A one-way analysis relating bmi to region. Use contrasts and appropriate post-hoc comparisons and discuss the results.

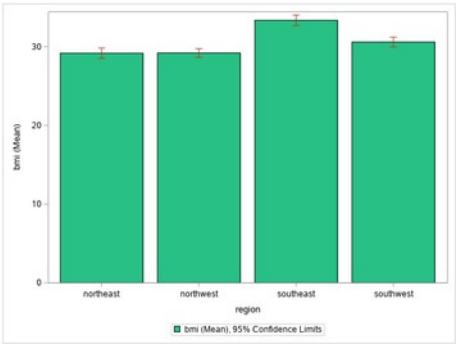


Figure 1

Figure 1 shows diagram with 4 regions which represents mean, and upper lines represent confidence limit. The mean summarizes the sample data, and confidence limits gives us some information about the level of variability in mean in populations. It gives us some information how accurate the sample mean is and idea of population mean might be. At quick look there is obvious differences between the bmi in southeast and other regions. The average of northeast, northwest and southwest with confidence limit implies the difference may not be significant. We discuss whether these differences is statistically different.

For one-way ANOVA we required to check some assumptions. First, the data is from four different groups(regions), it is assumed to have random independent samples(there is no dependency issue). Second, we need to check normality for samples. Third, standard deviations for populations should be equal.

The UNIVARIATE Procedure
Variable: bmi
region = northeast

Moments			
N	324	Sum Weights	324
Mean	29.1735031	Sum Observations	9452.215
Std Deviation	5.9375133	Variance	35.2540642
Skewness	0.22942025	Kurtosis	-0.2883316
Uncorrected SS	287141.286	Corrected SS	11387.0627
Coeff Variation	20.352418	Std Error Mean	0.32986185

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.992033	Pr < W	0.0796
Kolmogorov-Smirnov	D	0.036625	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.072784	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.549052	Pr > A-Sq	0.1617

Table 2

The UNIVARIATE Procedure
Variable: bmi
region = northwest

Moments			
N	325	Sum Weights	325
Mean	29.1997846	Sum Observations	9489.93
Std Deviation	5.13676496	Variance	26.3863543
Skewness	0.04429555	Kurtosis	-0.2965334
Uncorrected SS	285653.091	Corrected SS	8549.17878
Coeff Variation	17.5917906	Std Error Mean	0.28493645

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.995465	Pr < W	0.4656
Kolmogorov-Smirnov	D	0.026358	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.022911	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.175029	Pr > A-Sq	>0.2500

Table 1

The UNIVARIATE Procedure
Variable: bmi
region = southwest

Moments			
N	325	Sum Weights	325
Mean	30.5966154	Sum Observations	9943.9
Std Deviation	5.69183579	Variance	32.3969947
Skewness	0.16018381	Kurtosis	-0.1234235
Uncorrected SS	314746.31	Corrected SS	10496.6263
Coeff Variation	18.6028282	Std Error Mean	0.31572624

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.994927	Pr < W	0.3630
Kolmogorov-Smirnov	D	0.029848	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.035268	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.24266	Pr > A-Sq	>0.2500

Table 4

The UNIVARIATE Procedure
Variable: bmi
region = southeast

Moments			
N	364	Sum Weights	364
Mean	33.355989	Sum Observations	12141.58
Std Deviation	6.47764793	Variance	41.9599227
Skewness	0.22035026	Kurtosis	-0.2994397
Uncorrected SS	420225.861	Corrected SS	15231.4519
Coeff Variation	19.4197448	Std Error Mean	0.33952101

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.991089	Pr < W	0.0270
Kolmogorov-Smirnov	D	0.036843	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.057915	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.494575	Pr > A-Sq	0.2219

Table 3

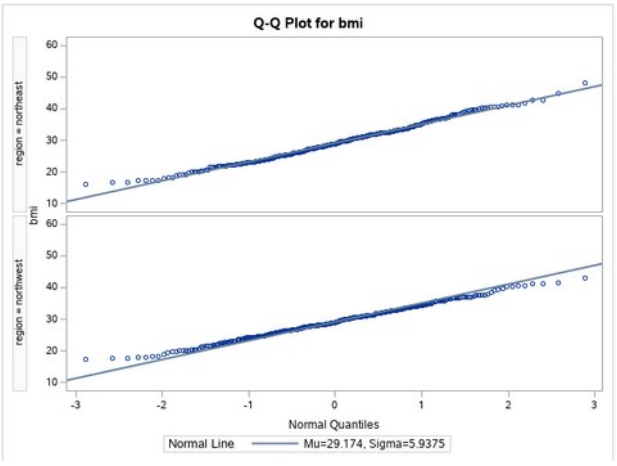
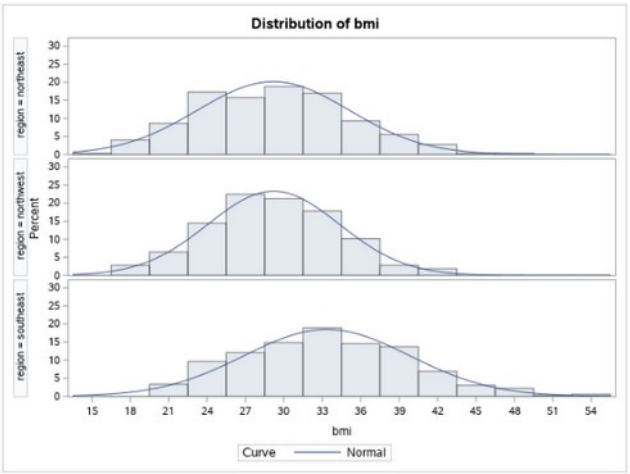


Figure 3



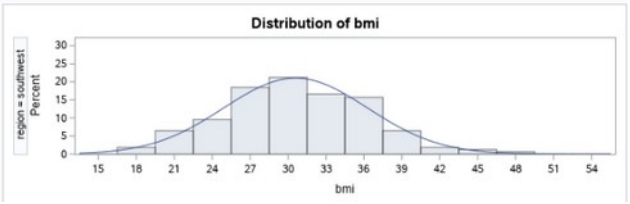


Figure 2

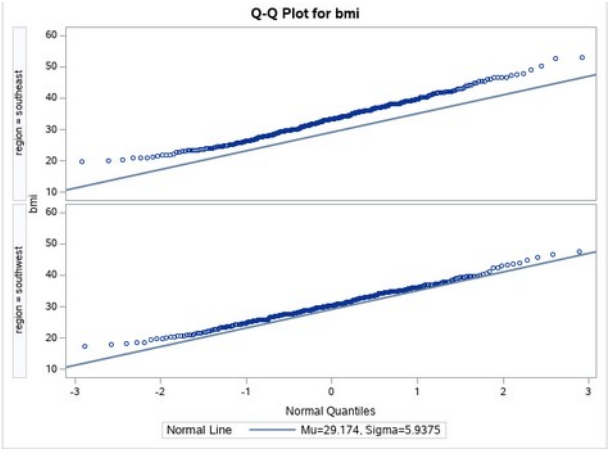


Figure 4

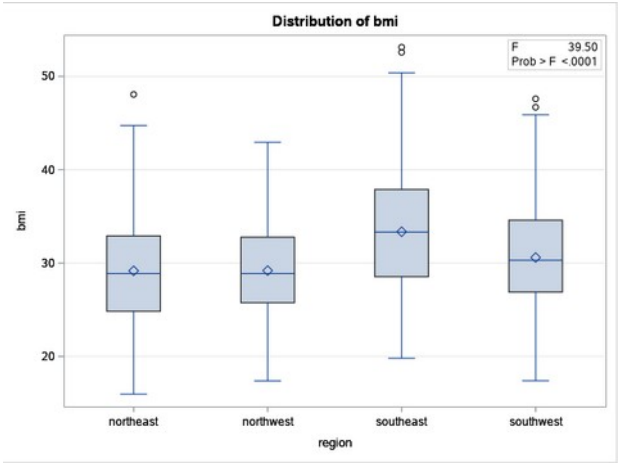


Figure 5

Test of normality for northeast, northwest and southwest show that the p-value for Normality tests are greater than 0.05 (we can not reject the assumption of normality) which indicate the normality of populations. Shapiro test for southeast is 0.02 which indicates the assumption of normality is rejected while other tests are greater than 0.05 which imply the normality of the population. (according to research: shapiro test mostly used for small samples and in this case we can decide according to other tests).
Distribution plots for northeast and northwest show somehow a symmetric plot which imply a normal distribution, but according to moment table a right-skewed(positive skewness) and light tail(negative kurtosis) is obvious in plot. Also, Q-Q plot indicate somehow no deviation from straight line in most part of the line.
Distribution plots for southeast show a symmetric pattern which imply a normal distribution.
Distribution plots for southwest show somehow a symmetric plot which imply a normal distribution, but according to moment table, there is a right-skewed and light tail. There is negligible deviation at both end of Q-Q plot.

According to tests the assumption of normality can not be rejected. Also Boxplots suggest a small variability in variance, but we should do homogeneity of variance test to measure.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4055.88063	1351.96021	39.50	<.0001
Error	1334	45664.31975	34.23112		
Corrected Total	1337	49720.20039			

R-Square	Coeff Var	Root MSE	bmi Mean
0.081574	19.08052	5.850737	30.66340

Source	DF	Type I SS	Mean Square	F Value	Pr > F
region	3	4055.880631	1351.960210	39.50	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
region	3	4055.880631	1351.960210	39.50	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	30.59661538	B	0.32454050	94.28	<.0001
region northeast	-1.42311230	B	0.45932358	-3.10	0.0020
region northwest	-1.39683077	B	0.45896958	-3.04	0.0024
region southeast	2.75937363	B	0.44650654	6.18	<.0001
region southwest	0.00000000	B	.	.	.

Table 5

Before interpreting above table we should check the equal variance assumption through a test of homogeneity in Table 6.

Levene's Test for Homogeneity of bmi Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
region	3	42988.4	14329.5	6.91	0.0001
Error	1334	2765947	2073.4		

Welch's ANOVA for bmi			
Source	DF	F Value	Pr > F
region	3.0000	35.83	<.0001
Error	739.5		

Table 6

By interpreting Table 6, we can examine results of a test of homogeneity of variance. P-value is less than 0.05, so we can reject the equal variance assumption. There are significant differences in variance across four regions. So, For our data we need to report F-ratio and P-value of Welch's ANOVA instead of main ones. The p-value in table 6 is smaller than 0.05 which imply the null hypothesis of equal means in all samples is rejected and hence there is a significant difference in the mean of bmi in some regions. F(3,739) = 35.83, P-value <0.0001.

The parameter estimate in table 5 indicates that all parameter are statistically significant(P-values < 0.05). The intercept (30.6) only corresponds to the line of southwest and imply the sample mean. Other estimates correspond to northeast, southeast and northwest explain the correction to region southwest. Parameter estimate shows that the bmi in northeast and northwest were significantly lower than southwest. Also, the southeast is significantly higher that southwest.

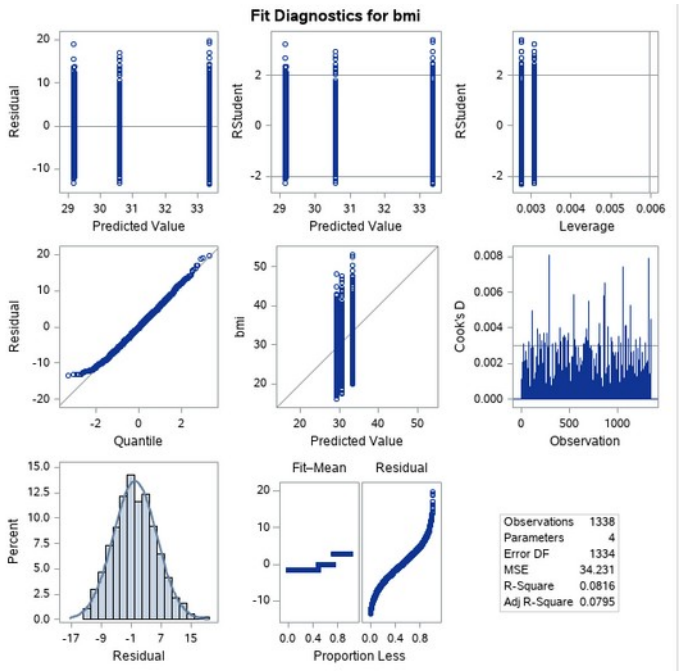


Figure 6

Figure 6 shows fit diagnostics, and interpretation of residual versus predicted value plots implies the unequal variability of residual in 4 different lines which relate to four regions and consistent with Levene test of equality of variance. Predicted value for two region are close to each other, which cause two line some how located on each other. Furthermore, the plot of Studentised residuals indicates some points outside the -2 and 2 bounds but most of them are close to bounds. The leverage plots indicate no point of high leverage. Cook's D shows some value above rule of thumb which is not extreme(not above the cause of concern). Q-Q plot and histogram imply the approximately normal distribution for residuals.

Comparisons significant at the 0.05 level are indicated by ***.				
region Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
southeast - southwest	2.7594	1.6109	3.9079	***
southeast - northwest	4.1562	3.0077	5.3047	***
southeast - northeast	4.1825	3.0330	5.3319	***
southwest - southeast	-2.7594	-3.9079	-1.6109	***
southwest - northwest	1.3968	0.2163	2.5774	***
southwest - northeast	1.4231	0.2416	2.6046	***
northwest - southeast	-4.1562	-5.3047	-3.0077	***
northwest - southwest	-1.3968	-2.5774	-0.2163	***
northwest - northeast	0.0263	-1.1552	1.2078	
northeast - southeast	-4.1825	-5.3319	-3.0330	***
northeast - southwest	-1.4231	-2.6046	-0.2416	***
northeast - northwest	-0.0263	-1.2078	1.1552	

Table 7

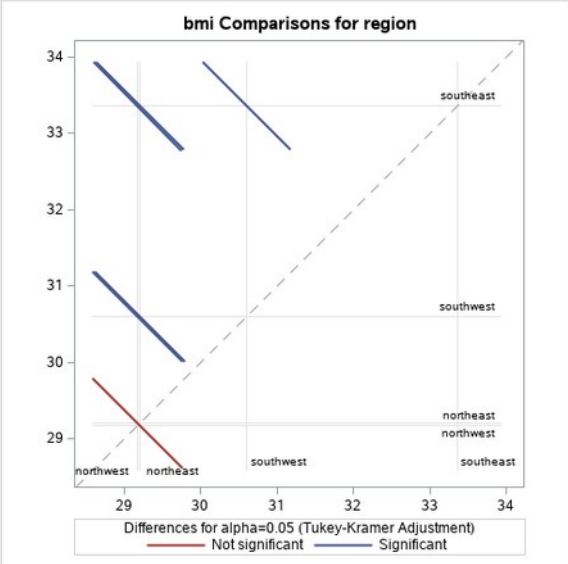


Figure 7

region	bmi LSMEAN	LSMEAN Number
northeast	29.1735031	1
northwest	29.1997846	2
southeast	33.3559890	3
southwest	30.5966154	4

Least Squares Means for effect region Pr > |t| for H0: LSMean(i)=LSMean(j) Dependent Variable: bmi				
i/j	1	2	3	4
1		0.9999	<.0001	0.0107
2	0.9999		<.0001	0.0127
3	<.0001	<.0001		<.0001
4	0.0107	0.0127	<.0001	

Table 8

ANOVA cannot indicate the significant difference between regions separately and we should apply Post-hoc test to indicate differences.
Table 7 shows the comparison significance at 5% level. The only means that is not significantly different is northwest and northeast. This is confirmed by table 8 and diffogram shown in Figure 7. In diffogram only the red line located on intersection of northwest and northeast and cross the zero line is not significant. Other lines did not cross zero lines which means are significant. It also indicate that bmi in southeast is significantly higher than other regions.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Northeast vs other regions	-5.63187975	1.12051737	-5.03	<.0001
Northeast vs Southeast	-4.18248592	0.44687042	-9.36	<.0001
Northeast vs Northwest	-0.02628153	0.45932358	-0.06	0.9544

Table 9

For contrast the mean for Northeast versus other region we put 3,-1,-1,-1 in code. Table 9 indicates the bmi average in Northeast differ significantly from bmi in the other regions(p-value<0.005). It also shows for Northeast versus Southeast, the weights for this comparison would be 1, 0, -1, 0, the bmi mean is statistically different. For contrast between Northeast and Northwest, by considering weights 1, -1, 0, 0 in code, the difference of means between these two regions is not statistically significant(P-value=0.9 >0.005) which confirm the difogram and least square mean in table 8.

A one-way ANCOVA relating bmi to region and age with age as a covariate, including appropriate post-hoc comparisons.

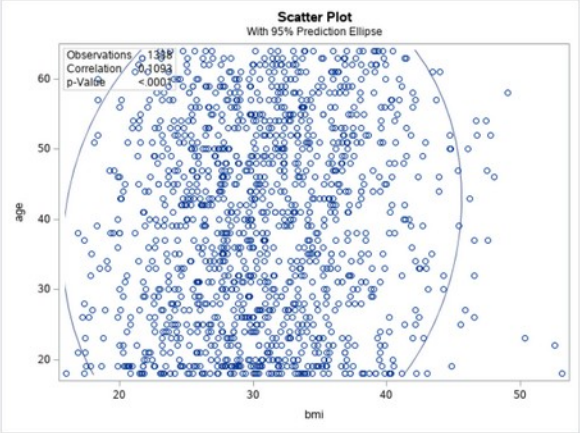


Figure 8

Pearson Correlation Coefficients, N = 1338 Prob > |r| under H0: Rho=0		
	bmi	age
bmi	1.00000	0.10927 <.0001
age	0.10927 <.0001	1.00000

Table 10

There is a weak positive correlation (small effect) between the bmi and age (r = 0.1). This correlation is statistically significant (P-value < 0.05). Also, the scatter plot implies a weak linear relationship between bmi as response and age as covariate, confirmed with pearson correlation coefficient.

Levene's Test for Homogeneity of age Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
region	3	5848.0	1949.3	0.07	0.9778
Error	1334	39295462	29456.9		

Table 11

We do one-way analysis of variance relating age and region to check the conditions for applying ANCOVA. First we should check the assumption of ANOVA, but we need to just check the relevant output to ANCOVA assumptions.
As LEVENE test indicates the satisfying of equality of variance in age between regions, we should read the following main table.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	47.3447	15.7816	0.08	0.9710
Error	1334	263878.3092	197.8098		
Corrected Total	1337	263925.6540			

Table 12

From Table 12, the difference in mean of age (covariate) between different regions is not significant (P-value > 0.05). So, we can assume that the age is independent of regions. So the independence assumption is satisfied.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4677.72950	1169.43238	34.61	<.0001
Error	1333	45042.47088	33.79030		
Corrected Total	1337	49720.20039			

R-Square	Coeff Var	Root MSE	bmi Mean
0.094081	18.95727	5.812943	30.66340

Source	DF	Type III SS	Mean Square	F Value	Pr > F
region	3	4084.053201	1361.351067	40.29	<.0001
age	1	621.848871	621.848871	18.40	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	28.68127106	B	0.55073871	52.08	<.0001
region northeast	-1.41404097	B	0.45636135	-3.10	0.0020
region northwest	-1.38428387	B	0.45601411	-3.04	0.0024
region southeast	2.78441408	B	0.44366060	6.28	<.0001
region southwest	0.00000000	B	.	.	.
age	0.04854456		0.01131603	4.29	<.0001

Table 13

According to test of ANCOVA, the overall model is statistically significant (p-value < 0.05), indicating regions and age explain something about response variable(bmi). Use partial sums of squares, Type III SS, factor and the covariate considered to provide significant contribution to the overall result (statistically significant at 5% level). Region and age are significant contributors to the model.
Table 13 shows the parameter estimate table. By looking to the last row we got the parameter estimate of age is about 0.048, which is a common slop of four regression lines (in figure 9) and is statistically significant(p-value< 0.05). The intercept (28.68) only corresponds to the line of region southwest. Estimate corresponding to northeast, northwest and southeast explain the correction to line southwest. As the table 13 implies northeast and northwest is lower than southwest, while southeast is above southwest which illustrate in figure 9. The estimate for northeast, northwest and southeast is statistically significant. Overall, Covariate can slightly improve the accountability of variables in model from 8% without age to about 9% with covariate. It diminished the effect.

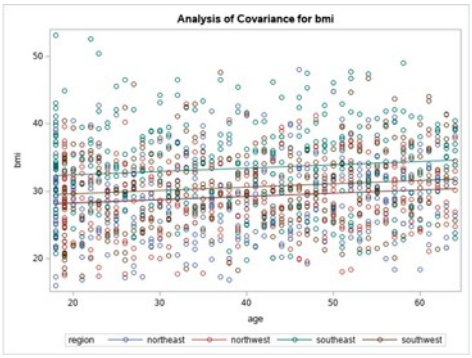


Figure 9

As figure 9 shows, the picture consist of 4 parallel lines. Each regression line relate to each region. Obviously they are parallel and have the same slope (we assume slops are equal) . According to the picture, southeast offer the highest bim score the regions.

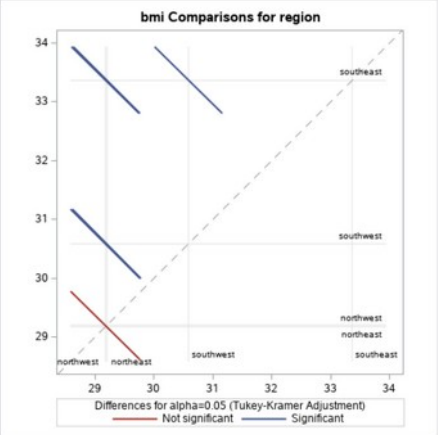


Figure 10

region	bmi LSMEAN	LSMEAN Number
northeast	29.1705179	1
northwest	29.2002750	2
southeast	33.3689730	3
southwest	30.5845589	4

Least Squares Means for effect region
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: bmi

i/j	1	2	3	4
1		0.9999	<.0001	0.0107
2	0.9999		<.0001	0.0131
3	<.0001	<.0001		<.0001
4	0.0107	0.0131	<.0001	

Table 14

Least square means table for dependent bmi and independent variable region and covariate age, shows the difference for all regions except for northeast and northwest, from each other is statistically significance. The Diffogram shows only line that cross zero line is on intersection of northwest and northeast, which is non significant. Other lines imply the significant difference. It also indicate that bmi in southeast is significantly higher than other regions.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	4912.23810	701.74830	20.83	<.0001
Error	1330	44807.96229	33.69020		
Corrected Total	1337	49720.20039			

R-Square	Coeff Var	Root MSE	bmi Mean
0.098798	18.92917	5.804326	30.66340

Source	DF	Type III SS	Mean Square	F Value	Pr > F
region	3	1074.349431	358.116477	10.63	<.0001
age	1	658.483542	658.483542	19.55	<.0001
age*region	3	234.508595	78.169532	2.32	0.0737

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	27.18045101	B	0.96658767	28.12	<.0001
region northeast	-0.63672341	B	1.36045884	-0.47	0.6398
region northwest	0.54625286	B	1.35906764	0.40	0.6878
region southeast	5.82418067	B	1.31463159	4.43	<.0001
region southwest	0.00000000	B	.	.	.
age	0.08658297	B	0.02309923	3.75	0.0002
age*region northeast	-0.01961392	B	0.03256578	-0.60	0.5471
age*region northwest	-0.04900143	B	0.03256075	-1.50	0.1326
age*region southeast	-0.07755982	B	0.03156195	-2.46	0.0141
age*region southwest	0.00000000	B	.	.	.

Table 15

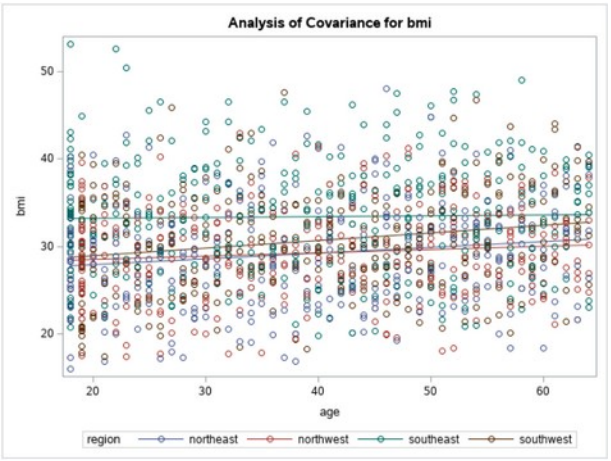


Figure 11

For testing homogeneity of slopes we rerun ANCOVA with additional term called interaction term(region*age). We use partial sums of squares, Type III SS, to assess region, age, region*age. As can see, the interaction term is not statistically significant (p-value=0.07 > 0.05). So the assumption of slope equality made previously, was justified. Thus, the interaction term is not statistically significant and we do not need this term in our model and we can got back to last model without this term.

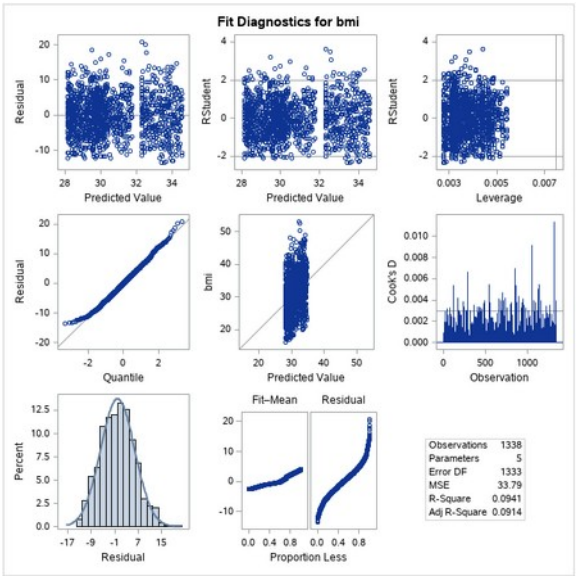


Figure 12

This is a plots of residuals for model of bmi and region, and age as covariate. Residual versus predicted value plot imply somehow the equal variance of residual. Also, small number of residual points are outside the bounds(-2 and 2) in plot of Studentised residuals. There is no evidence of high leverage. There are some value above rule of thumb in Cook D which are not extreme. Q-Q plot indicate small deviation from straight line on the both ends and histogram indicate the approximately normal distribution for residuals.

Create a new variable log_charges
log_charges=log(charges);
A one-way analysis of variance relating log_charges and children. Carry out appropriate post-hoc comparison.

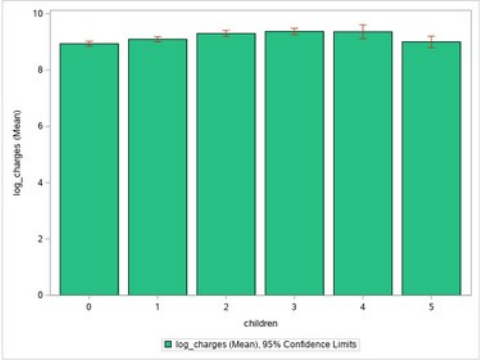


Figure 13

The bar chart shows the log-charges for number of children covered by health insurance. We can say that health insurance for three children is more expensive than no children case. We discuss whether this difference is statistically different.

Variable: log_charges
children = 0

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.953461	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.100662	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.554273	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	9.085941	Pr > A-Sq	<0.0050

Table 17

Variable: log_charges
children = 1

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.976792	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.069641	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.286683	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.917697	Pr > A-Sq	<0.0050

Table 16

Variable: log_charges
children = 2

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.952788	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.08583	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.531263	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.45466	Pr > A-Sq	<0.0050

Table 19

Variable: log_charges
children = 3

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.961079	Pr < W	0.0002
Kolmogorov-Smirnov	D	0.075928	Pr > D	0.0249
Cramer-von Mises	W-Sq	0.233389	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.628705	Pr > A-Sq	<0.0050

Table 18

Variable: log_charges
children = 4

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.967096	Pr < W	0.5727
Kolmogorov-Smirnov	D	0.107821	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.038841	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.253115	Pr > A-Sq	>0.2500

Table 20

Variable: log_charges
children = 5

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.953527	Pr < W	0.4830
Kolmogorov-Smirnov	D	0.11318	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.039301	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.28156	Pr > A-Sq	>0.2500

Table 21

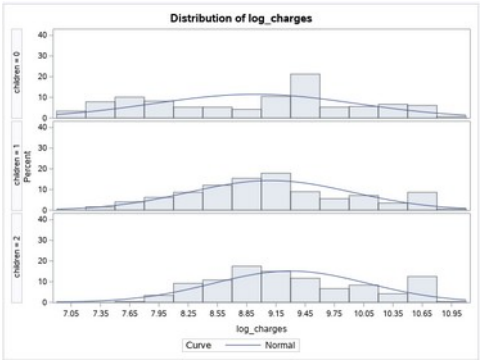


Figure 15

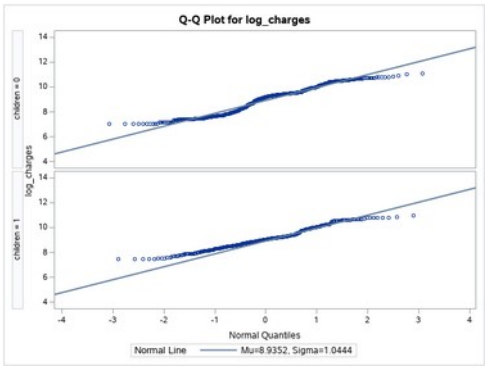


Figure 14

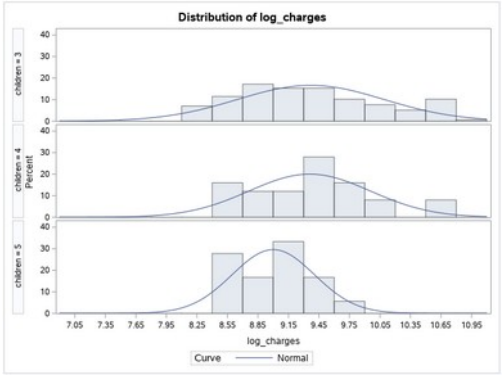


Figure 16

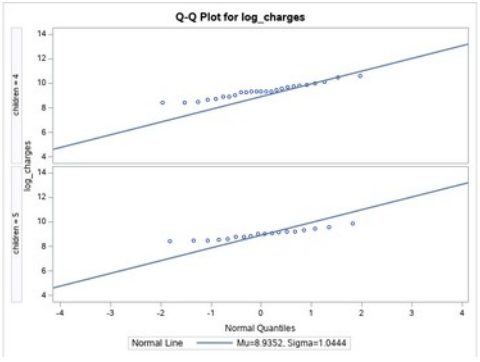


Figure 17

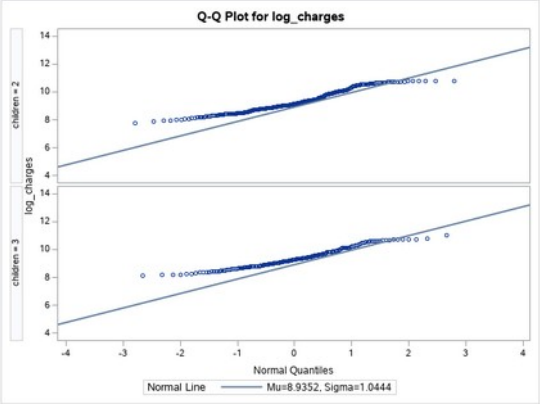


Figure 18

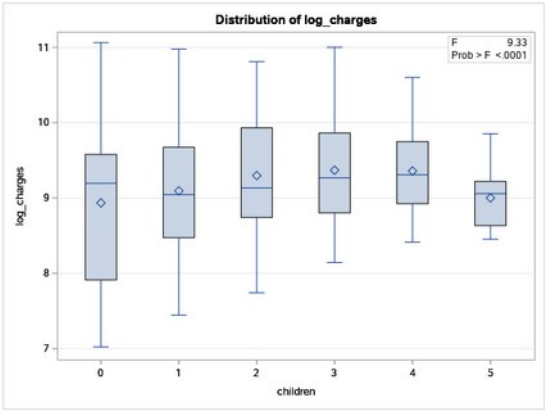


Figure 19

The P-value in Test of normality for 0, 1, 2 and 3 children is less than 0.05 which imply that the assumption of normality is rejected. The distribution plot and Q-Q plot also confirm that the distribution is not symmetric. According to Normality Test for 4 and 5 children, p-value is greater than significance level of 0.05 and imply the normal distribution. The distribution plot and Q-Q plot(there is a deviation from guide line at both ends) do not confirm it easily.
Boxplots suggest a variability in variance, but we should do homogeneity of variance test to measure.

For ANOVA interpretation, first we should check the equal variance assumption through a test of homogeneity.

Levene's Test for Homogeneity of log_charges Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
children	5	84.0153	16.8031	22.30	<.0001
Error	1332	1003.7	0.7535		

Welch's ANOVA for log_charges			
Source	DF	F Value	Pr > F
children	5.0000	10.55	<.0001
Error	116.2		

Table 22

The test of homogeneity of variance imply that P-value is less than 0.05, so we can reject the equal variance assumption. There is significant differences in variance of log-charge across all family with different number of children. So, For our data we need to report F-ratio and P-value of Welch's ANOVA instead of main ones. The p-value for Welch test in table 22 is smaller than 0.05 which imply the null hypothesis of equal means in all samples is rejected and hence there is a significant difference in the mean of log-charges in all families. P-value <0.0001.

As two assumption of ANOVA (normality and equal variance) is rejected, we should interpret with precaution.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	38.241013	7.648203	9.33	<.0001
Error	1332	1092.232746	0.819995		
Corrected Total	1337	1130.473759			

R-Square	Coeff Var	Root MSE	log_charges Mean
0.033827	9.952407	0.905536	9.098659

Source	DF	Type I SS	Mean Square	F Value	Pr > F
children	5	38.24101297	7.64820259	9.33	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
children	5	38.24101297	7.64820259	9.33	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	9.001097067	B	0.21343677	42.17	<.0001
children 0	-0.065923560	B	0.21675750	-0.30	0.7611
children 1	0.093713825	B	0.21928543	0.43	0.6692
children 2	0.297163455	B	0.22129595	1.34	0.1796
children 3	0.367816094	B	0.22534007	1.63	0.1029
children 4	0.357919718	B	0.27991969	1.28	0.2012
children 5	0.000000000	B	.	.	.

Table 23

The parameter estimate in table 23 indicates that all parameter except for intercept are not statistically significant(P-values > 0.05). The intercept (9) only corresponds to the line of family with 5 children and imply the sample mean. Other estimates correspond to 0,1,2,3 and 4 explain the correction to children 5.

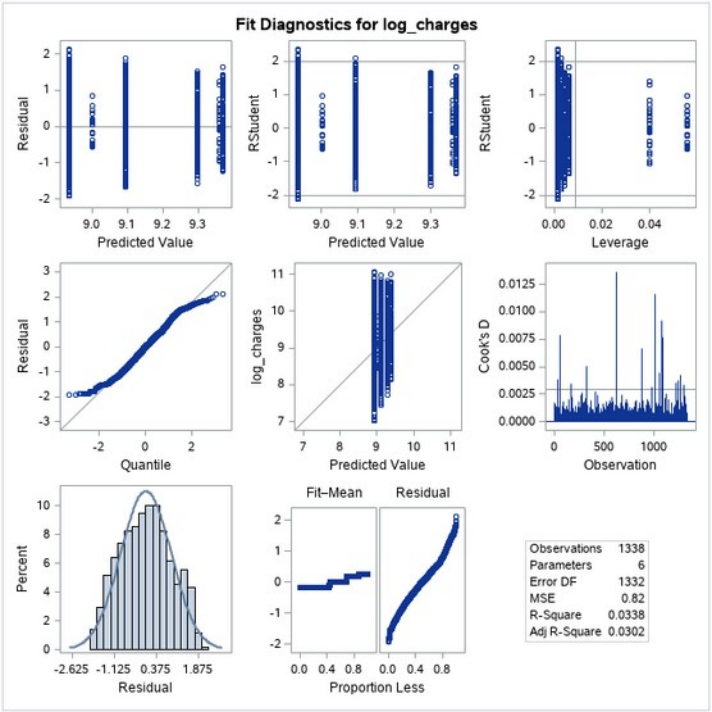


Figure 20

Figure 20 shows fit diagnostics, and interpretation of residual versus predicted value plots implies the unequal variability of residual (heteroscedasticity) in 6 different lines which increases from right to left which is align with the assumption of equal variances that has been violated . Furthermore, the plot of Studentised residuals indicates some points outside the -2 and 2 bounds. The leverage plot indicate some points(two group) with high leverage(greater than cause of concern). Cook's D shows some value above rule of thumb which is not extreme. Q-Q plot and histogram imply the not symmetric distribution. Q-Q implies the deviation at both ends to the guide line.

Comparisons significant at the 0.05 level are indicated by ***.				
children Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3 - 4	0.00990	-0.54659	0.56638	
3 - 2	0.07065	-0.19461	0.33592	
3 - 1	0.27410	0.02280	0.52540	***
3 - 5	0.36782	-0.27527	1.01091	
3 - 0	0.43374	0.20099	0.66649	***
4 - 3	-0.00990	-0.56638	0.54659	
4 - 2	0.06076	-0.48235	0.60386	
4 - 1	0.26421	-0.27222	0.80063	
4 - 5	0.35792	-0.44093	1.15677	
4 - 0	0.42384	-0.10415	0.95183	
2 - 3	-0.07065	-0.33592	0.19461	
2 - 4	-0.06076	-0.60386	0.48235	
2 - 1	0.20345	-0.01664	0.42354	
2 - 5	0.29716	-0.33439	0.92871	
2 - 0	0.36309	0.16444	0.56174	***
1 - 3	-0.27410	-0.52540	-0.02280	***
1 - 4	-0.26421	-0.80063	0.27222	
1 - 2	-0.20345	-0.42354	0.01664	
1 - 5	0.09371	-0.53210	0.71953	
1 - 0	0.15964	-0.01994	0.33921	
5 - 3	-0.36782	-1.01091	0.27527	
5 - 4	-0.35792	-1.15677	0.44093	
5 - 2	-0.29716	-0.92871	0.33439	
5 - 1	-0.09371	-0.71953	0.53210	
5 - 0	0.06592	-0.55267	0.68452	
0 - 3	-0.43374	-0.66649	-0.20099	***
0 - 4	-0.42384	-0.95183	0.10415	
0 - 2	-0.36309	-0.56174	-0.16444	***
0 - 1	-0.15964	-0.33921	0.01994	
0 - 5	-0.06592	-0.68452	0.55267	

Table 25

children	log_charges LSMEAN	LSMEAN Number
0	8.93517351	1
1	9.09481089	2
2	9.29826052	3
3	9.36891316	4
4	9.35901678	5
5	9.00109707	6

Least Squares Means for effect children
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: log_charges

i/j	1	2	3	4	5	6
1		0.1142	<.0001	<.0001	0.1984	0.9997
2	0.1142		0.0890	0.0232	0.7237	0.9982
3	<.0001	0.0890		0.9740	0.9996	0.7610
4	<.0001	0.0232	0.9740		1.0000	0.5772
5	0.1984	0.7237	0.9996	1.0000		0.7968
6	0.9997	0.9982	0.7610	0.5772	0.7968	

Table 24

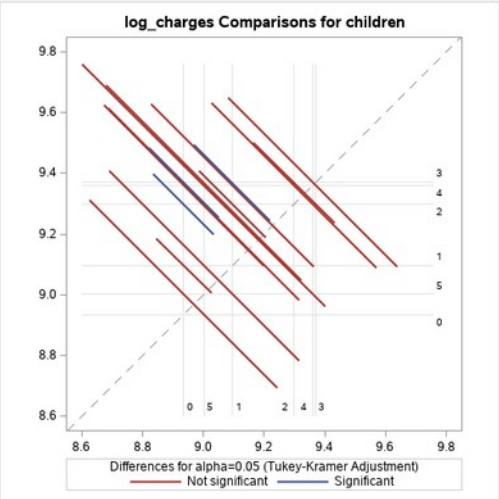


Figure 21

Post-hoc test indicate differences between regions separately. Table 25 shows the comparison significance at level 0.05 and indicate that only three comparisons are significantly different. This is confirmed by table 24 and diffogram shown in Figure 21. In diffogram only the blue line located on intersection of 1,3 and 0,2 and 0,3 do not cross the zero line and is significant. Other lines cross zero lines which means are not significant.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
0 children vs 1 children	-0.36308702	0.06960751	-5.22	<.0001
1 children vs 3 children	-0.27410227	0.08805534	-3.11	0.0019
1 children vs 2 children	-0.20344963	0.07711998	-2.64	0.0084

Table 26

For contrast the mean for 0 children versus 1 children we put 1 0 -1 0 0 0 in code. It indicates the charge average for family with 0 children differ significantly from charge in family with 1 child(p-value<0.005). It also shows for 1 children versus 3 children, the weights for this comparison would be 0 1 0 -1 0 0 and the charge mean is statistically different. For contrast between 1 children and 2 children, by considering weights 0 1 -1 0 0 0 in code, the difference of means between these two family is statistically significant(P-value=<0.005) which is different from least square mean table.

Discuss for evidence of interaction between children and smoker.

Analysis Variable : log_charges							
children	smoker	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	no	459	459	8.5984376	0.8709420	7.0227557	10.3175818
	yes	115	115	10.2791888	0.3951126	9.5286394	11.0630449
1	no	263	263	8.8162351	0.6418021	7.4448489	10.4676682
	yes	61	61	10.2958836	0.3886669	9.6394637	10.9779962
2	no	185	185	8.9817796	0.5744363	7.7424030	10.5162543
	yes	55	55	10.3627873	0.3877936	9.4594990	10.8112957
3	no	118	118	9.0502228	0.4847763	8.1441171	10.3110698
	yes	39	39	10.3331557	0.3648603	9.7062855	11.0024564
4	no	22	22	9.2547639	0.5446257	8.4128682	10.5072646
	yes	3	3	10.1235378	0.4232882	9.7949055	10.6011805
5	no	17	17	8.9509606	0.3545486	8.4527180	9.5804083
	yes	1	1	9.8534177	.	9.8534177	9.8534177

Table 27

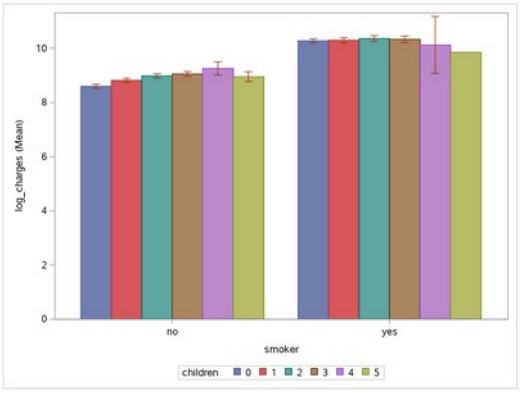


Figure 22

By looking at table and bar chart, it is obvious that there is differences in charges by smoking status across number of children. We discuss later, whether this difference is significant

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	538.291609	48.935601	109.58	<.0001
Error	1326	592.182150	0.446593		
Corrected Total	1337	1130.473759			

R-Square	Coeff Var	Root MSE	log_charges Mean
0.476165	7.344775	0.668276	9.098659

Source	DF	Type III SS	Mean Square	F Value	Pr > F
children	5	9.72899703	1.94579941	4.36	0.0006
smoker	1	37.79697258	37.79697258	84.63	<.0001
children*smoker	5	6.37746061	1.27549212	2.86	0.0143

Table 28

Table 28 shows, the model is statistically significant. The R-squared is weak so there is considerable variability in log-charge not explained by children and smoker. There is highly significant main effect due to children (p-value<0.05), and smoker status (p-value<0.05). There is also a significant interaction between children and smoker. (children*smoker p-value<0.05). Noteworthy, the increase of R2 from 3% to 47% with adding smoker and interaction of children*smoker shows the effect has been diminished.

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	9.853417720	B	0.66827605	14.74	<.0001
children 0	0.425771110	B	0.67117530	0.63	0.5260
children 1	0.442465907	B	0.67373145	0.66	0.5115
children 2	0.509369603	B	0.67432392	0.76	0.4502
children 3	0.479737948	B	0.67678946	0.71	0.4785
children 4	0.270120078	B	0.77165871	0.35	0.7264
children 5	0.000000000	B	.	.	.
smoker no	-0.902457162	B	0.68765038	-1.31	0.1896
smoker yes	0.000000000	B	.	.	.
children*smoker 0 no	-0.778294027	B	0.69117250	-1.13	0.2603
children*smoker 0 yes	0.000000000	B	.	.	.
children*smoker 1 no	-0.577191377	B	0.69417743	-0.83	0.4059
children*smoker 1 yes	0.000000000	B	.	.	.
children*smoker 2 no	-0.478550580	B	0.69526752	-0.69	0.4914
children*smoker 2 yes	0.000000000	B	.	.	.
children*smoker 3 no	-0.380475666	B	0.69864070	-0.54	0.5861
children*smoker 3 yes	0.000000000	B	.	.	.
children*smoker 4 no	0.033683284	B	0.80126588	0.04	0.9665
children*smoker 4 yes	0.000000000	B	.	.	.
children*smoker 5 no	0.000000000	B	.	.	.
children*smoker 5 yes	0.000000000	B	.	.	.

Table 29

The parameter estimate shows that intercept which relate to the mean log-charge for smoker families with 5 children is 9.85 and is statistically significant. The difference of mean log-charge between smoker families with 0,1,2,3,4 children with non smoker families is not significant. Furthermore, differences of mean log-charge between smokers and non smokers with 0,1,2,3,4 children relative to the difference between them with 5 children are not significant (p-value>0.05).

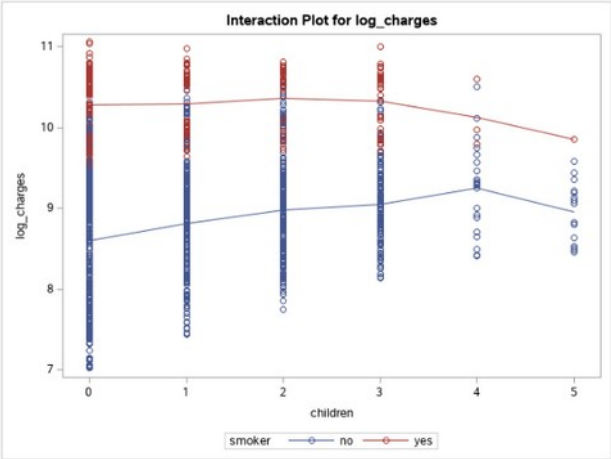


Figure 23

Interaction plot confirms that there is a significant interaction effect . Two lines are not parallel and there is a significant difference in smoker and non smoker in children 0 and 5. Furthermore, it confirms that there are significant main effect for smoker and non smoker. The mean log-charge for smokers are much higher than non smokers across all children numbers.

children	log_charges LSMEAN	LSMEAN Number
0	9.43881324	1
1	9.55605936	2
2	9.67228345	3
3	9.69168925	4
4	9.68915086	5
5	9.40218914	6

Least Squares Means for effect children
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: log_charges

i/j	1	2	3	4	5	6
1		0.3481	0.0024	0.0050	0.8369	1.0000
2	0.3481		0.5571	0.5043	0.9887	0.9978
3	0.0024	0.5571		0.9999	1.0000	0.9714
4	0.0050	0.5043	0.9999		1.0000	0.9622
5	0.8369	0.9887	1.0000	1.0000		0.9800
6	1.0000	0.9978	0.9714	0.9622	0.9800	

Table 30

smoker	log_charges LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
no	8.9420666	<.0001
yes	10.2079952	

Table 31

Post-hoc test indicates that there is a statistically significant differences in mean of log-charge in smokers and non smokers. Also the only significant difference is between 1, 3 children and 1,4 children.

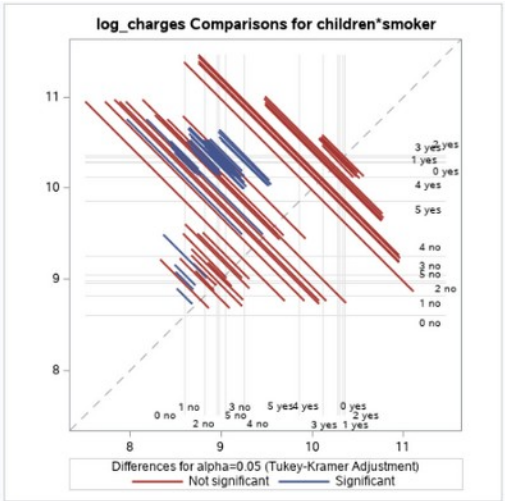


Figure 24

children	smoker	log_charges LSMEAN	LSMEAN Number
0	no	8.5984376	1
0	yes	10.2791888	2
1	no	8.8162351	3
1	yes	10.2958836	4
2	no	8.9817796	5
2	yes	10.3627873	6
3	no	9.0502228	7
3	yes	10.3331557	8
4	no	9.2547639	9
4	yes	10.1235378	10
5	no	8.9509606	11
5	yes	9.8534177	12

Least Squares Means for effect children*smoker
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: log_charges

i/j	1	2	3	4	5	6	7	8	9	10	11	12
1		<.0001	0.0016	<.0001	<.0001	<.0001	<.0001	<.0001	0.0005	0.0049	0.5969	0.7742
2	<.0001		<.0001	1.0000	<.0001	0.9998	<.0001	1.0000	<.0001	1.0000	<.0001	1.0000
3	0.0016	<.0001		<.0001	0.2912	<.0001	0.0702	<.0001	0.1228	0.0371	0.9997	0.9264
4	<.0001	1.0000	<.0001		<.0001	1.0000	<.0001	1.0000	<.0001	1.0000	<.0001	1.0000
5	<.0001	<.0001	0.2912	<.0001		<.0001	0.9994	<.0001	0.8119	0.1297	1.0000	0.9791
6	<.0001	0.9998	<.0001	1.0000	<.0001		<.0001	1.0000	<.0001	1.0000	<.0001	0.9998
7	<.0001	<.0001	0.0702	<.0001	0.9994	<.0001		<.0001	0.9768	0.2046	1.0000	0.9892
8	<.0001	1.0000	<.0001	1.0000	<.0001	1.0000	<.0001		<.0001	1.0000	<.0001	0.9999
9	0.0005	<.0001	0.1228	<.0001	0.8119	<.0001	0.9768	<.0001		0.6140	0.9621	0.9993
10	0.0049	1.0000	0.0371	1.0000	0.1297	1.0000	0.2046	1.0000	0.6140		0.1802	1.0000
11	0.5969	<.0001	0.9997	<.0001	1.0000	<.0001	1.0000	<.0001	0.9621	0.1802		0.9776
12	0.7742	1.0000	0.9264	1.0000	0.9791	0.9998	0.9892	0.9999	0.9993	1.0000	0.9776	

Table 32

Some differences of mean log-charge between smokers and non smokers families with various number of children are statistically significant. This is shown in diffogram by blue line which do not cross zero line. For example, mean log-charge in non smokers with 0 children are statistically different from smoker and non smoker with 1,2,3,4 children.

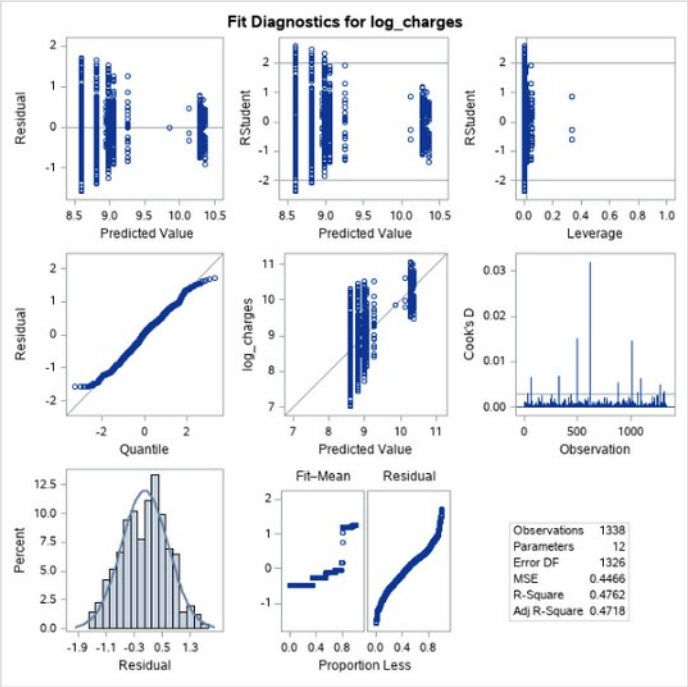


Figure 25

Fit diagnostics shows that residual versus predicted value plots implies the heteroscedasticity, variability decreases from left to right. Also, the plot of Studentised residuals indicates some points outside the -2 and 2 bounds. The leverage plot indicate bunch of points with high leverage. Cook's D shows some value above rule of thumb. Q-Q plot and histogram imply the not symmetric distribution. It implies the deviation at both ends of Q-Q plot line.

A summary of findings:

According to statistical tests, the variability of Bmi in different regions is not equal, and Bmi average in regions are significantly different. Bmi in southeast is significantly higher than other regions. Northeast and northwest do not show a significant difference with each other in term of Bmi.

Variability of insurance charge between families with different children is not equal. The average of insurance charge between families are significantly different. To be precise, here is a significantly difference in family with 1,3 and 0,2 and 0,3 in term of insurance charge. Furthermore, there is a significant interaction between children and smoker. This mean insurance charge in smoker families is significantly higher than non smoker families. There is a varied difference in mean of insurance charge

in smoker and non smoker families. The difference in non smoker and smoker families with 0 children is away more than this group with 5 children.