A Predictive Model for Personal Medical Insurance Charge

Zengxiaoran Kang, Fang Shu, Qiying Tao, Mingyue Wang

1. Introduction

Medical insurance is an essential part of people's self-protection plan as it covers a person's medical expenses for illness or injury. However, various factors of each individual may affect their medical insurance charges, such as age, body condition, and more. We constructed a prediction model to estimate insurance charges while providing more insights into the effect of each factor. This project utilized an insurance dataset obtained from the book *Machine Learning with R* by Brett Lantz. The data were analyzed using JMP to examine the relationship between features indicating individual characteristics and the corresponding health insurance charges.

2. Data Description

The data contains relative information about each person and each corresponding charge billed by health insurance companies. There are 1338 observations and 7 variables in this dataset. The response variable is the insurance charge which is a continuous variable. The predictor variables include three numerical continuous variables (age, bmi, children) and three categorical nominal variables (sex, smoker, region).

1. age: age of the primary beneficiary (in years)
2. sex: "male" or "female"
3. bmi: body mass index
4. children: number of dependents covered by health insurance
5. smoker: "yes" or "no"
6. region: the beneficiary's residential area in the US; "northeast", "northwest", "southeast", or "southwest"
7. charges: individual medical costs billed by health insurance (in dollars)

3. Objectives

1. To identify the most significant attributes relative to insurance charges.
2. To build a validated regression model for predicting insurance charges.

4. Data Cleaning

We discovered that we may acquire a better fitted model by creating a new categorical factor and removing influential observations using Cook's Distance. We then split the data into training and testing to examine the model's prediction. No missing or N/A values were found in the dataset.

4.1 Create a New Categorical Factor to Differentiate Charges

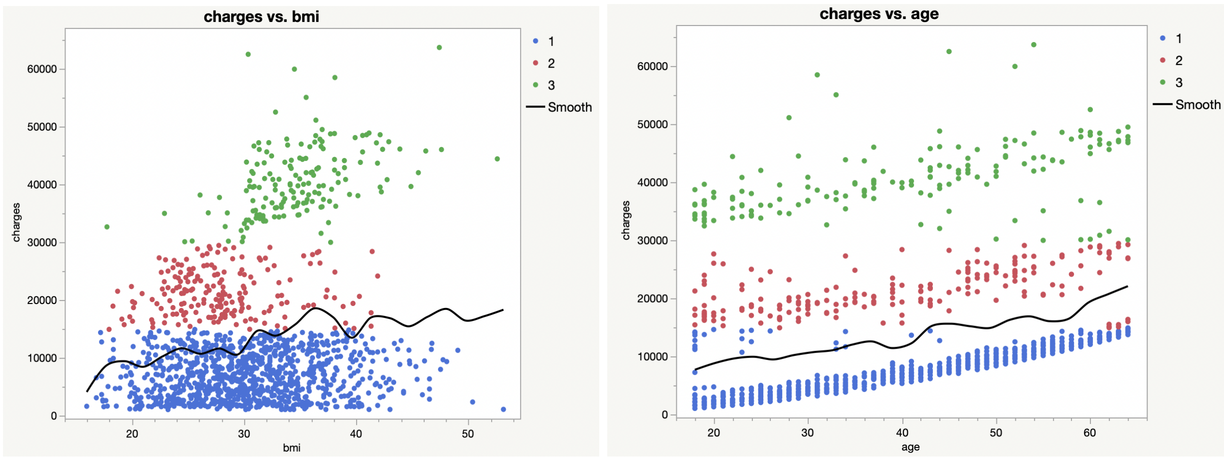


Figure 1. Scatterplot of charges vs bmi and charges vs age

From the scatterplots of the two quantitative variables, the two factors "bmi" and "age" showed a somewhat linear relationship with the response, "charges", albeit in clusters. This observation indicated a necessity to classify "charges" into three categories. Group 1 (low-charges) consisted of "charges" that were less than or equal to 15000 dollars. Group 2 (medium-charges) consisted of "charges" greater than 15000 but less than or equal to 30000 dollars. Group 3 (high-charges) consisted of "charges" that were greater than 30000 dollars.

4.2 Filter Influencers with Cook's Distance

We used Cook's distance to identify and remove influential observations in the set of predictor variables to obtain a better regression model. The threshold $\frac{4}{(n-2)}$, where n is the number of observations, was used to identify data points that may have a negative influence on modeling the response variable. A data point was treated as an influential observation if it had a Cook's distance greater than the calculated threshold; thus, a total of 113 points were excluded from the analysis. Figure 2 shows the removed data colored in red, and the remaining observations are shown in blue.

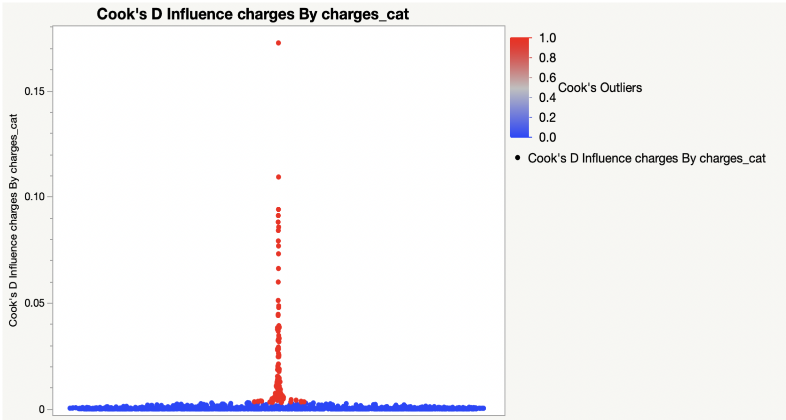


Figure 2. Cook's D plot for Observations

4.3 Split to Training and Testing data

We randomly split the filtered data into training and testing sets at a 7:3 ratio. We would fit and develop the model on training data and verify the result with the testing set.

5. Data Analysis and Results

We use forward selection to identify the significant main effects and interactions by observing the p-values and half-normal plots. We then fit the selected model on the training data and test the model on the testing data.

5.1 Forward Selection

All results are generated at the 95% confidence level. We first ran a simple model with all the 7 main effects. By observing the corresponding p-values and the half-normal plot, we found that all the main effects were significant. We then added all the two-way interactions into the model. The half-normal plot indicated that the interaction between bmi and smoker and the interaction between Charges categorical and bmi were significant. We then added all the three-way interactions into the model. Based on the half-normal plot, none of the three-way interactions were significant.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5990.8259	383.4616	15.62	<.0001*
age	261.22733	2.021884	129.20	<.0001*
sex[female]	233.25956	28.00805	8.33	<.0001*
bmi	194.22387	11.81539	16.44	<.0001*
children	433.67064	23.24988	18.65	<.0001*
smoker[no]	-1557.028	134.7902	-11.55	<.0001*
region[northeast]	389.58782	49.305	7.90	<.0001*
region[northwest]	108.60211	49.77388	2.18	0.0294*
region[southeast]	-273.5026	49.39525	-5.54	<.0001*
Charges Categorical[1]	-14118.25	137.9918	-102.3	<.0001*
Charges Categorical[2]	-645.1099	98.78021	-6.53	<.0001*
(bmi-30.9023)*smoker[no]	-369.5289	27.84547	-13.27	<.0001*
Charges Categorical[1]*(bmi-30.9023)	179.39016	30.29433	5.92	<.0001*
Charges Categorical[2]*(bmi-30.9023)	-53.78008	20.01168	-2.69	0.0073*

Summary of Fit

RSquare	0.959054
RSquare Adj	0.958652
Root Mean Square Error	2462.476
Mean of Response	13270.42
Observations (or Sum Wgts)	1338

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	13	1.8805e+11	1.447e+10	2385.481
Error	1324	8028457923	6063790	Prob > F
C. Total	1337	1.9607e+11		<.0001*

Figure 3. Parameter Estimates of the Selected Predictors and Summary of Fit for Selected Model

5.2 Effect Tests for Significant Predictors

LSMeans Differences Student's t

α= 0.050 t= 1.96272

LSMean[i]	Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	LSMean[j] northeast	northwest	southeast	southwest
northeast		0	280.986	663.09	614.275
		0	80.7646	80.7481	79.7844
		0	122.467	504.604	457.681
		0	439.504	821.576	770.87
northwest		-280.99	0	382.105	333.289
		80.7646	0	81.6409	80.04
		-439.5	0	221.866	176.193
		-122.47	0	542.343	490.386
southeast		-663.09	-382.1	0	-48.815
		80.7481	81.6409	0	79.1131
		-821.58	-542.34	0	-204.09
		-504.6	-221.87	0	106.462
southwest		-614.28	-333.29	48.8152	0
		79.7844	80.04	79.1131	0
		-770.87	-490.39	-106.46	0
		-457.68	-176.19	204.092	0

Level				Least Sq Mean
northeast	A			23114.692
northwest		B		22833.706
southwest			C	22500.416
southeast			C	22451.601

Levels not connected by same letter are significantly different.

Least Squares Means Table

Level	Least Sq Mean	Std Error	Lower 95%	Upper 95%	Mean
northeast	23114.692	78.453260	22960.710	23268.673	12971.3
northwest	22833.706	78.951810	22678.746	22988.666	10960.0
southeast	22451.601	80.291722	22294.011	22609.191	13258.4
southwest	22500.416	77.434037	22348.435	22652.398	11415.1

Figure 4. Pairwise Comparison of Regions

The insurance charge differs by region, with the northeast region having a higher average charge than the other three regions. The southwest region has the lowest average charge.

LSMeans Differences Student's t

α= 0.050 t= 1.96272

LSMean[i]	Mean[i]-Mean[j] / Std Err Dif / Lower CL Dif / Upper CL Dif	LSMean[j]: female	LSMean[j]: male
female	Mean[i]-Mean[j]	0	466.519
	Std Err Dif	0	56.0161
	Lower CL Dif	0	356.575
	Upper CL Dif	0	576.463
male	Mean[i]-Mean[j]	-466.52	0
	Std Err Dif	56.0161	0
	Lower CL Dif	-576.46	0
	Upper CL Dif	-356.58	0

Level			Least Sq Mean
female	A		22958.363
male		B	22491.844

Levels not connected by same letter are significantly different.

Least Squares Means Table

Level	Least Sq Mean	Std Error	Lower 95%	Upper 95%	Mean
female	22958.363	69.071910	22822.794	23093.932	11675.8
male	22491.844	66.114374	22362.080	22621.608	12671.1

Figure 5. Pairwise Comparison of Sex

The insurance charge differs by sex, with females having a higher average charge.

LSMeans Differences Student's t

α= 0.050 t= 1.96272

LSMean[i]	Mean[i]-Mean[j] / Std Err Dif / Lower CL Dif / Upper CL Dif	LSMean[j]: no	LSMean[j]: yes
no	Mean[i]-Mean[j]	0	-3114.1
	Std Err Dif	0	269.58
	Lower CL Dif	0	-3643.2
	Upper CL Dif	0	-2584.9
yes	Mean[i]-Mean[j]	3114.06	0
	Std Err Dif	269.58	0
	Lower CL Dif	2584.95	0
	Upper CL Dif	3643.17	0

Level			Least Sq Mean
yes	A		24282.132
no		B	21168.075

Levels not connected by same letter are significantly different.

Least Squares Means Table

Level	Least Sq Mean	Std Error	Lower 95%	Upper 95%	Mean
no	21168.075	136.74139	20899.690	21436.460	7676.2
yes	24282.132	158.78193	23970.488	24593.777	31480.5

Figure 6. Pairwise Comparison of Smokers and Non-Smokers

The insurance charge differs by whether a person smokes or not, with smokers having a higher charge.

5.3 Null Hypothesis

Selected Model: charges = $\beta_0+\beta_1 * age+\beta_2 * Charges\_Categorical+\beta_3 * bmi+\beta_4 * children +\beta_5 * smoker+\beta_6 * region+\beta_7 * bmi * smoker+\beta_8 * sex+\beta_9 * Charges\_Categorical * bmi$

H_0: All the β coefficients for the predictors are equal to 0: $\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=\beta_7=\beta_8=\beta_9=0$. In other words, none of the explanatory variables is statistically significant.

H_a: At least one of the β coefficients for the predictors is not equal to 0. In other words, at least one of the explanatory variables is significant.

Since the corresponding p-values for each explanatory variable is less than the alpha value of 0.05, we reject H_0. This means we have sufficient evidence to conclude that at least one of the β coefficients is significant. In fact, all the explanatory variables turned out to be significant.

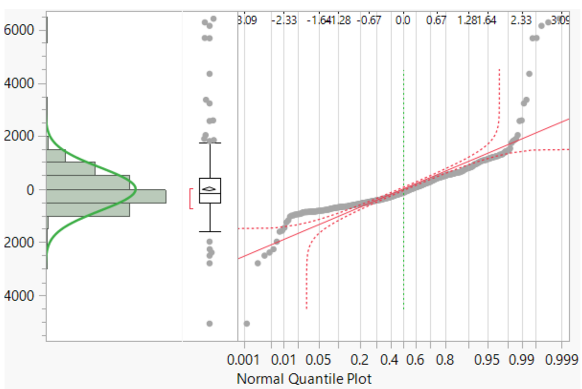
H_0: The model does not have any predictive ability.

H_a: The model has predictive ability.

Since the p-value < 0.0001, we rejected the null hypothesis at 95% level of significance, and concluded that our model is significant and has valid predictive power.

5.4 Normality Check

To see whether the observations are normally distributed, we plot the residuals and conduct a Goodness-of-Fit test. Since the p-value in the Shapiro-Wilk test is less than the alpha value of 0.05, we reject the null hypothesis and conclude that the residuals are not normally distributed.

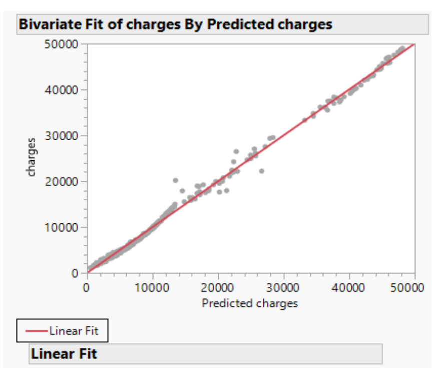


Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.7910013	<.0001*

Figure 7. Normality test of Residuals

5.5 Model fit on Testing Data



Summary of Fit

RSquare	0.995944
RSquare Adj	0.995933
Root Mean Square Error	789.3
Mean of Response	12721.07
Observations (or Sum Wgts)	371

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	5.6448e+10	5.645e+10	90608.15
Error	369	229884988	622994.55	Prob > F
C. Total	370	5.6678e+10		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-22.48081	58.91981	-0.38	0.7030
Predicted charges	1.0006038	0.003324	301.01	<.0001*

Figure 8. Model Fitting on Testing Data

By the Shapiro Wilk Normality test, the model cannot be developed since the residuals are not normally distributed. This could be due to left-over influential data points in the model. Also, performing a log-transformation on the response variable did not result in a normally distributed residual plot. However, if it is assumed that the normality assumption is satisfied, all of the explanatory variables in the model are significant. By testing our fitted model on the testing data, we found that the model was significant and had a high R^2 value of 0.9959.

6. Conclusion

Although the normality assumption is violated under the 0.05 alpha threshold, the model is proved to help budget a person's insurance charges as described above. In addition, we successfully rejected our null hypothesis and identified that the charge's category (low, medium, high standard), BMI, children, smoker, the residential region in the U.S., sex, and two interaction effects, BMI with smoker and charge's category with BMI, are significant towards our model. A major limitation of the study was that we manually created an additional variable charge category to describe the clustering of data, which is not given by the data providers and thus requires further consideration when deploying the model into practical use. Our model requires each individual to set an expectation towards his expenses on medical insurance before making a prediction. Perhaps other methods such as the K-means clustering can be applied to identify the charge's category. Furthermore, we removed about 7% of bad influencers with Cook's distance, which can be explored further for the interest of study.

7. References

Choi, Miri. "Medical Cost Personal Datasets." Kaggle, 21 Feb. 2018, https://www.kaggle.com/mirichoi0218/insurance.

Cook R.D. (2011) Cook's Distance. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_189

Checking the Normality of a Sample. (2020, August 11). University of New South Wales. https://stats.libretexts.org/@go/page/8267

B. HuitemaThe Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-experiments, and Single-case Studies
Wiley, Chichester (2011)