Forecasting Medical Expenses and Analyzing Variable Correlations

Charlie Allen, Michal Kryszkowski, Alex Ng, Rikesh Patel, and Kevin Sass

Dr. Gulasingam
DSC323

Introduction

It is well known that risk assessment is one of the essential functions performed in an insurance company. It is at the risk assessment stage that the basic parameters of the insurance process are shaped in the future. Incorrect verification and acceptance of inappropriate risks may lead to deterioration of the financial results of the insurance.

Risk assessment in this type of insurance includes both medical and non-medical risk factors.

Medical risk is related to the health of the insured and is shaped by many factors, including biological factors such as age, lifestyle, genetic factors, and behavior.

Non-medical risk factors include occupational and non-occupational risks. Occupational risks include factors that directly affect human health (dust, noise) as well as indirect effects (stress, rush). Non-professional risk is associated with non-professional activities performed by the insured in his spare time. Recently, it is worth noting the increase in the importance of this type of risk, which may result from the increase in free time and the popularization of active forms of rest.

Among individual risk factors in life insurance, it should be noted that the age of the insured is of particular importance. Expected mortality is closely related to this factor. Thus, as age increases, both the risk of illness and death increase. Because the age of the insured plays a crucial role in setting the tariff, it is worth emphasizing that it is usually not a criterion for

selecting applications for insurance. Only certain age ranges (very young or advanced people) are generally closed for insurance.

The second, also a significant risk factor affecting mortality, is the sex of the insured. Like age, it is not a selection criterion but a classification.

Medical factors are the third risk factor, including in particular:

- physical conditions (height, weight, digestive system, nervous system),
- the health burden of the person concerned (general health),
- family conditions (illness loads, age),
- hereditary loads

Other risk factors include:

- way to spend free time,
- the type of occupation
- sports practiced,
- place of residence,
- marital status,
- education.

Another factor that should be considered in Health Insurance is the type of plan the customer has. There are two main types of Health Insurance, individual vs family. Family has more benefits in terms of lowering the cost of Healthcare. Deductibles is the amount you have to pay before the insurance plan can start to pay. These deductibles affect the individual and the

family as a whole.[1] There are also premiums on insurance plans which affects the value of the deductible which is another consideration when calculating cost. With adding dependents on a Healthcare plan (Blue Cross of Michigan), you only pay for the three oldest children and any additional children added on the plan is at no additional cost. [2] This may only be true for Michigan Healthcare but there are reductions found in most Healthcare plans with additional dependents. There are many other factors to consider when talking about the cost of insurance but we have pulled a dataset that includes some risk factors that will affect cost of insurance.

Our goal in this project is to be able to predict the medical expenses that are charged to a consumer. We are using a large dataset to help to determine the legitimacy of any conclusion that we come up with. Our main goal is to determine if sex, age, BMI, number of children, smoker (or not), and what region they live in determine their cost of insurance. We want to see what tested variables are the most significant in raising the overall price.

Methodology

We spent some time working on finding data that would be relevant to the scope of this project. Our team looked through many datasets that had the right amount of data and variables. We found a good set about insurance located at https://www.kaggle.com/mirichoi0218/insurance , this set has 1,338 observations and 6 independent variables (3 numeric/3 text). This set did not need to have any need for any data cleaning, it was properly laid out in excel for easy data

---

1 https://www.healthcare.gov/glossary/deductible/#:~:text=Deductible,or%20coinsurance%20for%20covered%20services.
https://budgeting.thenest.com/mean-family-deductible-31054.html

2 https://www.bcbsm.com/index/health-insurance-help/faqs/topics/buying-insurance/family-size-impact-cost.html

processing. The only pre-processing step we had to make was to create dummy variables out of smoker, sex, and region. Region had to be split into four separate variables, so that it can be properly processed.

We first started out by creating a scatterplot matrix of all the variables vs charges. Then we created a correlation matrix of all the variables vs the dependent variable to check on the strength of connection of these variables. Next we created histograms of the three numeric variables (BMI, Age, and Number of Children).The last last step that we took in the exploration phase was to create box plots of two binary variables sex and smoker.

After finishing the exploratory analysis we then looked at the initial regression model to see if the data would need any adjustments. Along with the regression we ran tests for Cook's D and studentized residuals to see if there were any outliers that were going to interfere with a correct model equation. Then we ran stepwise, forward, and backwards tests on the regression to see if any variables should be removed. After running all the tests, we refitted the model into the final model and compared the performance with the initial model.

Analysis, Results and Findings

To begin our exploratory analysis, we decided to first look at the correlation between the predictor variables and the target variable (charges). This will give us an idea of which variables will be significant when we build our model later on. To do this, we created a correlation matrix (Figure 7). By looking at the column titled 'charges', we can see that the three predictor variables with the highest correlation coefficients are Smoker (0.78725), Age (0.299), and BMI (0.198). Of these, only Smoker seems to be significantly correlated with the target attribute as it has a

correlation coefficient above 0.5. The other predictor variables have correlation coefficients lower than 0.1, making that decidedly uncorrelated with the target attribute.

Now that we have a sense of which variables are correlated with the target attribute, we decided to look at the way in which these variables are correlated with the target attribute. To see this, we created a scatterplot matrix (Figure 8). Looking at the very last column, which is the column for the target attribute for Charges, we can see that BMI is the only variable that seems to have a somewhat linear relationship with Charges. Since the Smoker variable is binary, we can't see a linear relationship with Charges. That said, we can see that the Smoker variable switches from 0 (non-smoker) to 1 (smoker) around the middle of the Charges value, so there is a correlation between the two variables.

After analyzing correlation, we decided to take a look at the distribution of some of the predictor variables. First, we created histograms for the continuous variables. We can see that the distribution for charges is right skewed (Figure 6), with the majority of charges being around $6,000. We can also see that the potential outlier is above $60,000. For the Children variable and noticed that the majority of individuals have no dependents as the distribution is right skewed (Figure 5). The skewness can be explained by policies around the amount of dependents and if it is a family plan or individual plan. We noticed that the Age variable is right skewed with the majority of individuals being 18 years old (Figure 4). Finally, we saw that the distribution of BMI was normally distributed, with a peak at 29. Next, we created boxplots for the Sex variable (Figure 1), which is a binary variables that denotes whether or not an individual is Male or Female. Looking at the boxplot for Males and Females, we can see that the distribution of Charges is similar for both cases. We can see that, for both cases, the minimum charge is around

~$1000, maximum charge is above $60,000, and the 2nd quartile is around $10,000. Since the boxplot for both cases is so similar, we expect the Sex variable to have be insignificant in determining the Charges attribute. We also created boxplots for the Smoker variable (Figure 2) which has two cases, ‘doesn’t smoke’ and ‘does smoke’. We can see that the distribution for both cases is very different. For non-smokers, the minimum charge is around $1,000, maximum charge is near $40,000, and median is between $5000 and $10,000. For smokers, the minimum charge is around $10,000, maximum charge is above $60,000, and median is close to $35,000. Because these boxplots are quite different, the Smoker variable is likely to have a significant impact on determining Charges.

After running the initial model, with all independent variables accounted for, we can see that the VIFs are all close to 1 (Figure 9). Since the VIFs are not above 10 for any of the variables, we can confidently say that multicollinearity is not an issue. We ran the stepwise selection (Figure 17) on the initial dataset to find that the adjusted r^2 value was .7498 and removed d_sex, region_sw, region_se as they were found insignificant by the selection method.

Outliers have a profound effect on the accuracy of the model and model equation. So in order to make sure that the information is a reflection of the population, we must remove the points that cause these problems. We took the regression model and looked at the values of cook’s D greater than 0.003 and studentized residuals with values greater than ±3. We found that there was 19 points that were above these conditions (4, 35, 141, 220, 431, 469, 517, 544,

578, 820, 937, 1013, 1020, 1028, 1040, 1207, 1231, and 1301). We removed these points and got an adjusted r^2 value of .7497.

After running the stepwise regression test (Figure 13), SAS determined that age, bmi, children, and d_smoker variables were required for an accurate model. The adjusted R-Square value for stepwise (Figure 11) was 0.7492. Next we ran the adjusted R^2 testing method on the same model (Figure 12), which suggested that the variables that we want to keep in the model are age, bmi, children, d_smoker, region_sw, region_se. The result of adj R^2 testing (Figure 13) was that the adj R-square value was 0.7497.The Adj R^2 values between these 2 models are only 0.0005, but r-square testing had two values that were removed in stepwise testing. We decided that the stepwise prediction was better, even though there is a slight loss of R^2, the stepwise model has less variables in the final model.

If a noticeable trend is identified in the residual plots for the predictor variables, then it suggests a failure in assumptions and potential inadequacies in the model. For the final model, the residual plots appear to be randomly scattered, meaning that there is no trend or pattern (Figure 20). Based on this, we have no reason to believe that our final model's assumptions have failed or that the model itself is inadequate.

In order to determine the strongest predictors we look at the standardized estimates of each variable for the final model (Figure 18). D_smoker has by far the highest standardized estimate of 0.7936, this means that this is the strongest predictor of the independent variables.

However this variable is only strong if the person is a smoker, if they are not then age becomes the strongest predictor with a standardized estimate of 0.29743.

The final model of the final model would be as follows:

charges = -12174 + 256.55(age) + 325.06(bmi) + 497.19(children) + 23790(d_smoker)

Where d_smoker = 0 if non-smoker

Using the final regression model we performed a prediction with two very different scenarios. The first scenario was getting health insurance for a 25 year-old with a bmi of 22, no children, and doesn't smoke. Using the prediction (Figure 19), the predicted value is $1,391 and a 95% confidence interval between $672.87 and $2,110. This value of the first observation does fall between these two values.

The second scenario was of a 55 year-old with a bmi of 29, 3 children, and is a smoker. We chose this to have a more middle of the road test of the final model. If we use the second prediction of (Figure 19), we can see that the predicted value is $36,645 with a 95% confidence interval between $24,701 and $48,588.

We created training and test sets for both models retrieved after running stepwise and R-square testing. With this we wanted to see if both sets would be similar to test which one would make the better final model. Model 1 is from age, smoker, bmi, and number of children. After running the test on Model 1 using 990 observations (Figure 27), we can show that the test adj r-square value would be 0.7356. Model 2 is derived from age, smoker, bmi, number of children, region_se, region_sw. When we run the test on Model 2 on 990 observations (Figure

27), we can see that the adj r-square value would be 0.7351. We can see from this that Model 1 has a slightly higher adj r-square value, from this we can say that the model with only four variables is more exact.

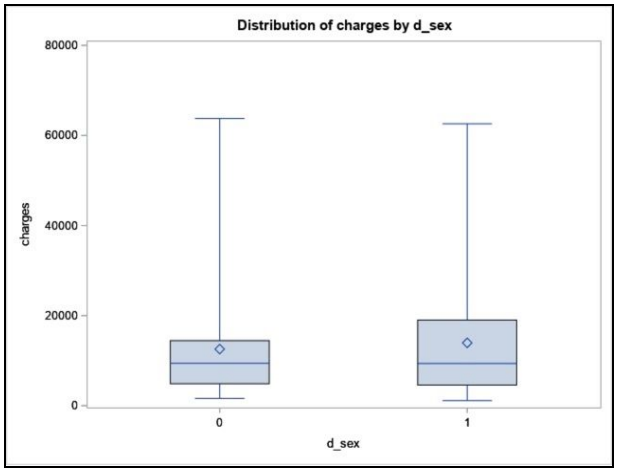
The best way to improve our final model would be to include attributes mentioned in our research that were not collected in this data set. These attributes, which describe potential risk factors of an individual, could improve our R^2 value and therefore improve our model. For example, collecting information regarding an individual's occupation type would inform an insurance firm of the health risks posed to said individual on a daily basis. A more dangerous form of occupation may determine higher medical charges. Another factor not considered in our model is whether or not an individual drinks alcohol. Knowing the alcohol drinking status of an individual could significantly improve our model, as we have already seen how impactful the "Smoker" attribute was in determining medical charges. Just as smoking increases chances of a multitude of medical conditions, which in turn increase medical charges, drinking does the same for a different set of medical conditions. Overall, including more attributes that describe an individual's medical risk factors would potentially increase the R^2 value of our final model.

Our model could have taken more cases into consideration as well. As this dataset represents the population of the United States, it would make sense to include more cases as the population of the country is so large. Including more cases would make our model more accurate in determining medical charges, as it would be more representative of Americans as a whole.

Overall, we are not overly satisfied with our final model. Since the adjusted R^2 value is 0.7492, our model only predicts 74.92% of variability in our data. If the improvements

mentioned above were made, such as including more attributes or more cases, than we could predict much more variability in our data using our model.

Appendix:



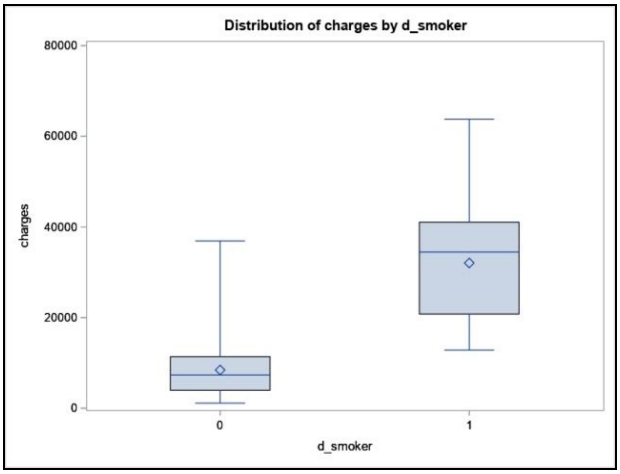
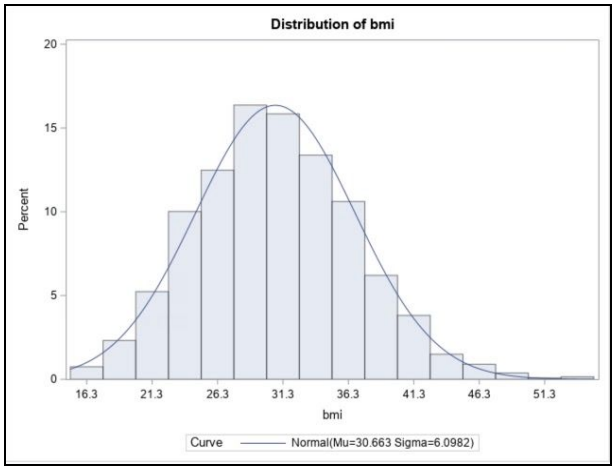


Figure 1, Figure 2 (left, right)



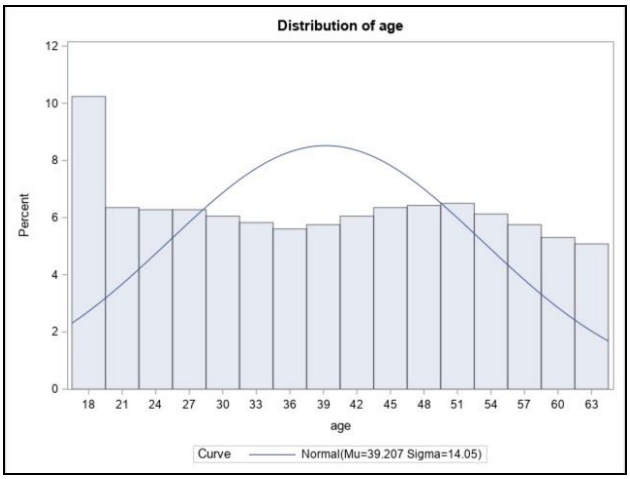
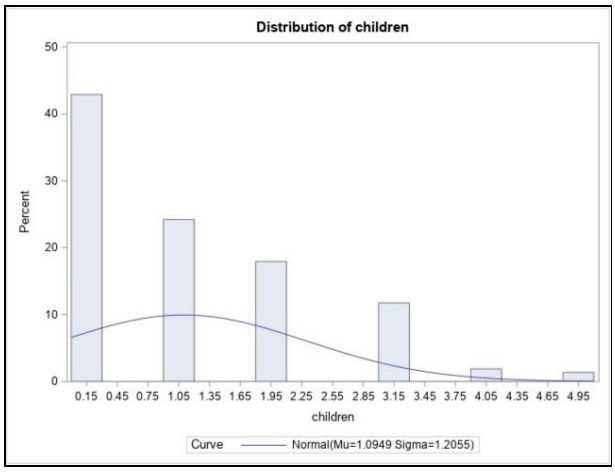


Figure 3, Figure 4 (left, right)



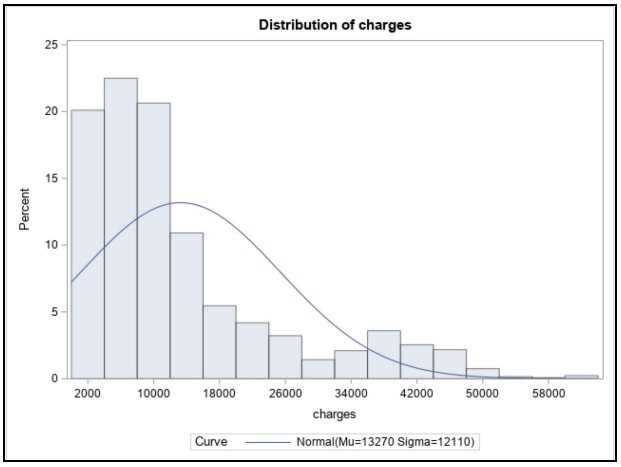


Figure 5, Figure 6(left, right)

Pearson Correlation Coefficients, N = 1338
Prob > |r| under H0: Rho=0

	age	bmi	children	charges	d_sex	d_smoker	region_sw	region_nw	region_se	region_ne
age	1.00000	0.10927 <.0001	0.04247 0.1205	0.29901 <.0001	-0.02086 0.4459	-0.02502 0.3605	0.01002 0.7143	-0.00041 0.9881	-0.01164 0.6705	0.00247 0.9279
bmi	0.10927 <.0001	1.00000	0.01276 0.6410	0.19834 <.0001	0.04637 0.0900	0.00375 0.8910	-0.00621 0.8206	-0.13600 <.0001	0.27002 <.0001	-0.13816 <.0001
children	0.04247 0.1205	0.01276 0.6410	1.00000	0.06800 0.0129	0.01716 0.5305	0.00767 0.7792	0.02191 0.4232	0.02481 0.3646	-0.02307 0.3992	-0.02281 0.4045
charges	0.29901 <.0001	0.19834 <.0001	0.06800 0.0129	1.00000	0.05729 0.0361	0.78725 <.0001	-0.04321 0.1141	-0.03990 0.1446	0.07398 0.0068	0.00635 0.8165
d_sex	-0.02086 0.4459	0.04637 0.0900	0.01716 0.5305	0.05729 0.0361	1.00000	0.07618 0.0053	-0.00418 0.8785	-0.01116 0.6835	0.01712 0.5316	-0.00243 0.9294
d_smoker	-0.02502 0.3605	0.00375 0.8910	0.00767 0.7792	0.78725 <.0001	0.07618 0.0053	1.00000	-0.03695 0.1768	-0.03695 0.1768	0.06850 0.0122	0.00281 0.9182
region_sw	0.01002 0.7143	-0.00621 0.8206	0.02191 0.4232	-0.04321 0.1141	-0.00418 0.8785	-0.03695 0.1768	1.00000	-0.32083 <.0001	-0.34626 <.0001	-0.32018 <.0001
region_nw	-0.00041 0.9881	-0.13600 <.0001	0.02481 0.3646	-0.03990 0.1446	-0.01116 0.6835	-0.03695 0.1768	-0.32083 <.0001	1.00000	-0.34626 <.0001	-0.32018 <.0001
region_se	-0.01164 0.6705	0.27002 <.0001	-0.02307 0.3992	0.07398 0.0068	0.01712 0.5316	0.06850 0.0122	-0.34626 <.0001	-0.34626 <.0001	1.00000	-0.34556 <.0001
region_ne	0.00247 0.9279	-0.13816 <.0001	-0.02281 0.4045	0.00635 0.8165	-0.00243 0.9294	0.00281 0.9182	-0.32018 <.0001	-0.32018 <.0001	-0.34556 <.0001	1.00000

Figure 7

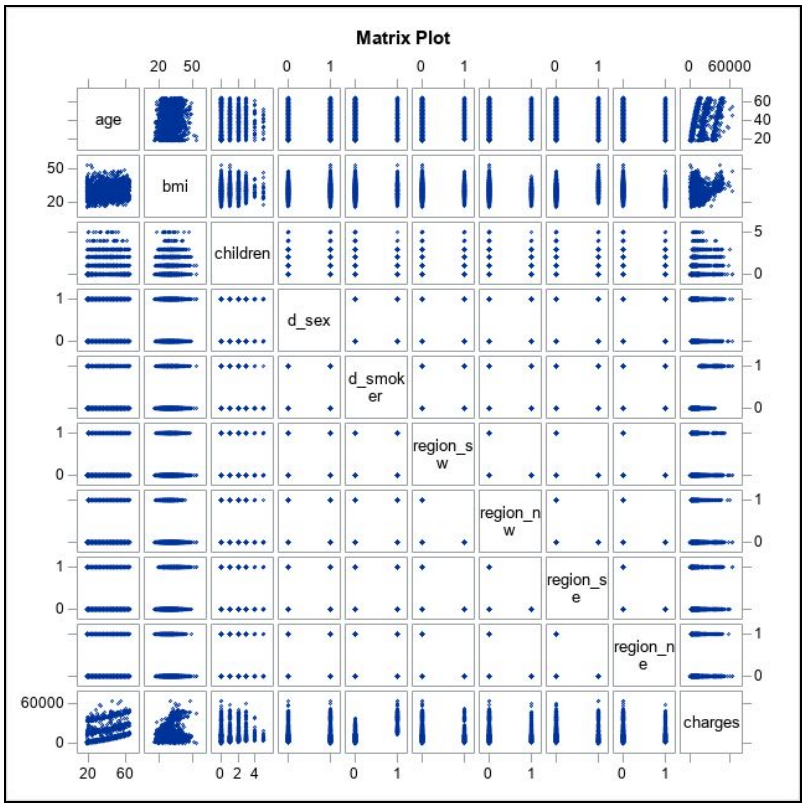


Figure 8

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-12003	999.42938	-12.01	<.0001	0	0
age	1	255.49525	11.99093	21.31	<.0001	0.29620	1.01674
bmi	1	341.24501	28.94761	11.79	<.0001	0.17100	1.10710
children	1	499.58391	139.52307	3.58	0.0004	0.04945	1.00354
d_sex	1	-132.28182	335.39360	-0.39	0.6933	-0.00546	1.00851
d_smoker	1	23830	415.63876	57.33	<.0001	0.79494	1.01145
region_sw	1	-897.62075	480.90696	-1.87	0.0622	-0.03187	1.53379
region_nw	1	-349.04759	481.17462	-0.73	0.4683	-0.01233	1.51932
region_se	1	-973.06748	482.85869	-2.02	0.0441	-0.03579	1.65937

Figure 9

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.451122E11	36278046619	985.25	<.0001
Error	1314	48383153714	36821274		
Corrected Total	1318	1.934953E11			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-12174	953.10373	6007013310	163.14	<.0001
age	256.55072	11.98228	16879765865	458.42	<.0001
bmi	325.05517	27.69350	5072910953	137.77	<.0001
children	497.18702	139.47232	467909975	12.71	0.0004
d_smoker	23790	413.66402	1.217857E11	3307.48	<.0001

Bounds on condition number: 1.0141, 16.113

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	d_smoker		1	0.6209	0.6209	679.420	2157.29	<.0001
2	age		2	0.1005	0.7214	152.800	474.60	<.0001
3	bmi		3	0.0261	0.7475	17.3136	136.11	<.0001
4	children		4	0.0024	0.7500	6.5906	12.71	0.0004

Figure 10(step-wise)

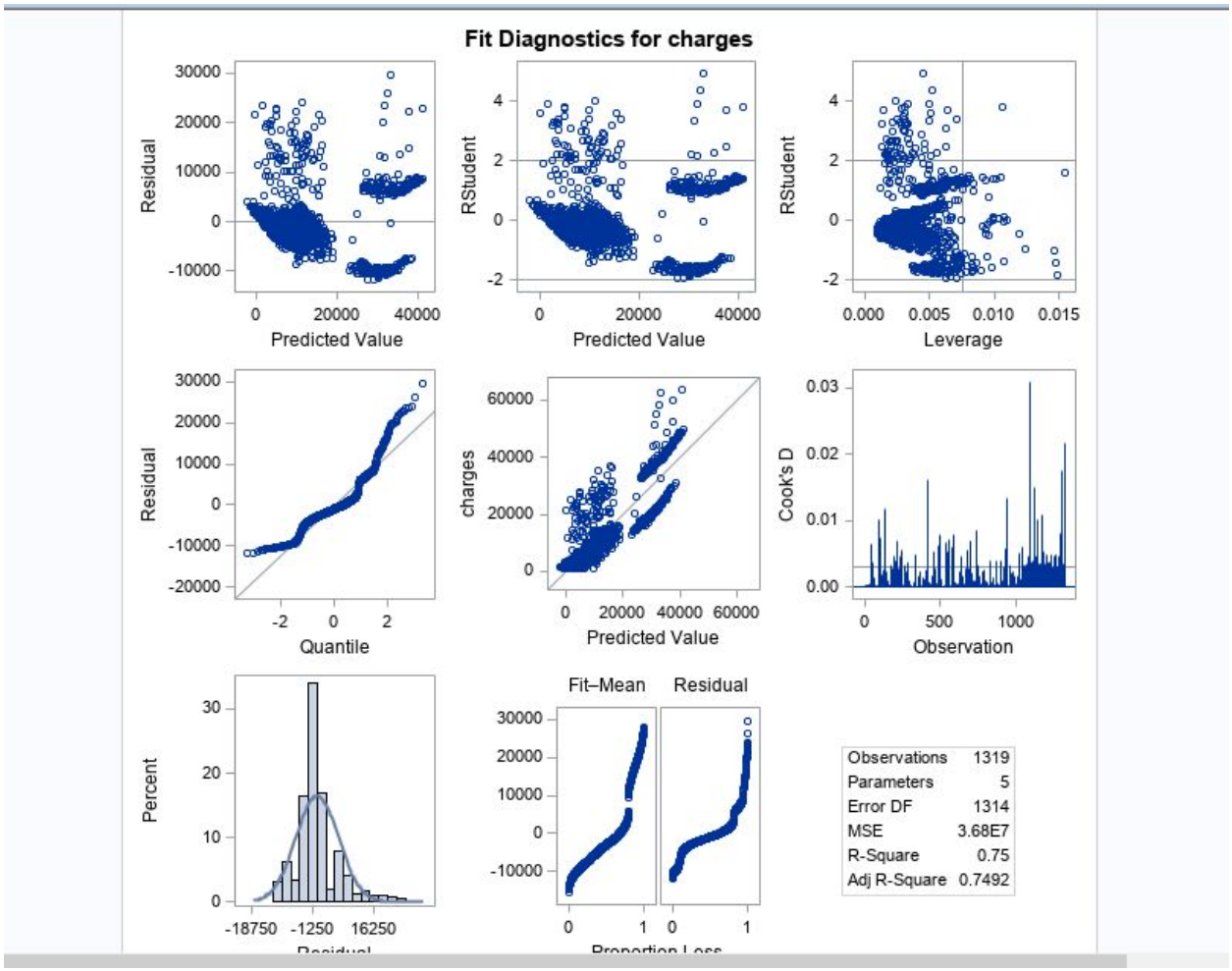


Figure 11 (step-wise)

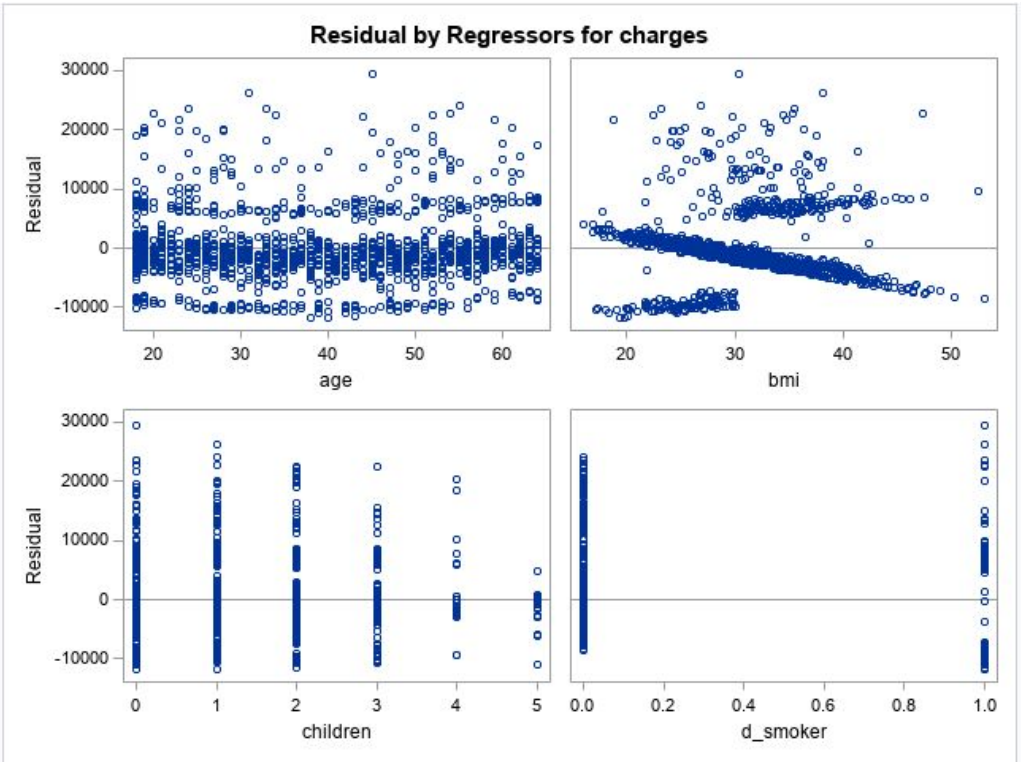


Figure 12 (Stepwise Residual)

R-Testing

The REG Procedure
Model: MODEL1
Dependent Variable: charges

Adjusted R-Square Selection Method

Number of Observations Read	1319
Number of Observations Used	1319

Number in Model	Adjusted R-Square	R-Square	Variables in Model
6	0.7497	0.7509	age bmi children d_smoker region_sw region_se
7	0.7497	0.7510	age bmi children d_smoker region_sw region_nw region_se
7	0.7496	0.7509	age bmi children d_sex d_smoker region_sw region_se
8	0.7495	0.7510	age bmi children d_sex d_smoker region_sw region_nw region_se
5	0.7494	0.7503	age bmi children d_smoker region_se
5	0.7492	0.7502	age bmi children d_smoker region_sw
6	0.7492	0.7503	age bmi children d_sex d_smoker region_se
4	0.7492	0.7500	age bmi children d_smoker

Figure 13 (Adjusted r^2 selection method)

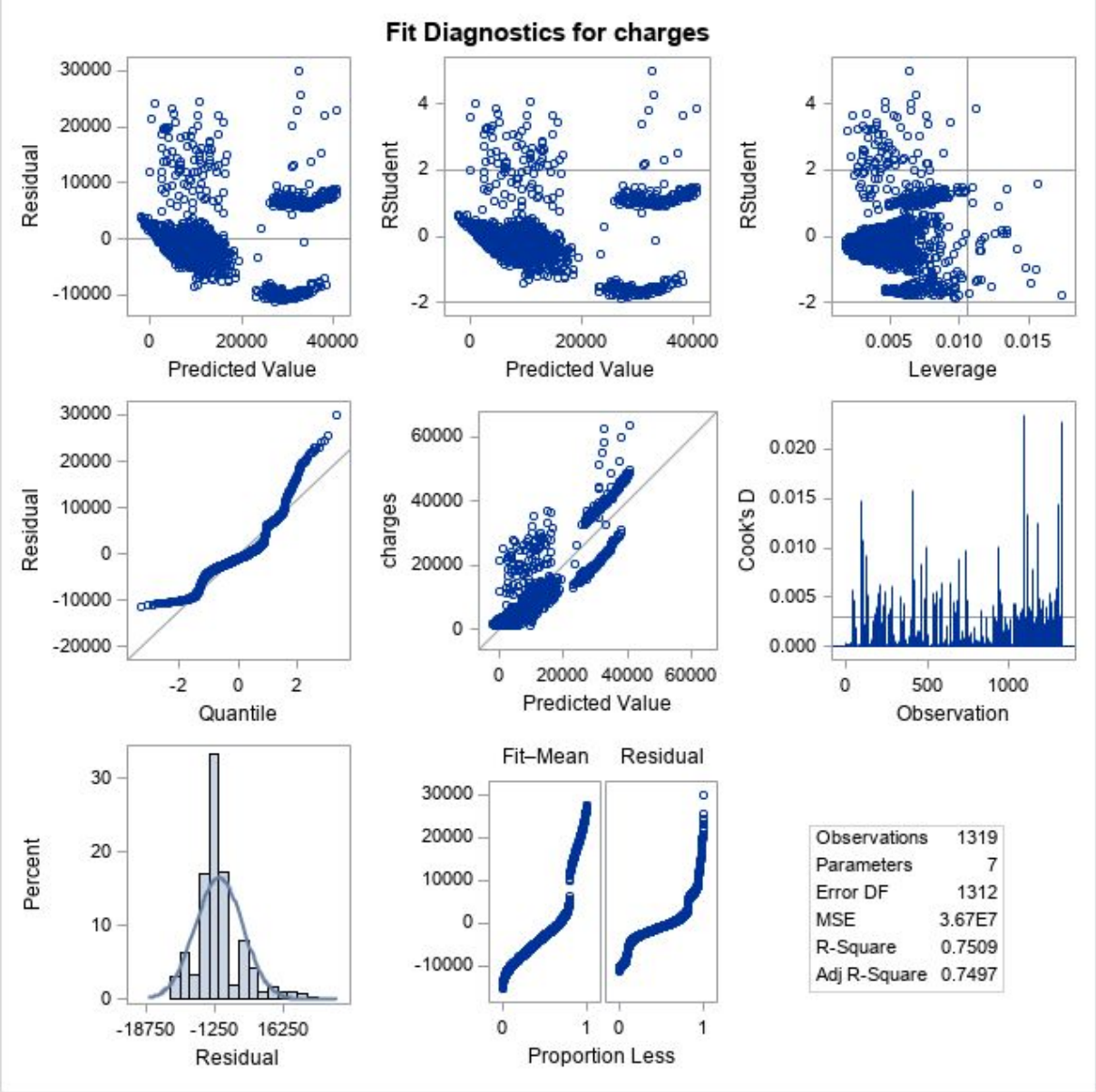


Figure 14 (r selection)

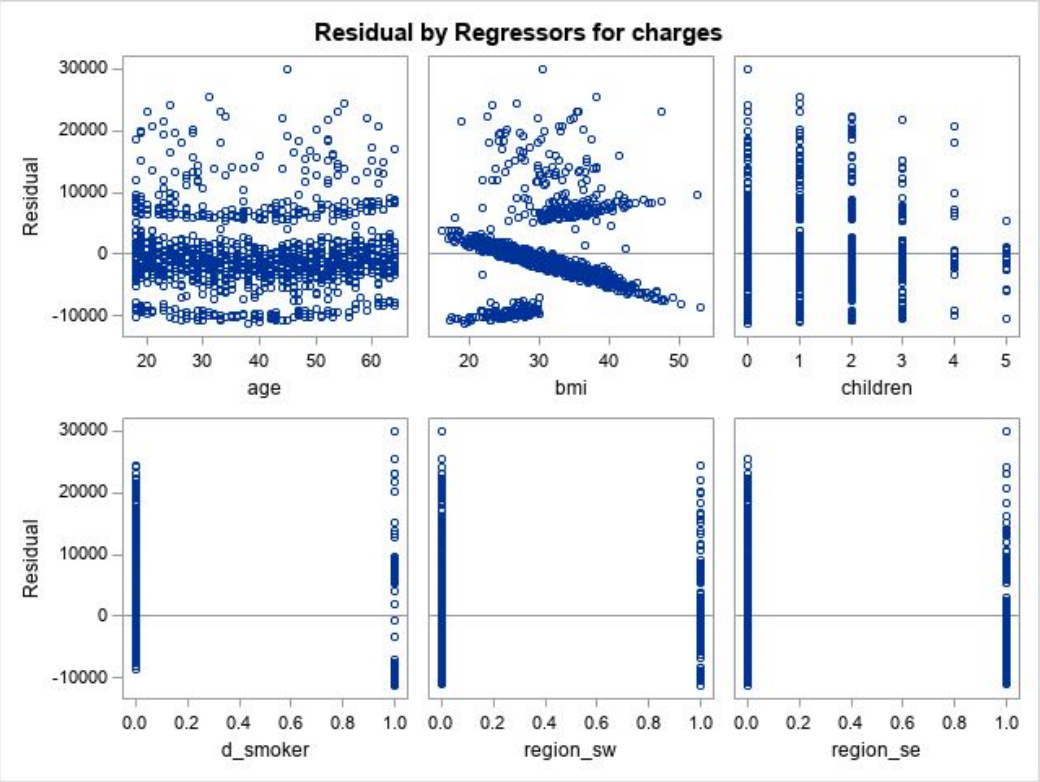


Figure 15 (R selection)

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance
Intercept	1	-12003	999.42938	-12.01	<.0001	.
age	1	255.49525	11.99093	21.31	<.0001	0.98353
bmi	1	341.24501	28.94761	11.79	<.0001	0.90326
children	1	499.58391	139.52307	3.58	0.0004	0.99647
d_sex	1	-132.28182	335.39360	-0.39	0.6933	0.99157
d_smoker	1	23830	415.63876	57.33	<.0001	0.98868
region_sw	1	-897.62075	480.90696	-1.87	0.0622	0.65198
region_nw	1	-349.04759	481.17462	-0.73	0.4683	0.65819
region_se	1	-973.06748	482.85869	-2.02	0.0441	0.60264

Figure 16

Collinearity Diagnostics											
			Proportion of Variation								
Number	Eigenvalue	Condition Index	Intercept	age	bmi	children	d_sex	d_smoker	region_sw	region_nw	region_se
1	5.01099	1.00000	0.00106	0.00394	0.00127	0.01196	0.01179	0.00891	0.00499	0.00487	0.00529
2	1.01859	2.21801	0.00002061	0.00014692	0.00000997	0.00389	0.00000371	0.04009	0.09127	0.08713	0.24599
3	1.00026	2.23823	1.872457E-7	2.59174E-10	0.00000597	0.00005281	0.00000389	0.00013313	0.24031	0.25532	0.00001274
4	0.74687	2.59023	0.00024113	0.00152	0.00046163	0.01842	0.00026422	0.92363	0.00194	0.00079688	0.03834
5	0.50984	3.13506	0.00023191	0.00069690	0.00037812	0.74016	0.26149	0.01875	0.00374	0.00592	0.00048368
6	0.42945	3.41590	0.00209	0.01393	0.00300	0.20118	0.68281	0.00019806	0.04723	0.04889	0.03811
7	0.19612	5.05472	0.00609	0.13773	0.00691	0.01496	0.01708	0.00079279	0.53737	0.52024	0.54834
8	0.07079	8.41347	0.06123	0.80394	0.14242	0.00371	0.02332	0.00348	0.06851	0.05594	0.11551
9	0.01709	17.12226	0.92904	0.03810	0.84555	0.00567	0.00323	0.00403	0.00464	0.02090	0.00794

Figure 17

Final Model

The REG Procedure
Model: MODEL1
Dependent Variable: charges

Number of Observations Read	1319
Number of Observations Used	1319

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.451122E11	36278046619	985.25	<.0001
Error	1314	48383153714	36821274		
Corrected Total	1318	1.934953E11			

Root MSE	6068.05354	R-Square	0.7500
Dependent Mean	13270	Adj R-Sq	0.7492
Coeff Var	45.72884		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-12174	953.10373	-12.77	<.0001	0	0
age	1	256.55072	11.98228	21.41	<.0001	0.29743	1.01405
bmi	1	325.05517	27.69350	11.74	<.0001	0.16289	1.01203
children	1	497.18702	139.47232	3.56	0.0004	0.04921	1.00160
d_smoker	1	23790	413.66402	57.51	<.0001	0.79360	1.00065

Figure 18 (final model)

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	1391	366.2415	672.8707	2110	-10534	13317	.
2	.	36645	494.1150	35676	37614	24701	48588	.

Figure 19

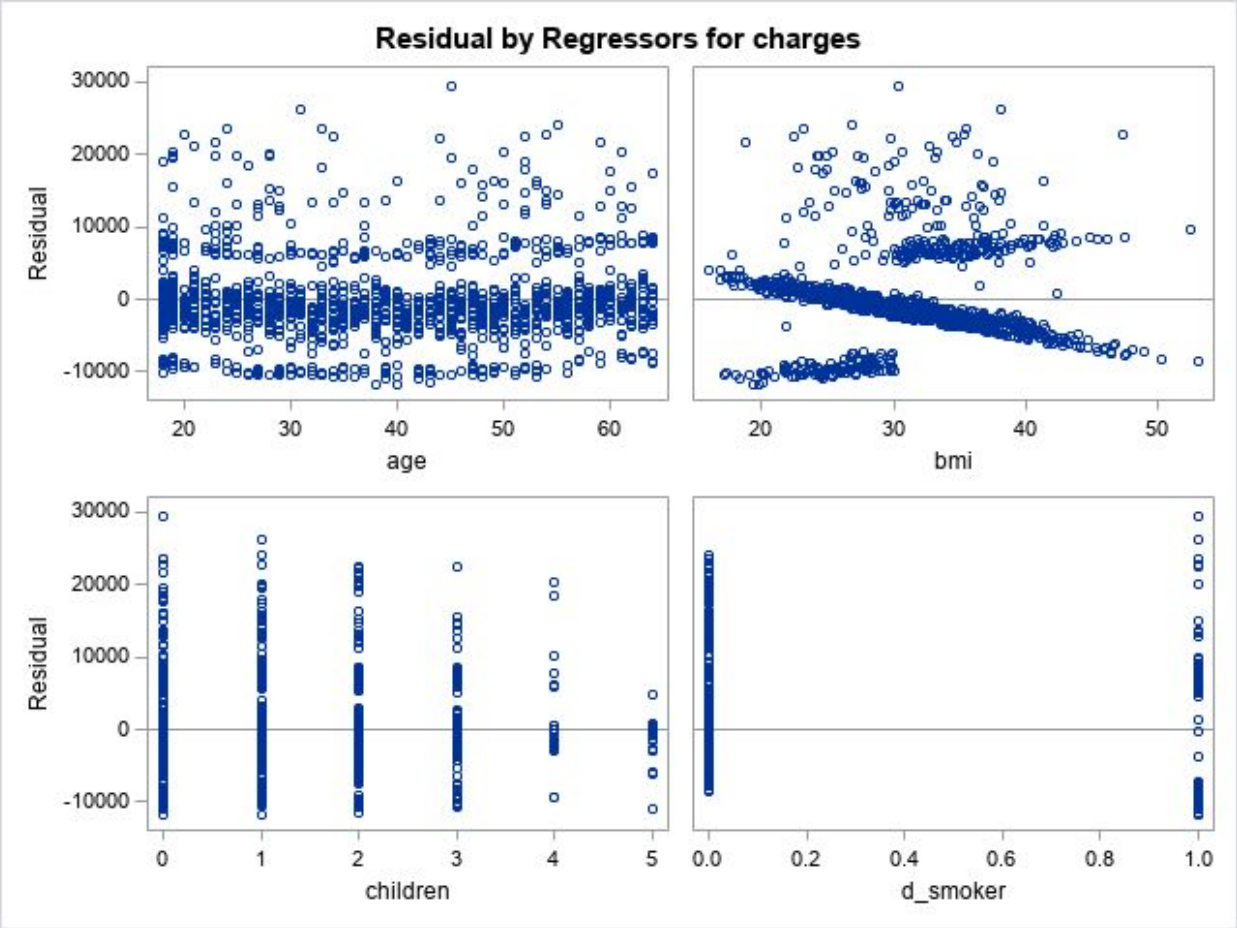


Figure 20 (Final Model Residual)

validation for model 1

The CORR Procedure

2 Variables:	charges yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
charges	329	13628	11819	4483771	1122	55135	
yhat	329	13644	10552	4488923	-1215	39735	Predicted Value of new_charges

Pearson Correlation Coefficients, N = 329
Prob > |r| under H0: Rho=0

	charges	yhat
charges	1.00000	0.85835 <.0001
yhat Predicted Value of new_charges	0.85835 <.0001	1.00000

Figure 21(validation)

validation for model 1

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	329	6077.50	4138.86

Figure 22(validation)

validation - test set

The REG Procedure
Model: MODEL1
Dependent Variable: new_charges

Number of Observations Read	1319
Number of Observations Used	990
Number of Observations with Missing Values	329

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.112518E11	27812958782	753.24	<.0001
Error	985	36370710916	36924580		
Corrected Total	989	1.476225E11			

Root MSE	6076.55985	R-Square	0.7536
Dependent Mean	13150	Adj R-Sq	0.7526
Coeff Var	46.20821		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-11150	1102.98666	-10.11	<.0001
age	1	259.69498	13.84444	18.76	<.0001
bmi	1	286.11997	31.83879	8.99	<.0001
children	1	473.82888	162.60965	2.91	0.0037
d_smoker	1	24160	478.79893	50.46	<.0001

Figure 23(validation)

validation for model 2

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	329	6086.59	4183.07

Figure 24(validation)

validation for model 2

The CORR Procedure

2 Variables:	charges yhat

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
charges	329	13628	11819	4483771	1122	55135	
yhat	329	13671	10593	4497621	-1585	40361	Predicted Value of new_charges

Pearson Correlation Coefficients, N = 329 Prob > |r| under H0: Rho=0		
	charges	yhat
charges	1.00000	0.85805 <.0001
yhat Predicted Value of new_charges	0.85805 <.0001	1.00000

Figure 25(validation)

The REG Procedure
Model: MODEL2
Dependent Variable: new_charges

Number of Observations Read	1319
Number of Observations Used	990
Number of Observations with Missing Values	329

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1.114459E11	18574315789	504.71	<.0001
Error	983	36176651309	36802290		
Corrected Total	989	1.476225E11			

Root MSE	6066.48912	R-Square	0.7549
Dependent Mean	13150	Adj R-Sq	0.7534
Coeff Var	46.13163		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-11368	1115.23693	-10.19	<.0001
age	1	258.86813	13.82930	18.72	<.0001
bmi	1	308.91874	33.43748	9.24	<.0001
children	1	469.49898	162.38060	2.89	0.0039
d_smoker	1	24229	479.76993	50.50	<.0001
region_sw	1	-633.63398	480.36965	-1.32	0.1875
region_se	1	-1084.10397	485.91205	-2.23	0.0259

Figure 26(validation for model 2)

	Model 1: 4 predictors	Model 2: 6 predictors
Train	RMSE: 6076.55985 R^2: 0.7536 adj-R^2: 0.7526 GOF: OK Residuals: OK	RMSE: 6066.48912 R^2: .7549 adj-R^2: .7534 GOF: OK Residuals: OK
Test	RMSE: 6077.50 MAE: 4138.86 R^2: 0.85835^2=0.7367 adj-R^2: 0.7356 cv-R^2: 0.0169	RMSE: 6086.59 MAE: 4183.07 R^2: .85805^2=.7362 adj-R^2: .7351 cv-R^2; .7549-.7362=.0187

Figure 27