	1310.674
region[T.northwest]	-412.2586	1067.754	-0.386	0.700	-2514.958	1690.441
region[T.southeast]	-1841.6078	997.094	-1.847	0.066	-3805.159	121.944
region[T.southwest]	-439.5869	1067.763	-0.412	0.681	-2542.305	1663.131
age	271.4410	26.589	10.209	0.000	219.080	323.802
bmi	1469.5383	60.417	24.323	0.000	1350.560	1588.517

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2294.7649	855.371	-2.683	0.007	-3973.203	-616.326
sex[T.male]	-517.9605	285.355	-1.815	0.070	-1077.894	41.973
children[T.1]	1216.0355	288.546	4.214	0.000	649.840	1782.231
region[T.northwest]	-581.7301	407.580	-1.427	0.154	-1381.498	218.038
region[T.southeast]	-1049.4136	416.237	-2.521	0.012	-1866.168	-232.659
region[T.southwest]	-1407.3913	407.534	-3.453	0.001	-2207.069	-607.714
age	264.8890	10.245	25.854	0.000	244.785	284.993
bmi	20.6755	24.770	0.835	0.404	-27.928	69.279

5.1.2 Sequential and Partial ANOVA Tests Based on the sequential ANOVA test for the smoker population (**Figure 17.1)**, the best model would be charges regressed against sex, region, age and BMI. This is divergent from the conclusion in the individual t-test which only takes age and bmi as significant predictors. According to the partial ANOVA test (**Figure 17.2**), only age and BMI are significant predictors which coincides with the individual t-test.

Figure 17 *ANOVA Results for Smoker Population*

	df	sum_sq	mean_sq	F	PR(>F)
sex	1.0	3.514863e+08	3.514863e+08	10.322102	1.482891e-03
children	1.0	9.850627e+07	9.850627e+07	2.892835	9.018856e-02
region	3.0	1.271766e+09	4.239219e+08	12.449318	1.260568e-07
age	1.0	4.470341e+09	4.470341e+09	131.280539	8.251155e-25
bmi	1.0	2.014555e+10	2.014555e+10	591.614563	1.659852e-68
Residual	256.0	8.717264e+09	3.405181e+07	NaN	NaN

Figure 17.1 *Sequential ANOVA test results*

	sum_sq	df	F	PR(>F)
sex	8.886940e+06	1.0	0.260983	6.098860e-01
children	1.297548e+06	1.0	0.038105	8.453871e-01
region	1.379231e+08	3.0	1.350130	2.586160e-01
age	3.548813e+09	1.0	104.218035	9.484256e-21
bmi	2.014555e+10	1.0	591.614563	1.659852e-68
Residual	8.717264e+09	256.0	NaN	NaN

Figure 17.2 *Partial ANOVA test results*

Figure 18 *ANOVA Results for Non smoker Population*

	df	sum_sq	mean_sq	F	PR(>F)
sex	1.0	1.214951e+08	1.214951e+08	5.675404	1.738228e-02
children	1.0	5.571611e+08	5.571611e+08	26.026680	3.996768e-07
region	3.0	2.577836e+08	8.592785e+07	4.013950	7.456212e-03
age	1.0	1.468851e+10	1.468851e+10	686.144956	1.066699e-116
bmi	1.0	1.491545e+07	1.491545e+07	0.696746	4.040700e-01
Residual	1046.0	2.239204e+10	2.140731e+07	NaN	NaN

Figure 18.1 *Sequential ANOVA test results*

	sum_sq	df	F	PR(>F)
sex	7.053164e+07	1.0	3.294746	6.978814e-02
children	3.802098e+08	1.0	17.760749	2.721238e-05
region	2.816267e+08	3.0	4.385212	4.467105e-03
age	1.430978e+10	1.0	668.453027	2.308547e-114
bmi	1.491545e+07	1.0	0.696746	4.040700e-01
Residual	2.239204e+10	1046.0	NaN	NaN

Figure 18.2 *Partial ANOVA test results*

According to the sequential ANOVA test, sex, children, region and age are all significant predictors for health insurance cost in the nonsmoker population. The partial ANOVA test results in a similar conclusion except sex is no longer a significant predictor.

5.2 Checking Underlying Data and Model Assumptions

5.2.1 Multicollinearity Detection.

Figure 19 *VIF Scores for smokers (left) and nonsmokers (right)*

	VIF Factor	features
0	33.433144	Intercept
1	1.034587	sex[T.male]
2	1.025799	children[T.1]
3	1.457830	region[T.northwest]
4	1.692959	region[T.southeast]
5	1.496411	region[T.southwest]
6	1.026407	age
7	1.129563	bmi

	VIF Factor	features
0	36.023646	Intercept
1	1.001358	sex[T.male]
2	1.004069	children[T.1]
3	1.527743	region[T.northwest]
4	1.637164	region[T.southeast]
5	1.535168	region[T.southwest]
6	1.021533	age
7	1.109714	bmi

Since there was a lack of significant multicollinearity in the entire dataset, we hypothesized that there wouldn't be an issue once partitioned into two datasets. However, we calculated the VIF scores for each dataset in case new trends arose from the partition.

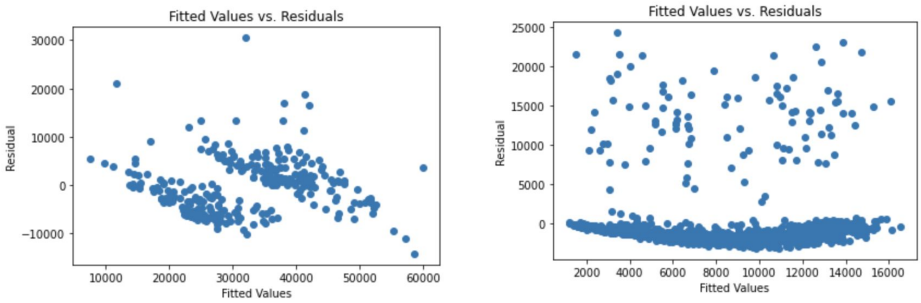
In **Figure 19,** the VIF scores for each predictor variable in both subsets are ~1-2 indicating that there is no significant multicollinearity in both datasets.

5.2.2 Heteroscedasticity Detection.

In **Figure 20,** we created the fitted values vs. residuals plot for both distributions to see if there is a horizontal band around the 0 line. For the smoker population, the fitted values vs. residuals plot looks a lot better. While it may not be a perfectly horizontal band, we think that the shape is a good indication of there not being a significant heteroscedasticity problem.

However, for the nonsmoker population, **Figure 20** shows that the nonsmoker model may possibly be missing a binary predictor value and should be separated into two subgroups. There is a clear horizontal band around 0 for the data clustered below but a lot of points with high error lingering at the top.

Figure 20 *Fitted values vs. residuals plot for smokers (left) and nonsmokers (right)*

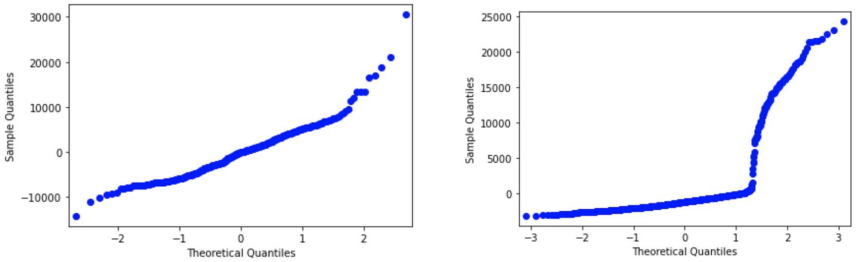


Once we ran through the Breusch-Pagan test for the smoker population, we received a p-value of 0.808. This indicates that we do not have a significant heteroscedasticity issue for the smoker population. Similarly, we receive a p-value of 0.274 for the nonsmoker population and conclude that we do not have a significant heteroscedasticity issue for the nonsmoker population.

5.2.3 Normality of Residuals Detection

The Jarque-Bera test and the Omnibus test for both populations both resulted in a p-value that was less than 0.05, suggesting that there is a significant violation to our model assumption "normality of residuals". We plotted the QQ plot to determine whether or not the model has a normal distribution. For the smoker population, since the resulting plot is approximately linear on the diagonal of the plot, then it suggests that the error term of the model does follow a normal distribution. Conversely, the QQ-plot for the nonsmoker population does not follow a straight line on the diagonal of the plot. It also seems to look bimodal. This further confirms our suspicions that there could be a missing binary predictor value.

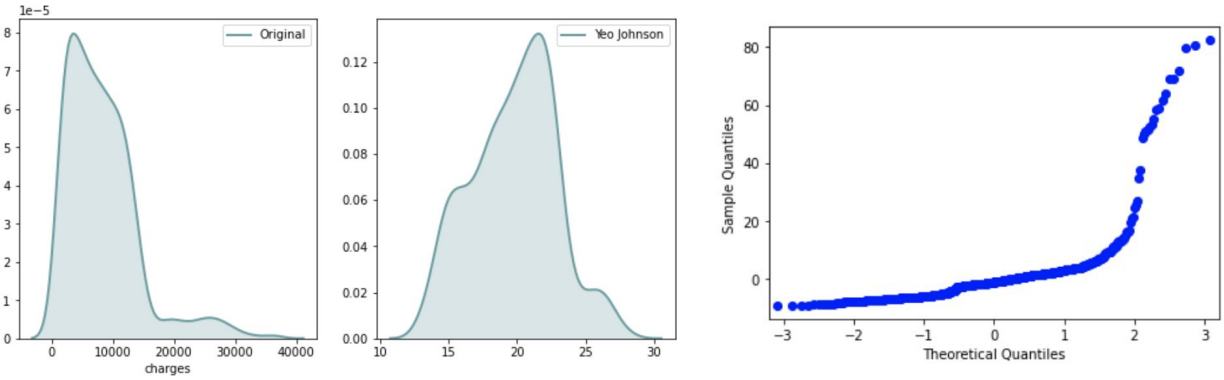
Figure 21 *QQ plot for smokers (left) and nonsmokers (right)*



5.2.3 Normality of Residuals for Nonsmoker Population Solution

We implemented the Yeo-Johnson method which is a technique that is similar to a Box-cox transformation, however, it does not require that the values be strictly positive.

Figure 22 Nonsmoker data. *Distribution plots for the original and the yeo-johnson transformed plots (left and center), QQ plot (right)*



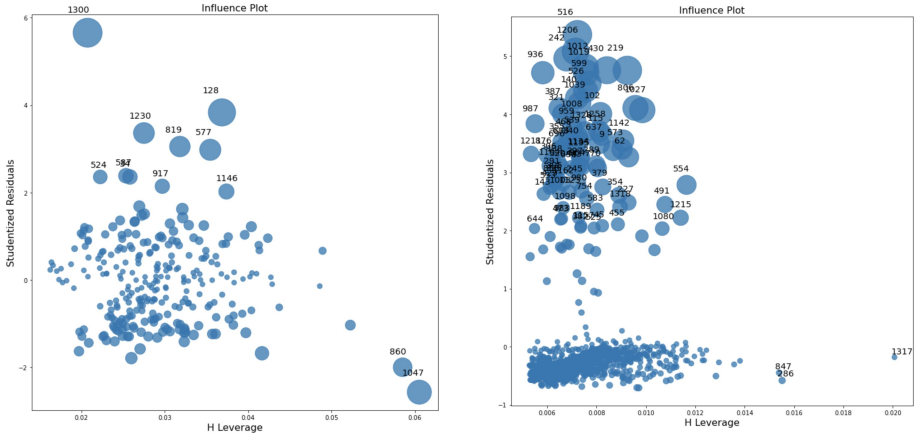
Omnibus:	1013.934	**Durbin-Watson:**	2.020
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	45448.172
Skew:	5.074	**Prob(JB):**	0.00
Kurtosis:	34.888	**Cond. No.**	313.

As shown from the QQ plot and the Jarque-Bera and Omnibus tests, the Yeo-Johnson did not help to significantly fix the non-normality of residuals issue. We will revisit this by conducting more exploratory data analysis. However, for now, since our sample size is greater than 30 times # of predictors, we can assume normality based on the central limit theorem.

5.2.4 Effects of Influential Points on Models

For our partitioned datasets, we ran through the process of dropping the intersection of the points identified by the externalized studentized residual test and the Cook's distance test once again in order to examine the effects of the influential points on our model assumptions. **Figure 23** visually illustrates the influence plot for both partitions.

Figure 23 *Influence plot for smokers (left) and nonsmokers (right) where x-axis is the leverage, y-axis is the external studentized residuals, and the size of the points represent Cook's distance.*



The first thing we investigated was the effect on heteroscedasticity and the model fit. For the nonsmoker population, the Breush-Pagan test results in the p-value decreasing from 0.274 to 0.256 once we drop the influential points. However, the adjusted r-squared value goes from 0.409 to 0.889, resulting in a significant increase in model fit after dropping the influential points. We still reject the null for the Omnibus and Jarque-Bera tests which we can safely disregard since we have enough data to assume normality of residuals. For the smoker population, there is a dramatic decrease in the p-value for the Breush-Pagan test from 0.808 to 0.353. Similar to the results found in the nonsmoker population, the adjusted R-squared value increases from 0.75 to 0.823

5.2.4.1 Linearity Test Comparison

Figure 24 *Scatterplots between residuals and predictor variables*

Fig 24.1 *Smokers with influential points*

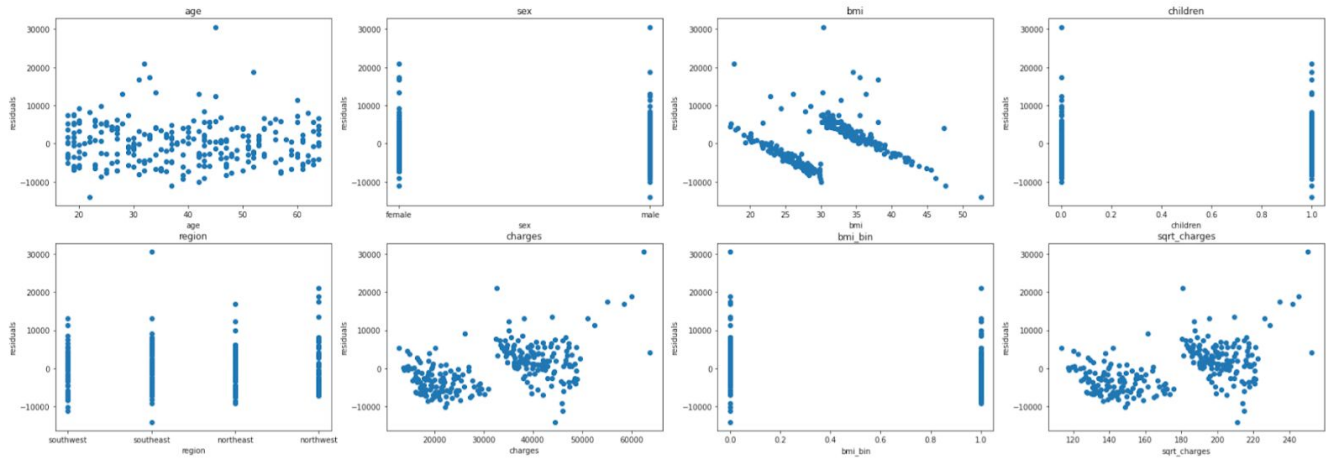


Fig 24.2 *Smokers without influential points*

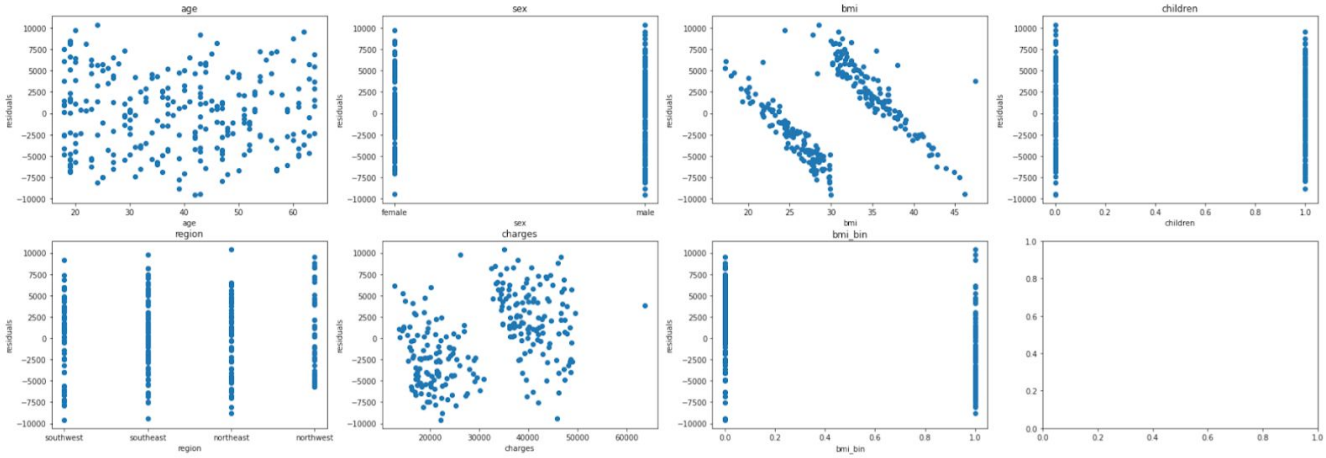


Fig 24.3 *Nonsmokers with influential points*

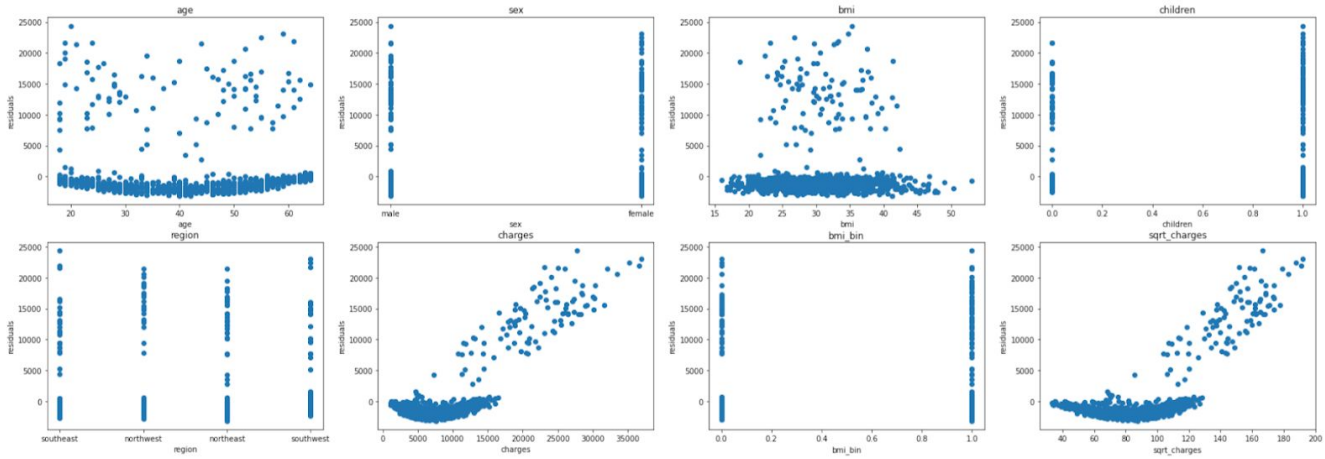
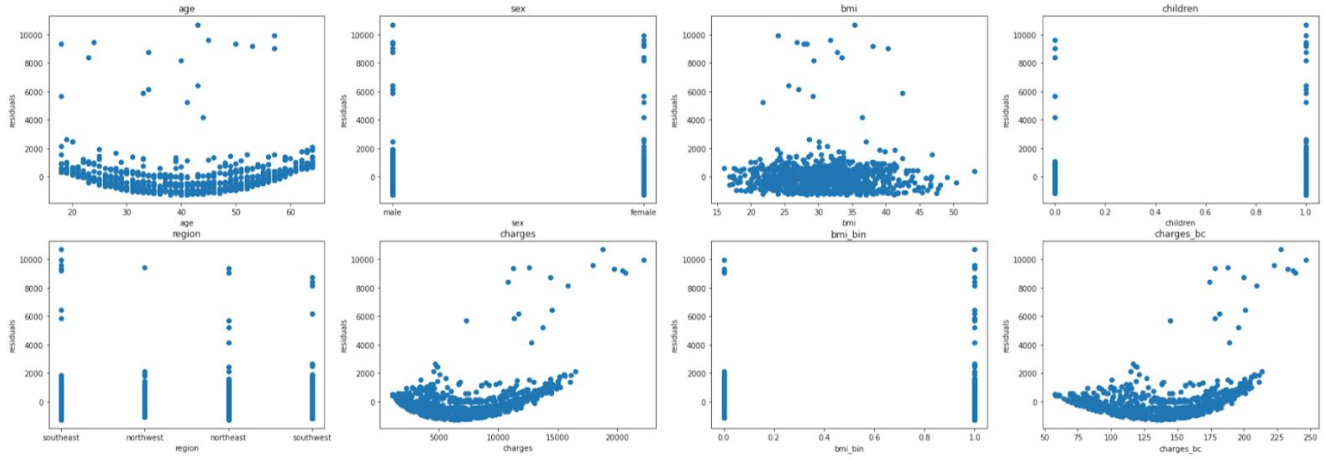


Fig 24.4 *Nonsmokers without influential points*



Finally, we tested for non-linearity using a scatterplot between the residuals and the predictor variables. We found that in most of the models, age and bmi do not seem to have a linear relationship with the response variable. Per their scatterplots, there does not seem to be an easy fix, thus we are reluctant to try complex solutions like a polynomial regression as we may lose inference and add multicollinearity. In this case, we will leave the predictors as they are and make a statement about the non-linear relationship in our conclusion. Thus, our predictions with our final model will probably have some amount of error, especially when we extrapolate beyond the range of the sample data.

5.3 Final Model Selection

During our model selection process we first looked at the individual t-values from the summary table for each predictor variable. If the p-value of that predictor variable is below 0.05, it suggests the significance of the predictor in the full model. This gives us a good idea of what predictor variables may end up in our final model, but not a definite conclusion. From there we moved onto using the sequential ANOVA test and partial ANOVA test results to further analyze the model.

The sequential ANOVA test indicates the significance of a predictor variable, given the predictors before it in the table are already in the model. However, since the conclusion changes when we change the order of the table, we prefer the partial ANOVA test. This tests the significance of a predictor variable given all the rest of the predictors are already in the model. This will give us an even better idea of what our final, best model is as it is a more complex analysis of the significance of the predictor variables. However, we still need to use one more selection process to finalize our results.

The previous tests will give us a general idea of which predictor variables may be included in our final model, however, the most important step in the model selection process is analyzing the Adjusted R-squared, Mallow's CP, AIC, and BIC values for each possible model. The Adjusted R-squared and Mallow Cp values give us an idea of what initial candidate models we should analyze. The Mallow Cp value measures the mean squared error of the prediction, so for models with the same number of predictors, we are looking for ones with a Cp value that is close to, or equal to the number of predictors included in the model

(we can not use this when evaluating the full model). We use this in tandem with models that have a high Adjusted R-squared value to find the subset we want to test.

We then calculate the AIC and BIC scores for each model. Both of these values measure the amount of information lost by a given model. Thus, we are looking for the models with the lowest AIC and BIC scores as this indicates a higher quality model. In our analysis we place more importance on a lower BIC score compared to a lower AIC score, as the BIC has a larger penalty for the number of parameters. However, in some cases, like with the Non-smokers model (with the influential points included), we prefer a model with the second lowest BIC score as it has the lowest AIC score and matches the final model for the Non-smokers model without the influential points.

For each subset of our regressions (including the models with and without the influential points), we made the following conclusions on the best model:

Figure 25 *Results of significance from all tests for the* ***original smoker*** *data*

Fig 25.1 *Individual Independent T-test*
Best Model: charges ~ age + bmi

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.191e+04	2062.919	-10.621	0.000	-2.6e+04	-1.78e+04
sex[T.male]	-470.5580	736.968	-0.639	0.524	-1921.850	980.735
children[T.1]	-69.3894	728.149	-0.095	0.924	-1503.314	1364.535
region[T.northwest]	-811.5576	1054.437	-0.770	0.442	-2888.033	1264.918
region[T.southeast]	-1846.5982	990.048	-1.865	0.063	-3796.274	103.077
region[T.southwest]	-539.4356	1074.030	-0.502	0.616	-2654.494	1575.623
age	266.0817	25.839	10.298	0.000	215.198	316.965
bmi	1461.7217	59.539	24.551	0.000	1344.474	1578.970

Fig 25.2 *Sequential ANOVA T-test*
Best Model: charges ~ sex+region+age + bmi

	df	sum_sq	mean_sq	F	PR(>F)
sex	1.0	2.611407e+08	2.611407e+08	7.782206	5.673011e-03
children	1.0	5.120797e+07	5.120797e+07	1.526039	2.178403e-01
region	3.0	1.301563e+09	4.338545e+08	12.929218	6.841652e-08
age	1.0	4.725949e+09	4.725949e+09	140.837154	3.561830e-26
bmi	1.0	2.022563e+10	2.022563e+10	602.740251	3.118547e-69
Residual	256.0	8.590368e+09	3.355613e+07	NaN	NaN

Fig 25.3 *Partial ANOVA T-test*
Best Model: charges ~ age + bmi

	sum_sq	df	F	PR(>F)
sex	1.368045e+07	1.0	0.407689	5.237158e-01
children	3.047316e+05	1.0	0.009081	9.241546e-01
region	1.272951e+08	3.0	1.264499	2.870425e-01
age	3.558439e+09	1.0	106.044402	4.934691e-21
bmi	2.022563e+10	1.0	602.740251	3.118547e-69
Residual	8.590368e+09	256.0	NaN	NaN

Fig 25.4 *Adjusted R2, Mallows CP, AIC and BIC*
Best Model: charges ~ age + bmi

	Number of predictors	Adj_R2	Cp	AIC	BIC	Predictors
5	2	0.749762	2.173535	5326.129806	5336.857653	age, bmi
21	3	0.750500	2.412817	5328.298499	5349.754193	age, bmi, region
7	3	0.749166	3.794278	5327.744690	5342.048487	age, sex, bmi
23	4	0.749935	4.001370	5329.869919	5354.901563	age, sex, bmi, region
13	3	0.748804	4.169553	5328.125765	5342.429562	age, bmi, children
29	4	0.749546	4.403079	5330.280649	5355.312293	age, bmi, children, region
15	4	0.748198	5.793497	5329.743897	5347.623642	age, sex, bmi, children
31	5	0.748968	6.000000	5331.860555	5360.468147	age, sex, bmi, children, region
4	1	0.648291	107.075509	5415.003548	5422.155446	bmi
12	2	0.648004	107.972943	5416.209392	5426.937239	bmi, children
6	2	0.647577	108.416072	5416.528856	5427.256704	sex, bmi
20	2	0.647451	108.547268	5418.592586	5436.472332	bmi, region
14	3	0.647382	109.214509	5417.661716	5431.965513	sex, bmi, children
28	3	0.647042	109.566092	5419.877120	5441.332815	bmi, children, region
22	3	0.646687	109.934380	5420.142952	5441.598647	sex, bmi, region

For the Smokers model with all influential points left in, we started our model selection process by first looking at the individual t-values for each predictor. From **Figure 25.1,** the t-values seem to indicate that only the age and bmi predictor values have a significant impact on the response variable. Thus we took note that a charges ~ age + bmi model may be the best. We then checked the sequential ANOVA test **(Figure 25.2)** and the partial ANOVA test **(Figure 25.3)** for the model. The sequential ANOVA test seems to indicate that a charges ~ sex + region + age + bmi is a good model while the partial ANOVA test seems to support our t-value conclusion of a charges ~ age + bmi being best. Finally, since we had only a small amount of predictor variables, we decided to use a table to display all the Adjusted R-squared, Mallow's Cp, AIC, and BIC values for each possible model **(Figure 25.4)** . Per the table, we decided that our conclusion from the t-values and partial ANOVA table to be correct. We chose as our final, best model to be charges ~ age + bmi as it has the lowest AIC and BIC scores.

Figure 26 *Results of significance from all tests for the* ***original nonsmoker*** *data*

Fig 26.1 *Individual Independent T-test*
Best Model: charges ~ children + region + age

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2258.3929	855.857	-2.639	0.008	-3937.786	-579.000
sex[T.male]	-534.4616	285.180	-1.874	0.061	-1094.052	25.129
children[T.1]	1243.4114	288.143	4.315	0.000	678.008	1808.815
region[T.northwest]	-580.8801	406.923	-1.427	0.154	-1379.358	217.598
region[T.southeast]	-1039.4280	415.911	-2.499	0.013	-1855.543	-223.313
region[T.southwest]	-1419.7125	406.961	-3.489	0.001	-2218.266	-621.159
age	264.3445	10.236	25.824	0.000	244.258	284.431
bmi	19.9868	24.785	0.806	0.420	-28.647	68.621

Fig 26.2 *Sequential ANOVA T-test*
Best Model: charges ~ sex+children+region+age

	df	sum_sq	mean_sq	F	PR(>F)
sex	1.0	2.611407e+08	2.611407e+08	7.782206	5.673011e-03
children	1.0	5.120797e+07	5.120797e+07	1.526039	2.178403e-01
region	3.0	1.301563e+09	4.338545e+08	12.929218	6.841652e-08
age	1.0	4.725949e+09	4.725949e+09	140.837154	3.561830e-26
bmi	1.0	2.022563e+10	2.022563e+10	602.740251	3.118547e-69
Residual	256.0	8.590368e+09	3.355613e+07	NaN	NaN

Fig 26.3 *Partial ANOVA T-test*
Best Model: sex+children+region

	sum_sq	df	F	PR(>F)
sex	1.368045e+07	1.0	0.407689	5.237158e-01
children	3.047316e+05	1.0	0.009081	9.241546e-01
region	1.272951e+08	3.0	1.264499	2.870425e-01
age	3.558439e+09	1.0	106.044402	4.934691e-21
bmi	2.022563e+10	1.0	602.740251	3.118547e-69
Residual	8.590368e+09	256.0	NaN	NaN

Fig 26.4 *Adjusted R2, Mallows CP, AIC and BIC*
Best Model: charges ~ sex+children+region+age

	Number of predictors	Adj_R2	Cp	AIC	BIC	Predictors
27	4	0.407557	4.649623	20787.853057	20822.575492	age, sex, children, region
31	5	0.407359	6.000000	20789.197997	20828.880778	age, sex, bmi, children, region
25	3	0.406168	6.110407	20789.327690	20819.089776	age, children, region
29	4	0.405937	7.517123	20790.731254	20825.453688	age, bmi, children, region
11	3	0.402037	13.429053	20794.643630	20814.485021	age, sex, children
9	2	0.400728	14.758896	20795.951215	20810.832258	age, children
15	4	0.401515	15.344529	20796.559536	20821.361274	age, sex, bmi, children
13	3	0.400195	16.692971	20797.885831	20817.727222	age, bmi, children
19	3	0.397550	21.379405	20804.514176	20834.276263	age, sex, region
23	4	0.397385	22.655167	20805.796857	20840.519291	age, sex, bmi, region
17	2	0.396156	22.867176	20805.954528	20830.756266	age, region
21	3	0.395956	24.202669	20807.298371	20837.060458	age, bmi, region
3	2	0.392607	29.161080	20810.138783	20825.019826	age, sex
1	1	0.391295	30.516059	20811.416521	20821.337217	age
7	3	0.392095	31.044150	20812.024362	20831.865752	age, sex, bmi

For the non-smokers model with all data left in, we began our model selection process by looking at the individual t-values for each predictor. From the summary table above, the t-values indicate that only children, region, and age have a significant impact on the response variable. Thus, we chose charges ~ children + region + age as our model. We then checked the sequential and partial ANOVA tests for the model. The sequential ANOVA test seems to indicate that a charges ~ sex + children + region is a good model while the partial ANOVA test seems to conclude that charges ~ sex + children + region + age may be best. Finally we looked at a combination of Adjusted R-squared, Mallow's Cp, AIC, and BIC values for each possible model to make our final conclusion. Per the table, we chose our final, best model to be charges ~ sex + children + region + age as it has the lowest AIC and one of the lowest BIC scores. While charges ~ children + region + age has a lower BIC value, our chosen model has a higher Adjusted R-squared value, lower AIC score, and is supported by our partial ANOVA test.

Figure 27 *Results of significance from all tests for the* ***smoker dropped*** *data*

Fig 27.1 *Individual Independent T-test*
Best Model: charges ~ bmi + region + age

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.29e+04	1670.570	-13.708	0.000	-2.62e+04	-1.96e+04
sex[T.male]	-518.1649	593.459	-0.873	0.383	-1687.097	650.767
children[T.1]	63.3954	588.958	0.108	0.914	-1096.672	1223.463
region[T.northwest]	-1705.0651	854.959	-1.994	0.047	-3389.073	-21.058
region[T.southeast]	-1880.0160	795.994	-2.362	0.019	-3447.881	-312.151
region[T.southwest]	-661.3133	863.081	-0.766	0.444	-2361.319	1038.692
age	254.4690	20.780	12.246	0.000	213.538	295.400
bmi	1494.4271	49.189	30.381	0.000	1397.539	1591.315

Fig 27.2 *Sequential ANOVA T-test*
Best Model: charges ~ sex + region + age + bmi

	df	sum_sq	mean_sq	F	PR(>F)
sex	1.0	2.212802e+08	2.212802e+08	10.573837	1.308150e-03
children	1.0	5.302417e+07	5.302417e+07	2.533751	1.127244e-01
region	3.0	1.499917e+09	4.999723e+08	23.891093	1.346043e-13
age	1.0	4.703645e+09	4.703645e+09	224.762893	1.709222e-36
bmi	1.0	1.931614e+10	1.931614e+10	923.018461	4.653978e-85
Residual	245.0	5.127150e+09	2.092714e+07	NaN	NaN

Fig 27.3 *Partial ANOVA T-test*
Best Model: charges ~ age + bmi

	sum_sq	df	F	PR(>F)
sex	1.595383e+07	1.0	0.762351	3.834487e-01
children	2.424692e+05	1.0	0.011586	9.143694e-01
region	1.492707e+08	3.0	2.377625	7.046766e-02
age	3.138124e+09	1.0	149.954739	3.237261e-27
bmi	1.931614e+10	1.0	923.018461	4.653978e-85
Residual	5.127150e+09	245.0	NaN	NaN

Fig 27.4 *Adjusted R2, Mallows CP, AIC and BIC*
Best Model: charges ~ age + bmi

	Number of predictors	Adj_R2	Cp	AIC	BIC	Predictors
21	3	0.830302	2.754616	4987.351110	5008.551446	age, bmi, region
23	4	0.830134	4.003550	4988.574718	5013.308444	age, sex, bmi, region
29	4	0.829614	4.760419	4989.348775	5014.082501	age, bmi, children, region
5	2	0.827604	5.704326	4988.396120	4998.996288	age, bmi
31	5	0.829449	6.000000	4990.562753	5018.829869	age, sex, bmi, children, region
7	3	0.827273	7.176401	4989.867024	5004.000582	age, sex, bmi
13	3	0.826928	7.679815	4990.371579	5004.505137	age, bmi, children
15	4	0.826607	9.132875	4991.823353	5009.490300	age, sex, bmi, children
28	3	0.726006	155.023564	5108.559526	5129.759863	bmi, children, region
30	4	0.726179	155.165754	5109.373533	5134.107259	sex, bmi, children, region
12	2	0.725431	155.472730	5106.144144	5116.744312	bmi, children
20	2	0.725215	155.790214	5108.311500	5125.978448	bmi, region
14	3	0.725467	155.811220	5107.097533	5121.231091	sex, bmi, children
22	3	0.725186	156.220936	5109.315688	5130.516025	sex, bmi, region
4	1	0.724457	156.516201	5106.050115	5113.116894	bmi

For the Smokers model with all influential points removed, we started our model selection process by first looking at the individual t-values for each predictor. From the summary table above, the t-values seem to indicate that only the children, region, and age predictor values have a significant impact on the response variable. Thus we took note that a charges ~ region + age + bmi model may be the best. We then checked the sequential ANOVA test and the partial ANOVA test for the model. The sequential ANOVA test seems to indicate that a charges ~ sex + region + age + bmi is a good model while the partial ANOVA test seems to conclude that charges ~ age + bmi may be best. Finally we looked at a combination of Adjusted R-squared, Mallow's Cp, AIC, and BIC values for each possible model to make our final conclusion. Per the table, we concluded that our partial ANOVA test to be accurate and that our final, best model is charges ~ age + bmi as it has the lowest AIC and BIC scores.

Figure 28 *Results of significance from all tests for the* ***nonsmoker dropped*** *data*

Fig 28.1 *Individual Independent T-test*
Best Model: charges ~ sex + children + region + age

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3313.4701	258.345	-12.826	0.000	-3820.453	-2806.488
sex[T.male]	-479.2049	86.578	-5.535	0.000	-649.108	-309.302
children[T.1]	760.2767	87.426	8.696	0.000	588.711	931.843
region[T.northwest]	-455.5915	124.282	-3.666	0.000	-699.485	-211.698
region[T.southeast]	-637.7596	127.694	-4.994	0.000	-888.349	-387.170
region[T.southwest]	-724.2268	123.648	-5.857	0.000	-966.876	-481.578
age	267.0626	3.128	85.376	0.000	260.924	273.201
bmi	8.5499	7.491	1.141	0.254	-6.151	23.251

Fig 28.2 *Sequential ANOVA T-test*
Best Model: charges ~ sex + children + region + age

	df	sum_sq	mean_sq	F	PR(>F)
sex	1.0	1.082798e+08	1.082798e+08	59.446605	3.116537e-14
children	1.0	2.632721e+08	2.632721e+08	144.538823	3.983990e-31
region	3.0	1.189056e+08	3.963521e+07	21.760097	1.216071e-13
age	1.0	1.357623e+10	1.357623e+10	7453.477047	0.000000e+00
bmi	1.0	2.372742e+06	2.372742e+06	1.302657	2.540113e-01
Residual	966.0	1.759533e+09	1.821463e+06	NaN	NaN

Fig 28.3 *Partial ANOVA T-test*
Best Model: charges ~ sex + children + region + age

	sum_sq	df	F	PR(>F)
sex	5.580164e+07	1.0	30.635615	4.010654e-08
children	1.377481e+08	1.0	75.624981	1.454323e-17
region	7.192255e+07	3.0	13.162048	1.989243e-08
age	1.327684e+10	1.0	7289.109888	0.000000e+00
bmi	2.372742e+06	1.0	1.302657	2.540113e-01
Residual	1.759533e+09	966.0	NaN	NaN

Fig 28.4 *Adjusted R2, Mallows CP, AIC and BIC*
Best Model: charges ~ sex + children + region + age

	Number of predictors	Adj_R2	Cp	AIC	BIC	Predictors
27	4	0.887998	5.303283	16811.728485	16845.898364	age, sex, children, region
31	5	0.888033	6.000000	16812.415924	16851.467215	age, sex, bmi, children, region
25	3	0.884597	33.765207	16839.868956	16869.157424	age, children, region
29	4	0.884601	34.696909	16840.825556	16874.995435	age, bmi, children, region
11	3	0.883936	39.493266	16843.443791	16862.969436	age, sex, children
15	4	0.883817	41.486145	16845.436893	16869.843950	age, sex, bmi, children
9	2	0.880650	67.021094	16869.634832	16884.279066	age, children
13	3	0.880527	69.020746	16871.634505	16891.160150	age, bmi, children
19	3	0.879353	79.194418	16883.151309	16912.439777	age, sex, region
23	4	0.879392	79.779324	16883.829636	16917.999515	age, sex, bmi, region
3	2	0.876024	107.144596	16906.679318	16921.323551	age, sex
17	2	0.876020	107.179981	16908.703125	16933.110182	age, region
21	3	0.876027	108.008743	16909.639456	16938.927924	age, bmi, region
7	3	0.875900	109.107420	16908.645639	16928.171284	age, sex, bmi
1	1	0.872794	134.289595	16930.730962	16940.493784	age

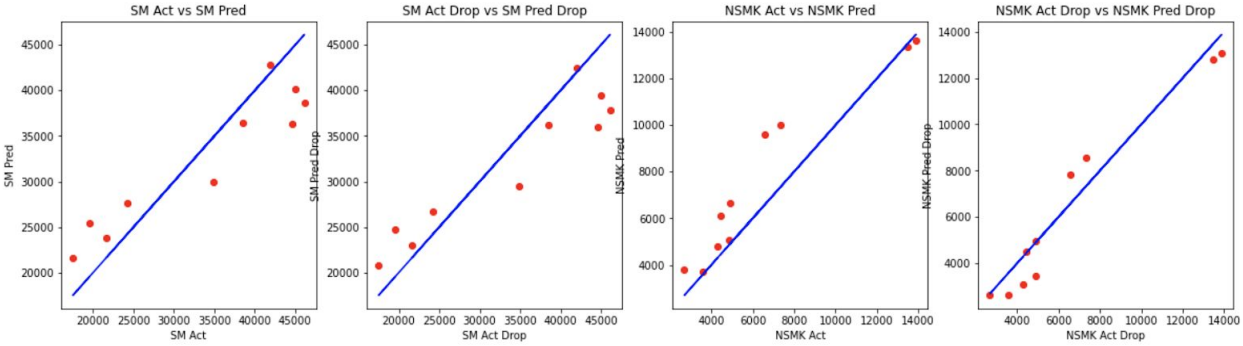
For the non-smokers model with all influential points removed, we started our model selection process by first looking at the individual t-values for each predictor. From the summary table above, the t-values seem to indicate that the sex, children, region, and age predictor values have a significant impact on the response variable. Thus we took note that a charges ~ sex + children + region + age model may be the best. We then checked the sequential ANOVA test and the partial ANOVA test for the model. The sequential ANOVA test seems to indicate that a charges ~ sex + children + region + age is a good model and the partial ANOVA test seems to come to the same conclusion. Finally we looked at a combination of Adjusted R-squared, Mallow's Cp, AIC, and BIC values for each possible model to make our final conclusion. Per the table, we found all of our previous conclusions to be accurate and that our final, best model is charges ~ sex + children + region + age as it has the lowest AIC and BIC scores.

5.4 Final Model Performance

Figure 29 *Model predictions on testing dataset and their respective r-squared and adjusted r-squared values*

	SM Act	SM Pred	SM Act Drop	SM Pred Drop	NSMK Act	NSMK Pred	NSMK Act Drop	NSMK Pred Drop
0	19539.24	25467.89	19539.24	24713.82	3597.60	3704.39	3597.60	2639.35
1	17496.31	21578.65	17496.31	20813.00	4296.27	4784.73	4296.27	3107.02
2	38511.63	36398.99	38511.63	36187.90	2690.11	3804.33	2690.11	2613.74
3	24180.93	27596.21	24180.93	26698.13	13462.52	13340.68	13462.52	12790.18
4	41949.24	42821.68	41949.24	42507.05	4877.98	5050.97	4877.98	3432.39
5	21659.93	23845.22	21659.93	23011.97	6555.07	9613.84	6555.07	7842.70
6	45008.96	40180.00	45008.96	39395.43	4906.41	6642.79	4906.41	4979.85
7	46130.53	38612.20	46130.53	37786.29	7345.73	9984.27	7345.73	8581.35
8	34838.87	29946.78	34838.87	29492.77	13880.95	13606.12	13880.95	13057.72
9	44585.46	36319.91	44585.46	35997.57	4433.39	6112.72	4433.39	4502.58
R2	NaN	0.79	NaN	0.79	NaN	0.83	NaN	0.94
adjR2	NaN	0.73	NaN	0.72	NaN	0.70	NaN	0.89

Figure 30 *Actual response values plotted against forecasted response values*



In **Figure 29** we display the actual response values preserved from our test data and the respective forecasted values for each of the four models (smoker all data, smoker dropped influential points, nonsmoker all data, nonsmoker dropped influential points). At the bottom of the table, we output the r-squared and adjusted r-squared values for each model. From this table, we see that the model containing all data points has a better fit to the test data than the model with the influential points dropped. Conversely, for the nonsmoker population, the model with the influential points dropped outperform the model contains all data points.

6. Conclusions

Research question 1: What are the differences in significant predictors of health insurance cost for individuals who smoke and those who do not?

This question was answered by obtaining the regression weights and the p-values for these weights. In **Figure 28,** the regression weights, standard errors and p-values for all the predictors are provided. For the nonsmoker population BMI was not a significant predictor of health insurance charges ($\beta = 8.55$, $p = 0.254$). However, AGE, SEX, CHILDREN and REGION were all significant predictors. For the smoker population, (**Figure 25)**, the significant predictors were AGE and BMI ($\beta = 266.08$, $p=0$, $\beta = 1461.72$, $p=0$, respectively). SEX, CHILDREN, and REGION were not significant predictors. We also know from

our exploratory analysis that, on average, the smoker population pays significant more than the nonsmoker population. The only overlap in significant predictors between the two populations is AGE.

Research question 2: For a 45 year old with a BMI of 35 in the subset population of smokers, what is the expected cost of their health insurance?

The model containing all data points outperformed the one without the influential points. Therefore, we used this model to forecast the expected health insurance for a 45 year old smoker with a BMI of 35. The forecasted value is $39843.23. From our model performance analysis, we observed that the model received an adjusted r-squared value of 0.73 for the testing data. This means that 73% of the model explains the variation in health insurance costs around the mean.

Research question 3: For a 45 year old male living in the northwest with children in the subset population of nonsmokers, what is the expected cost of their health insurance?

Unlike the smoker population, the model without the influential points outperformed the model containing all data points. Therefore, we used the model without the influential points to forecast the expected health insurance for a 45 year old male living in the northwest with children within the nonsmoker population. The forecasted value is $8783.35 From our model performance analysis, we observed that the model received an adjusted r-squared value of 0.89 for the testing data. This means that 89% of the nonsmoker model with dropped outliers explains the variation in health insurance costs around the mean.

Research question 4: How does the cost of health insurance change for an individual in the nonsmoker population when they have a child?

From our model summary table seen in **Figure 28**, we can conclude that those in the nonsmoker population, who have children pay, on average, $760.83 more on health insurance than those without children.

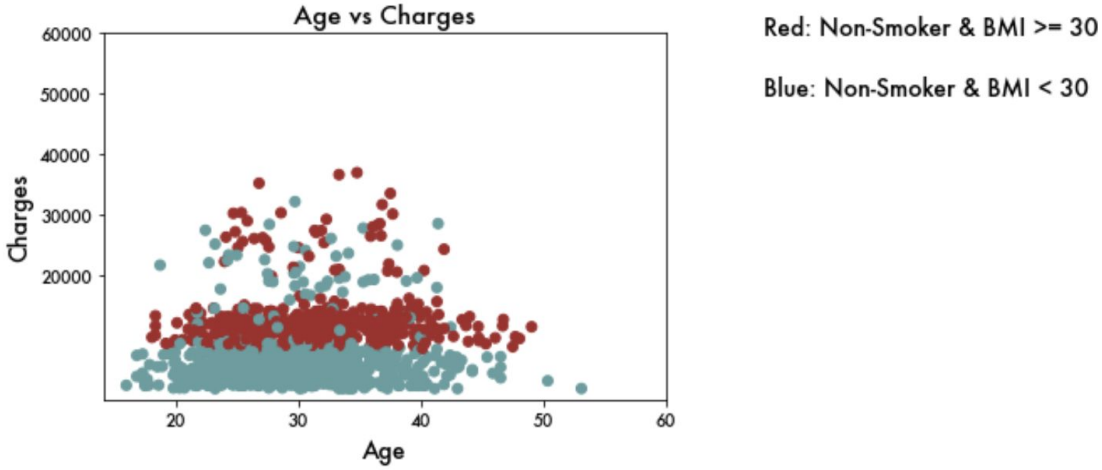
Research question 5: How does health insurance price change in respect to an increase in body mass index within the smoker population?

For those in the smoker population, every increase in bmi score results in a $1425.28 increase in health insurance cost on average (**Figure 26)**.

7. Discussions and Summary

As we were not able to confirm that our models conform to the assumption of linearity, we expect that some of our answers to the research questions may have some form of error. We expect that the errors would be even more serious if we try to extrapolate using our models. If we were to go through the analysis again, we may try a polynomial regression in an effort to fix some of these non-linear relationships.

Figure 31 *Age vs. Charges for the nonsmoker population partitioned by BMI >= 30 and BMI <30*



Furthermore, during the model diagnostics in section 5.2, we noted that the QQ plot for the nonsmoker model appeared to reveal a bimodal distribution. Additionally, **Figure 20**, displayed the fitted value vs. residual plot for the nonsmoker model and we noticed that while there is a clear horizontal band around 0 for a portion of the data, there was still a significant amount of points lingering at the top. We believe that the nonsmoker model is missing a binary predictor and should be further split into two subsets. In **Figure 31,** we go through further exploratory analysis in an attempt to find variables that could potentially partition the data appropriately. If given more time, we would integrate additional

data sources to discover the potential missing binary predictors in the nonsmoker data. We would also examine the effect of dropping the union set of influential points from Cook's distance and the externally studentized residuals rather than dropping the intersection set.

From our regression analysis, arose a few interesting observations. The first is that the significance of the children on health insurance costs differs between smokers and non-smokers. We expected children, as dependents, to have a significant impact on the costs of health insurance. This is true for the non-smoker population, but children were not a significant predictor for smokers. This could be for many reasons, one of which could be that individuals that smoke most likely have to pay large insurance premiums. Insurance companies often charge higher premiums for smokers as they are generally seen as a higher risk than most non-smokers.

For each of our smoker models, both with and without the influential points, age and BMI are the only significant predictors. This makes sense, as one would think that insurance companies would have higher risk individuals (those with poor health), pay more on average than individuals with less risk. The number of children, region, and sex of an individual would thus be less important in determining how much one pays for insurance. However, age and BMI play a significant role in health, as we tend to deteriorate in health as both our age and weight increases. We confirm this relationship as age shows up in every model and BMI shows up in all but one. We can conclude that age and BMI most likely have a higher impact on health insurance costs than sex, region and the number of children on average.

It was interesting to find that for non-smokers, the sex, number of children, region, and age were all significant in prediciting the cost of health insurance. We did not expect the region one lives in to show up as significant, but one could assume that eating habits and pollution could play a part. Some regions could have a higher level of pollution or common eating habits that support an unhealthy way of living. These both could have an impact on whether or not an insurance company views an individual as high risk, which would lead to higher insurance premium payments for those individuals.