Abhijay Paturi

Professor Lodovico Pizzati

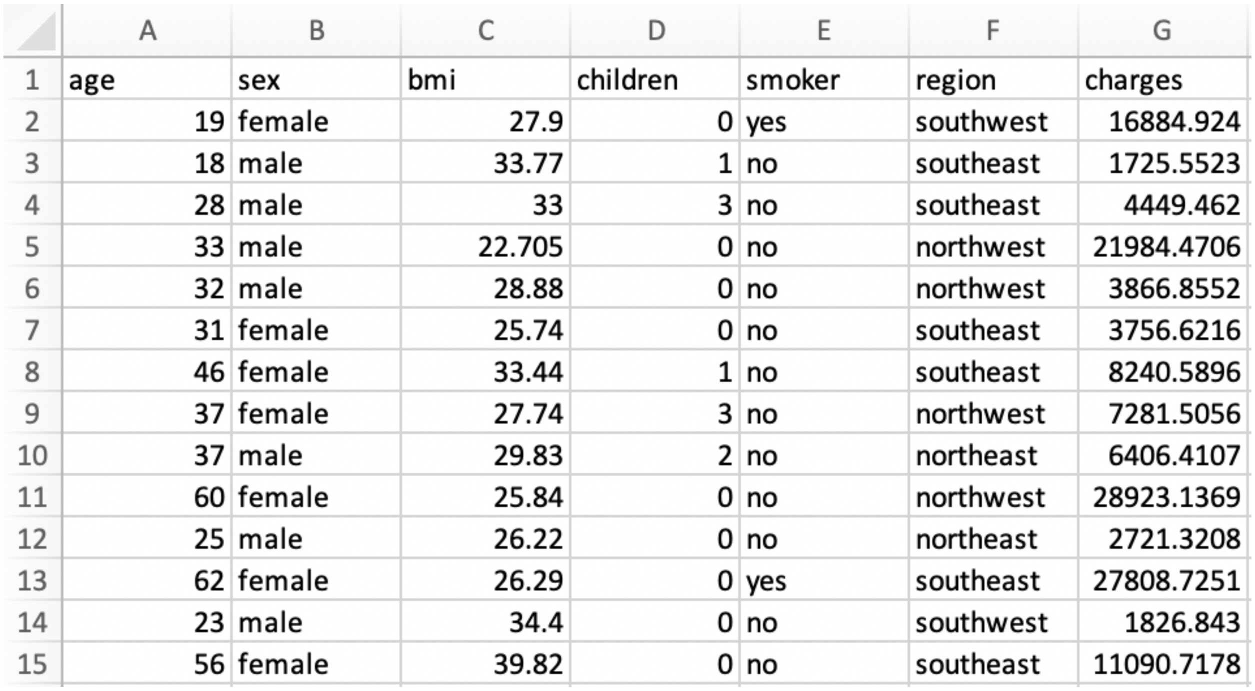
ECON317

9 May 2022

Multiple Linear Regression Analysis Explaining Multiple Factors that Affect the Annual Cost of Health Insurance Premium in America

In the paper, the personal data collected about individuals in America is being analyzed to predict the annual premium that a health insurance company would charge an individual in dollars. Thus, the dependent variable in this analysis is the premium an individual would need to pay the health insurance company yearly. Many independent variables about an individual's life must be analyzed to determine the annual cost of health insurance premiums. The six main independent variables in the analysis are the age, body mass index (BMI), and sex of the individual; the number of children the individual cares for; the region where the individual resides; and whether the individual is a smoker or not. BMI is measured in kilograms per meters^2 ($\frac{kg}{m^2}$). The four regions that the dataset downloaded from Kaggle contains are southeast, southwest, northeast, and northwest. A health insurance premium is the upfront payment that health insurance policyholders must pay, typically every month, to maintain an active policy and thus health coverage. Health insurance companies must analyze an individual's physical, familial, and geospatial data to identify how healthy the individual is and recommend the appropriate policy. If the health insurance company determines that an individual has unhealthy habits or is at risk for certain diseases, it will recommend a plan with a far higher annual premium. Without health insurance, medical bills become unaffordable for millions across America.

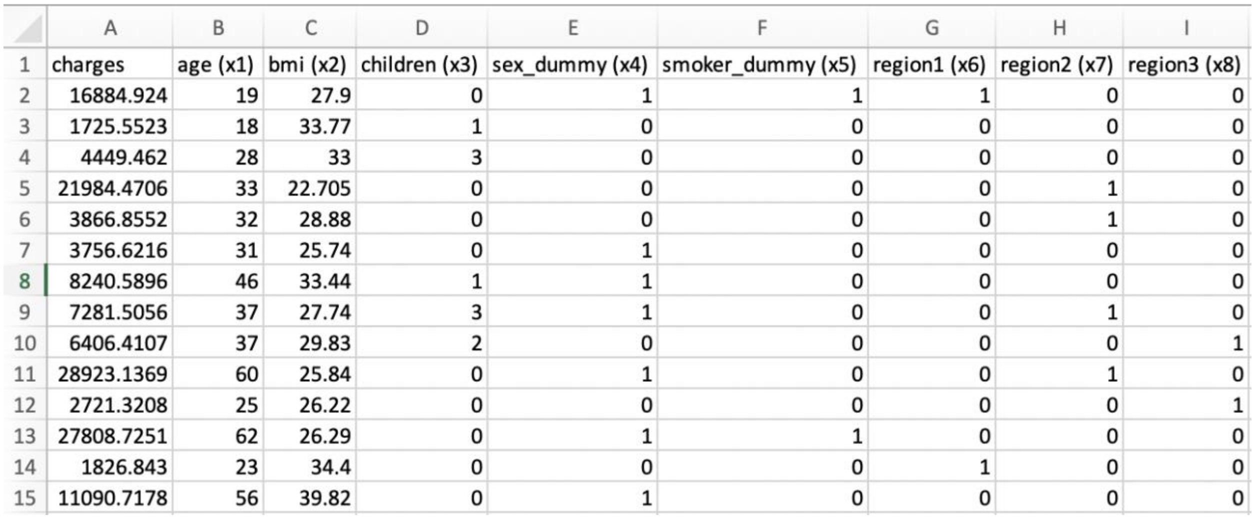
The purpose of the paper is to interpret the results of the derived final estimated multiple linear regression equation associated with the relatively highest R^2 and independent variables that are statistically significant using the selected health insurance dataset. There are 1,338 observations and three numerical and three categorical variables in the dataset. Figure 1 displays the first 14 observations and the names of the independent variables from the original dataset.



	A	B	C	D	E	F	G
1	age	sex	bmi	children	smoker	region	charges
2	19	female	27.9	0	yes	southwest	16884.924
3	18	male	33.77	1	no	southeast	1725.5523
4	28	male	33	3	no	southeast	4449.462
5	33	male	22.705	0	no	northwest	21984.4706
6	32	male	28.88	0	no	northwest	3866.8552
7	31	female	25.74	0	no	southeast	3756.6216
8	46	female	33.44	1	no	southeast	8240.5896
9	37	female	27.74	3	no	northwest	7281.5056
10	37	male	29.83	2	no	northeast	6406.4107
11	60	female	25.84	0	no	northwest	28923.1369
12	25	male	26.22	0	no	northeast	2721.3208
13	62	female	26.29	0	yes	southeast	27808.7251
14	23	male	34.4	0	no	southwest	1826.843
15	56	female	39.82	0	no	southeast	11090.7178

Figure 1. Observations 1-14 from the original health insurance dataset.

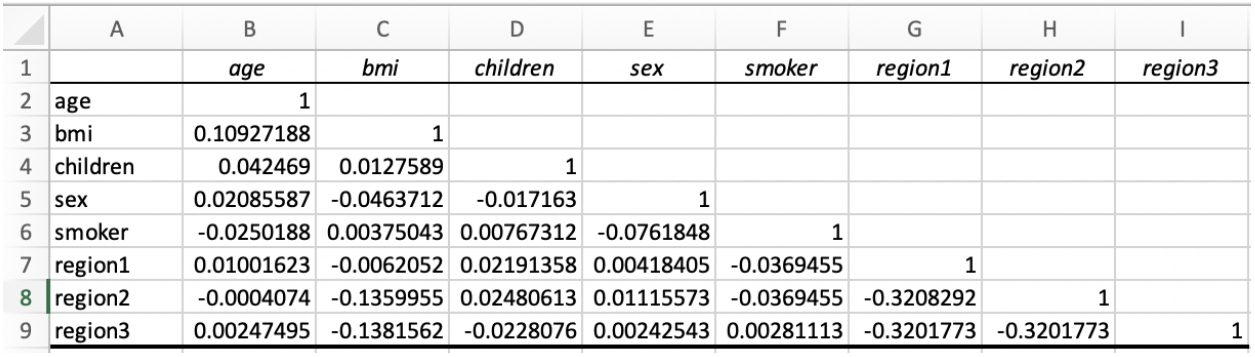
The three categorical variables are sex, smoker, and region. Dummy variables were created for each of the three categorical variables. A dummy variable can only take on the value of 0 or 1. For sex, only one dummy variable called sex_dummy was created. If the sex was female, sex_dummy was assigned a value of 1. If the sex was male, sex_dummy was given a value of 0. For the smoker variable in Column E in Figure 1, only one dummy variable called smoker_dummy was created too. If the individual was a smoker and thus had a value of “yes” in Column E in Figure 1, smoker_dummy was given a value of 1. If the individual was not a smoker and thus had a value of “no” in Column E in Figure 1, smoker_dummy was assigned a value of 0. One dummy variable with values of 0 or 1 can only account for two different categorical levels. Similarly, two dummy variables with values of 0 or 1 can only account for three different levels. Thus, for the region variable in Column F in Figure 1, where there are four levels, three dummy variables called region1, region2, and region3 were created. If the individual was from the southwest region, the dummy variables region1, region2, and region3 were assigned values 1, 0, and 0, correspondingly. If the individual was from the northwest region, the dummy variables region1, region2, and region3 were assigned values 0, 1, and 0, correspondingly. If the individual was from the northeast region, the dummy variables region1, region2, and region3 were assigned values 0, 0, and 1, correspondingly. If the individual was from the southeast region, the dummy variables region1, region2, and region3 were assigned values 0, 0, and 0, correspondingly. Figure 2 displays the first 14 observations with dummy variables.



	A	B	C	D	E	F	G	H	I
1	charges	age (x1)	bmi (x2)	children (x3)	sex_dummy (x4)	smoker_dummy (x5)	region1 (x6)	region2 (x7)	region3 (x8)
2	16884.924	19	27.9	0	1	1	1	0	0
3	1725.5523	18	33.77	1	0	0	0	0	0
4	4449.462	28	33	3	0	0	0	0	0
5	21984.4706	33	22.705	0	0	0	0	1	0
6	3866.8552	32	28.88	0	0	0	0	1	0
7	3756.6216	31	25.74	0	1	0	0	0	0
8	8240.5896	46	33.44	1	1	0	0	0	0
9	7281.5056	37	27.74	3	1	0	0	1	0
10	6406.4107	37	29.83	2	0	0	0	0	1
11	28923.1369	60	25.84	0	1	0	0	1	0
12	2721.3208	25	26.22	0	0	0	0	0	1
13	27808.7251	62	26.29	0	1	1	0	0	0
14	1826.843	23	34.4	0	0	0	1	0	0
15	11090.7178	56	39.82	0	1	0	0	0	0

Figure 2. Observations 1-14 from the health insurance dataset with dummy variables.

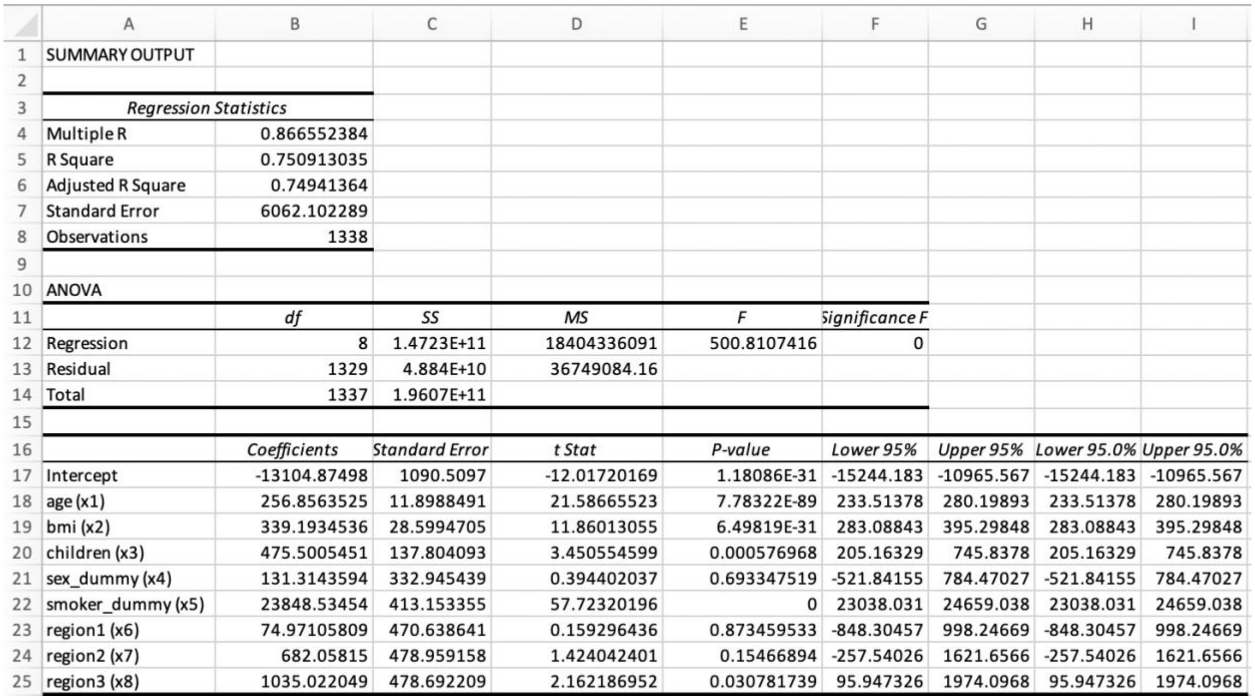
After converting the categorical variables into dummy variables, the independent variables were tested for multicollinearity. Multicollinearity is when two highly correlated independent variables can cause the estimated slope coefficients to be inaccurate and some independent variables to seem statistically insignificant when they are significant. As a result, it could misrepresent the data and lead to misleading conclusions. If the absolute value of the sample correlation coefficient was higher than 0.7, the two independent variables were determined to be too highly correlated. As seen in Figure 3, a correlation matrix was created to identify the sample correlation coefficients. The largest absolute value of the sample correlation coefficients was 0.3208292 in cell G8 in Figure 3. Thus, there was no multicollinearity.



	A	B	C	D	E	F	G	H	I
1		*age*	*bmi*	*children*	*sex*	*smoker*	*region1*	*region2*	*region3*
2	age	1							
3	bmi	0.10927188	1						
4	children	0.042469	0.0127589	1					
5	sex	0.02085587	-0.0463712	-0.017163	1				
6	smoker	-0.0250188	0.00375043	0.00767312	-0.0761848	1			
7	region1	0.01001623	-0.0062052	0.02191358	0.00418405	-0.0369455	1		
8	region2	-0.0004074	-0.1359955	0.02480613	0.01115573	-0.0369455	-0.3208292	1	
9	region3	0.00247495	-0.1381562	-0.0228076	0.00242543	0.00281113	-0.3201773	-0.3201773	1

Figure 3. The correlation matrix that identifies the sample correlation coefficients.

The multiple linear regression analysis was then performed without removing any independent variables since there was no multicollinearity in the data. Figure 4 displays the output of the multiple linear regression analysis performed on Microsoft Excel. The output contains general regression statistics about the overall multiple linear regression analysis, the ANOVA table, and statistical information such as the coefficient and P-value of each independent variable.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.866552384							
5	R Square	0.750913035							
6	Adjusted R Square	0.74941364							
7	Standard Error	6062.102289							
8	Observations	1338							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	8	1.4723E+11	18404336091	500.8107416	0			
13	Residual	1329	4.884E+10	36749084.16					
14	Total	1337	1.9607E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	-13104.87498	1090.5097	-12.01720169	1.18086E-31	-15244.183	-10965.567	-15244.183	-10965.567
18	age (x1)	256.8563525	11.8988491	21.58665523	7.78322E-89	233.51378	280.19893	233.51378	280.19893
19	bmi (x2)	339.1934536	28.5994705	11.86013055	6.49819E-31	283.08843	395.29848	283.08843	395.29848
20	children (x3)	475.5005451	137.804093	3.450554599	0.000576968	205.16329	745.8378	205.16329	745.8378
21	sex_dummy (x4)	131.3143594	332.945439	0.394402037	0.693347519	-521.84155	784.47027	-521.84155	784.47027
22	smoker_dummy (x5)	23848.53454	413.153355	57.72320196	0	23038.031	24659.038	23038.031	24659.038
23	region1 (x6)	74.97105809	470.638641	0.159296436	0.873459533	-848.30457	998.24669	-848.30457	998.24669
24	region2 (x7)	682.05815	478.959158	1.424042401	0.15466894	-257.54026	1621.6566	-257.54026	1621.6566
25	region3 (x8)	1035.022049	478.692209	2.162186952	0.030781739	95.947326	1974.0968	95.947326	1974.0968

Figure 4. The output of the multiple linear regression analysis after checking for multicollinearity. The output contains the ANOVA table, overall regression statistics such as the R^2, and the statistics about each independent variable.

The regression statistics near the top of the summary output of the multiple linear regression analysis contains statistical information about the overall model. The R^2 of 0.750913035 in cell B5 in Figure 4 indicates that the independent variables in this multiple linear regression model can explain around 75.09% of the variability in the annual cost of health insurance premium. However, the independent variables in this multiple linear regression model cannot explain around 24.91% of the variability in the costs of health insurance premium, making this model a moderately good fit for the used dataset.

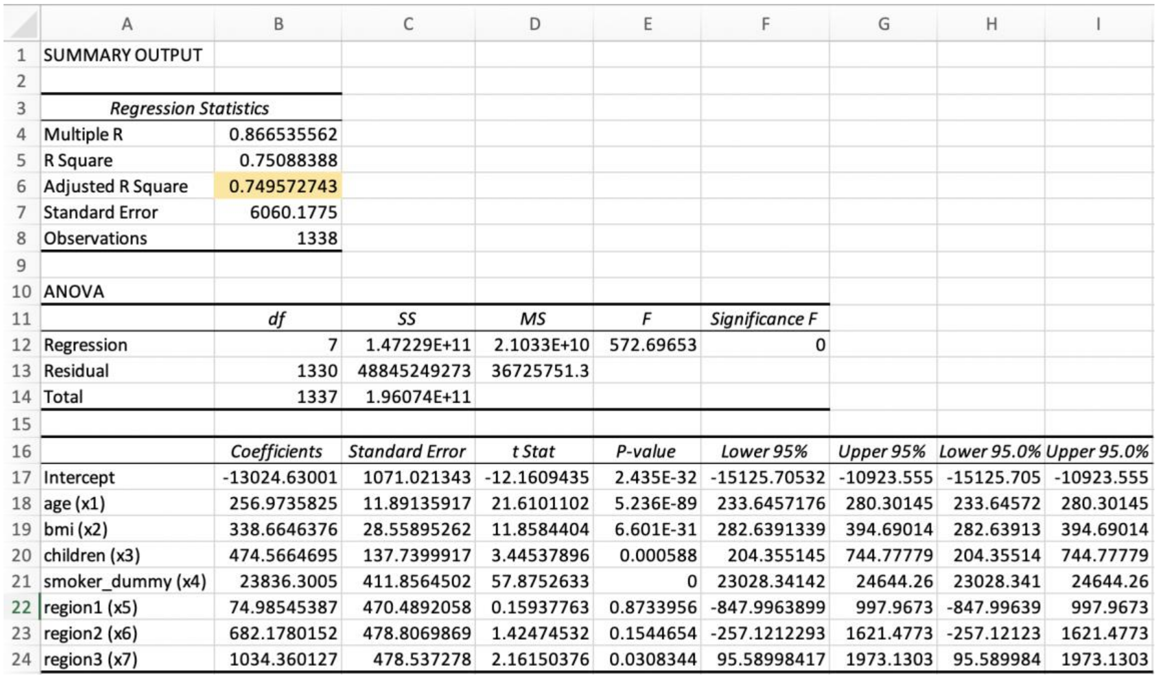
In the ANOVA table in Figure 4, the F test statistic in cell E12, derived by dividing MSR by MSE, is used to perform an F test. In the F test, the null hypothesis is that the coefficients of all eight independent variables in the multiple linear regression model are equal to zero: $\mathrm{H}_0\colon \beta_1 =$

$\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$. The alternative hypothesis is that at least one of the eight coefficients is not equal to zero: H$_a$: $\beta_1 \neq 0$ or/and $\beta_2 \neq 0$ or/and $\beta_3 \neq 0$ or/and $\beta_4 \neq 0$ or/and $\beta_5 \neq 0$ or/and $\beta_6 \neq 0$ or/and $\beta_7 \neq 0$ or/and $\beta_8 \neq 0$. The alpha in this model is 10% or 0.10. According to Figure 4, the P-value corresponding to an F value of around 500.81 with 8 and 1329 degrees of freedom in the numerator and denominator, correspondingly, is 0. Using the P-value approach, since the P-value, 0, is lower than the alpha, 0.10, the null hypothesis that the coefficients of all eight of the independent variables are equal to 0 can be rejected. Thus, the F test indicates an overall statistically significant relationship between the independent variables and the annual cost of health insurance premium.

According to cells A17:A25 and B17:B25 in Figure 4, the estimated multiple linear regression equation is thus ŷ $= -13104.87498 + 256.8563525 * age + 339.1934536 * bmi + 475.5005451 * children + 131.3143594 * sex\_dummy + 23848.53454 * smoker\_dummy + 74.97105809 * region1 + 682.05815 * region2 + 1035.022049 * region3$. For each independent variable, a t-test is performed. The t-test will reveal if each of the independent variables is statistically significant. In the t-test, the null hypothesis is that the variable's coefficient equals 0 for each independent variable: H$_0$: $\beta_i = 0$ (i is a value ranging from 1-8 representative of the coefficient for each of the eight independent variables). The alternative hypothesis is that the variable's coefficient does not equal to 0: H$_0$: $\beta_i \neq 0$ (i is a value ranging from 1-8 representative of the coefficient for each of the eight independent variables). The test statistic t in cells D17:D25 in Figure 4 is derived by dividing the sample statistic of b_i, the coefficient, by the standard deviation of b_i for each of the eight independent variables. The alpha for the t-test remains 10% or 0.10.

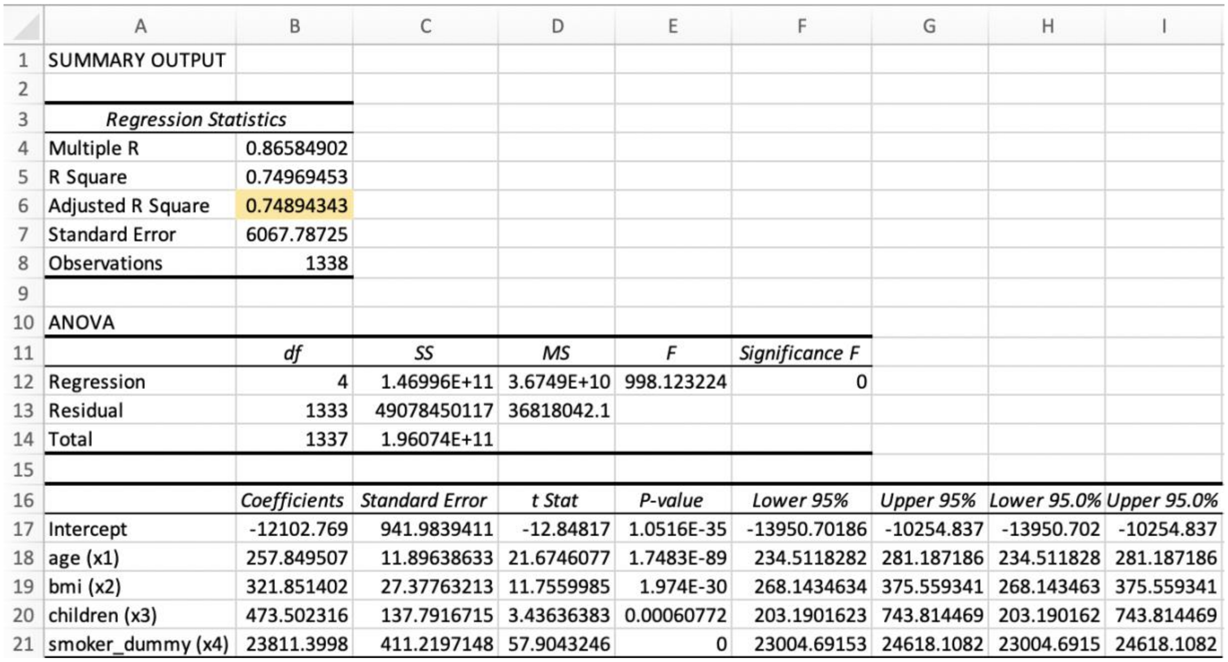
According to cells E17:E20, E22, and E25 in Figure 4, the coefficients of the intercept and age, bmi, children, smoker_dummy, and region3 independent variables had P-values lower than 0.10. The P-values were 1.18086E-31, 7.78322E-89, 6.49819E-31, 0.000576968, 0, and 0.030781739, respectively. Since the P-values are lower than alpha, the null hypothesis that the coefficients of the intercept and these five independent variables are equal to zero can be rejected. Thus, the t-test indicates that the intercept and the age, bmi, children, smoker_dummy, and region3 independent variables have a statistically significant relationship with the annual cost of health insurance premium. However, according to cells E21 and E23:E24 in Figure 4, the coefficients of the sex_dummy, region1, and region2 independent variables had P-values greater than 0.10. The P-values were 0.693347519, 0.873459533, and 0.15466894 respectively. Thus, there is only around 31%, 12.6%, and 84.5% confidence, correspondingly, that these coefficients are not equal to 0. Since the P-values are greater than alpha, the null hypothesis that the coefficients of these three independent variables are equal to zero cannot be rejected. Thus, the t-test shows that the sex_dummy, region1, and region2 independent variables are not statistically significant.

Since there is not enough evidence to conclude a significant relationship between each of the three statistically insignificant independent variables and the annual cost of health insurance premium, other multiple linear regression models were constructed by eliminating combinations of these variables.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.866535562							
5	R Square	0.75088388							
6	Adjusted R Square	0.749572743							
7	Standard Error	6060.1775							
8	Observations	1338							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	7	1.47229E+11	2.1033E+10	572.69653	0			
13	Residual	1330	48845249273	36725751.3					
14	Total	1337	1.96074E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	-13024.63001	1071.021343	-12.1609435	2.435E-32	-15125.70532	-10923.555	-15125.705	-10923.555
18	age (x1)	256.9735825	11.89135917	21.6101102	5.236E-89	233.6457176	280.30145	233.64572	280.30145
19	bmi (x2)	338.6646376	28.55895262	11.8584404	6.601E-31	282.6391339	394.69014	282.63913	394.69014
20	children (x3)	474.5664695	137.7399917	3.44537896	0.000588	204.355145	744.77779	204.35514	744.77779
21	smoker_dummy (x4)	23836.3005	411.8564502	57.8752633	0	23028.34142	24644.26	23028.341	24644.26
22	region1 (x5)	74.98545387	470.4892058	0.15937763	0.8733956	-847.9963899	997.9673	-847.99639	997.9673
23	region2 (x6)	682.1780152	478.8069869	1.42474532	0.1544654	-257.1212293	1621.4773	-257.12123	1621.4773
24	region3 (x7)	1034.360127	478.537278	2.16150376	0.0308344	95.58998417	1973.1303	95.589984	1973.1303

Figure 5. The output of the multiple linear regression analysis without adding sex as an independent variable.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.86584902							
5	R Square	0.74969453							
6	Adjusted R Square	0.74894343							
7	Standard Error	6067.78725							
8	Observations	1338							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	4	1.46996E+11	3.6749E+10	998.123224	0			
13	Residual	1333	49078450117	36818042.1					
14	Total	1337	1.96074E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	-12102.769	941.9839411	-12.84817	1.0516E-35	-13950.70186	-10254.837	-13950.702	-10254.837
18	age (x1)	257.849507	11.89638633	21.6746077	1.7483E-89	234.5118282	281.187186	234.511828	281.187186
19	bmi (x2)	321.851402	27.37763213	11.7559985	1.974E-30	268.1434634	375.559341	268.143463	375.559341
20	children (x3)	473.502316	137.7916715	3.43636383	0.00060772	203.1901623	743.814469	203.190162	743.814469
21	smoker_dummy (x4)	23811.3998	411.2197148	57.9043246	0	23004.69153	24618.1082	23004.6915	24618.1082

Figure 6. The output of the multiple linear regression analysis without adding sex, region1, region2, and region3 as independent variables.

The adjusted R^2 is used to compare how well the independent variables explain the variability in the annual cost of health insurance premium across different models. R^2 increases as more independent variables are added to the model. However, the adjusted R^2 counterbalances this increase by decreasing the R^2 value for every added independent variable. The higher the adjusted R^2 value, the more the independent variables in the multiple linear regression model can explain the variability in the annual cost of health insurance premium in America. Comparing the adjusted R^2 value in cell B6 in Figure 4 with that in cell B6 in Figures 5 and 6, it is the greatest in Figure 5.

In Figure 5, sex was not included as an independent variable in the multiple linear regression model as it was identified as statistically insignificant using the t-test in Figure 4. When sex was eliminated, the adjusted R^2 value, 0.749572743, became slightly higher than the adjusted R^2 value in Figure 4, 0.74931364. Since the adjusted R^2 is used to compare different multiple linear regression models, it indicates that Figure 5, the model without sex as an independent variable, best fits the health insurance data. Thus, according to cells A17:A25 and B17:B25 in Figure 5, the new estimated multiple linear regression equation is $\hat{y} = -13024.63001 + 256.97735825 * age + 338.6646376 * bmi + 474.5664695 * children + 23846.3005 * smoker\_dummy + 74.98545387 * region1 + 682.1780152 * region2 + 1034.360127 * region3$. Though the P-values corresponding to the new test statistics for region1, 0.87339556, and region2, 0.15446543, are still higher than the alpha, 0.10, the estimated multiple linear regression equation derived from Figure 5 captures the most variability in the annual cost of health insurance premium in America.

The coefficient of -13024.63001 for the intercept is negative in cell B17 in Figure 5. It indicates that the firm supplying health insurance would owe the individual around 13,024.63

dollars annually if the individual had an age of zero, BMI of zero, zero children, was not a smoker, and was from the northwest based on the estimated multiple linear regression equation from Figure 5. However, a living individual cannot have a BMI and age of zero. In America, an individual may remain under their parent's health insurance plan until the individual is 26 years old. Additionally, the average healthy BMI of a young adult in America is 22 $\frac{kg}{m^2}$. According to cell B18 in Figure 5, the coefficient for the independent variable age indicates that for every increase in an individual's age by one year, an individual's annual cost of health insurance premium increases by 256.9735825 dollars. Thus, for a 26-year-old, since $256.9735825\ dollars * 26\ \approx 6{,}681.31\ dollars$, the annual cost of health insurance premium would be around 6,681.31 dollars when just considering age. According to cell B19 in Figure 5, the coefficient for the independent variable bmi shows that for every increase in an individual's BMI by 1.0, an individual's annual cost of health insurance premium increases by 338.6646376 dollars. Thus, for an individual with a BMI of 22 $\frac{kg}{m^2}$, since $338.6646376\ dollars * 22\ \approx 7{,}450.62\ dollars$, the annual cost of health insurance premium would be around 7,450.62 dollars when just considering BMI. So, for an individual with normal BMI and age, since $6{,}681.31 +\ 7{,}450.62 = 14{,}131.93\ dollars > 13{,}024.63\ dollars$, the negative magnitude of the intercept is countered. In fact, the annual cost of health insurance premium becomes positive. Thus, the model supports the accepted logic that an average individual must pay an insurance company for health coverage services.

Furthermore, according to cell B20 in Figure 5, the coefficient of 474.5664695 for the independent variable children indicates that for every additional child an individual has, the annual cost of health insurance premium increases by 474.5664695 dollars. However, one shortcoming of the model with the higher adjusted R2 in Figure 5 is that it does not consider that women may have post-delivery complications. In effect, women may potentially have higher premium costs than men. However, the adjusted R^2 of the model is lesser when including sex as an independent variable. According to cell B21 in Figure 5, the coefficient of 23,836.3005 for the independent variable smoker_dummy shows that if an individual is a smoker, has a value of 1 for the dummy variable smoker, the annual cost of health insurance premium increases by 23,836.3005 dollars. Additionally, according to cells E22 and E23 in Figure 5, the P-values from the t-test for the independent variables region1 and region2 are higher than the alpha of 0.10. Thus, the cost of annual health insurance premium for an individual from the southwest or northwest is not different from that of an individual from the southeast. However, based on cells B24 and E24 in Figure 5, since the P-value from the t-test for the independent variable region3 is lower than the alpha, being from the northeast increases an individual's annual cost of health care insurance premium by 1,034.360127 dollars.

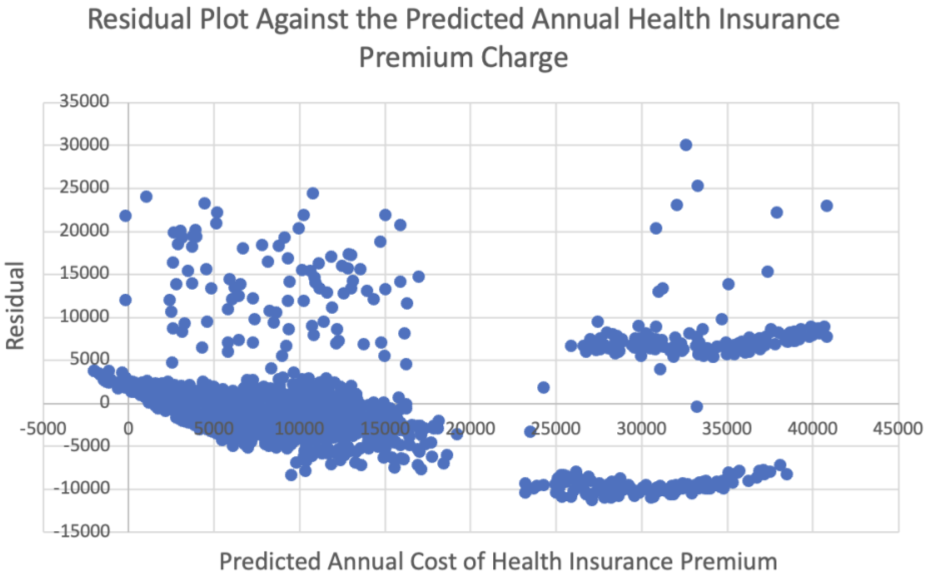


Figure 7. Residual plot against the predicted annual cost of health insurance premium.

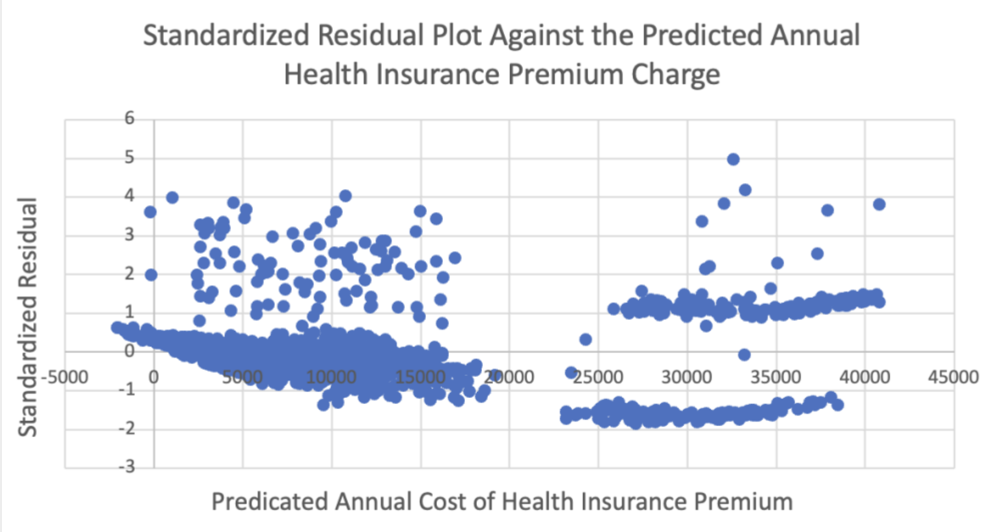


Figure 8. Standardized residual plot against the predicted annual cost of health insurance premium.

As seen in Figures 7 and 8, residual analysis was performed to determine if linear regression analysis is appropriate for this health insurance dataset. Since a multiple linear regression analysis was used, it is more appropriate to employ a residual plot against the annual cost of health insurance premium. The residuals in the plot in Figure 7 do not form a horizontal pattern. In addition, many of the standardized residuals in the standardized residual plot are not between -2 and 2, indicating that the variance of the residuals is not constant for every predicted annual cost of health insurance premium. Thus, there is not enough evidence to confirm that multiple linear regression analysis is appropriate for this dataset. To further explore which independent variables were leading to the nonlinearity, the residuals were plotted against each continuous numeric independent variable. The residuals were not plotted against dummy variables because they only have values of 0 and 1. Similarly, the residuals were not plotted against the independent variable children because it only has discrete values 0, 1, 2, 3, 4, and 5.

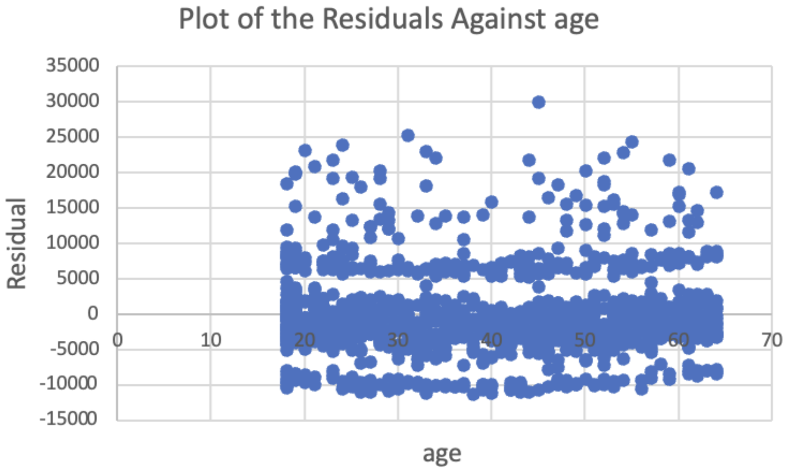


Figure 9. The residual plot of the residual against the age independent variable.

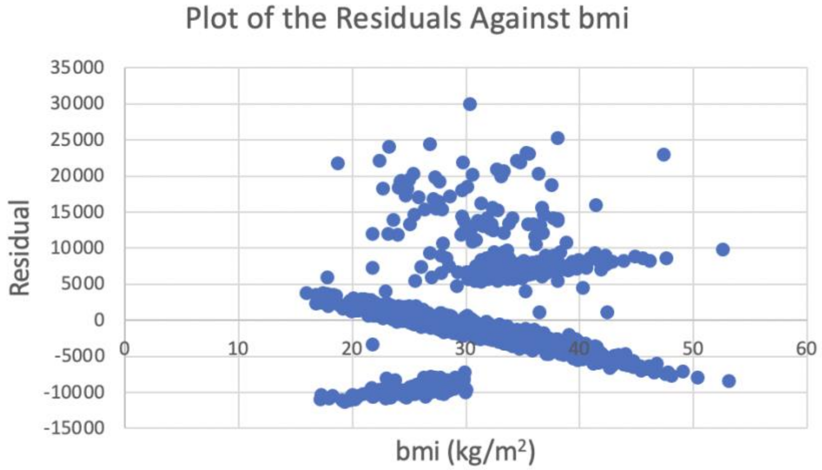


Figure 10. The residual plot of the residual against the bmi independent variable.

The residuals in Figure 9 form a horizontal pattern, indicating that the variance of the residuals is very similar for all values of the age variable. Thus, the age data is linear. However, the residuals in Figure 10 do not form a horizontal pattern, indicating that the variance of the residuals is nonconstant for all values of the bmi variable. Thus, the BMI data has nonlinearity.

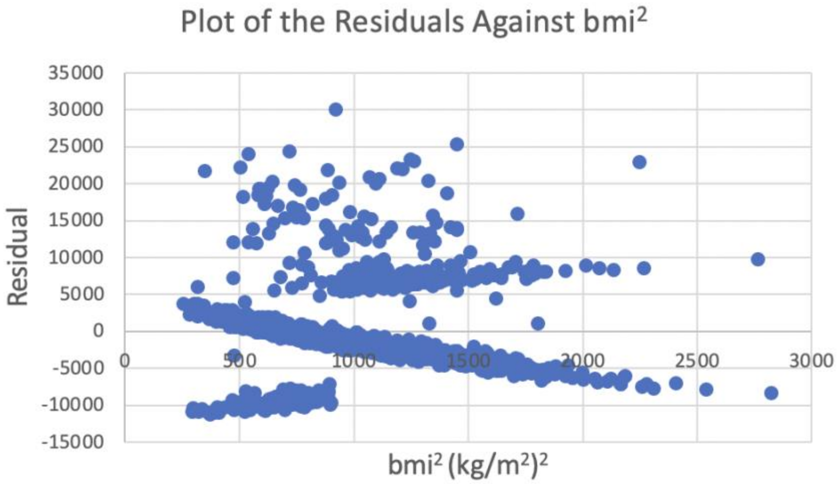


Figure 11. The residual plot of the residual against the newly added bmi^2 independent variable.

The values of the bmi variable were squared and assigned to a newly independent variable, bmi2, to reduce the nonlinearity in the BMI data. The residuals in Figure 11 still do not form a horizontal pattern, indicating that the variance of the residuals is still nonconstant for all values of bmi^2. I do not know how to address this nonlinearity issue in the data using my limited current knowledge. However, it must be addressed to be confident that a multiple linear regression analysis is appropriate to explain the annual cost of health insurance premium using this dataset. According to cell B5 in Figure 5, another limitation to this analysis is that the independent variables in the model only explain around 75.09% of the variability in the annual cost of health insurance premium in America. A solution is to add more independent variables. Categorical variables such as an individual's occupation could affect the annual cost of health insurance premium. For instance, the premium could increase if the individual works in physically or emotionally demanding environments like construction sites. Another potentially significant categorical variable would identify whether the individual has a group health insurance plan or not. A potentially significant numerical independent variable would be the number of pre-existing health conditions the individual has. However, I could not find data about these suggested independent variables for the individuals in this dataset.

The R^2 for the model with the highest adjusted R^2 was 0.749572743. Thus, the estimated multiple linear regression equation $\hat{y} = -13024.63001 + 256.97735825 * age + 338.6646376 * bmi + 474.5664695 * children + 23846.3005 * smoker\_dummy + 74.98545387 * region1 + 682.1780152 * region2 + 1034.360127 * region3$ is a moderately good fit for this health insurance data. The P-value of the test statistic F from the F-test in cell F12 in Figure 5 is lower than the alpha of 0.10. Thus, there is an overall significance between the independent variables and dependent variable, the annual cost of health insurance premium. However, the independent variables sex_dummy, region1, and region2 were not statistically significant in explaining the annual cost of health insurance premium based on corresponding t-tests. The independent variables age, bmi, children, smoker_dummy, and region3 have a statistically significant relationship with the annual cost of health insurance premium. Lastly, the residual analysis showed that a multiple linear regression model might not be the most effective for this health insurance dataset since the residuals did not form a horizontal pattern when plotted against the predicted annual costs of health insurance premium. Thus, methods besides squaring the values of an independent variable that I have not yet learned in econometrics must be employed to reduce the nonlinearity within the data. Another approach would be to conclude that multiple linear regression analysis is not appropriate and analyze the health insurance data using a different statistical model. Additionally, more independent variables, such as the individual’s occupation and pre-existing health conditions, need to be analyzed to attempt to explain the nearly 24.91% variability in the annual cost of health insurance premium that the current model in Figure 5 cannot explain. Thus, the multiple linear regression analysis is not complete.