QMB 6305 Regression Project

About Data:

Source: https://github.com/stedy/Machine-Learning-with-R-datasets

Variables

1) Age – Age of the person – Continuous variable
2) BMI – BMI of the person which provides the understanding about weight of a person relative to height - Continuous
3) Smoker – Is the person a smoker – Nominal variable
4) Charges – Individual medical costs billed by health insurance - Continuous

The dependent variable (Y) is "Charges" billed by the health insurance company. Each record is for an individual. The independent variables are "Age", "BMI" of the person in that observation and "Smoker" which says whether that person is a smoker or a nonsmoker.

Levels in the binary variable ("Smoker") – 1- Smoker, 0 – Nonsmoker.

Number of Observations: 100

Ratio of Binary variable - Smoker – 50:50
The original data had ~1330 rows of which 100 rows with even split in the binary variable is extracted

DATA:

age	bmi	smoker	charges
18	33.77	0	1725.552
28	33	0	4449.462
33	22.705	0	21984.47
32	28.88	0	3866.855
31	25.74	0	3756.622
46	33.44	0	8240.59
37	27.74	0	7281.506
37	29.83	0	6406.411
60	25.84	0	28923.14
25	26.22	0	2721.321
23	34.4	0	1826.843
56	39.82	0	11090.72
19	24.6	0	1837.237
52	30.78	0	10797.34
23	23.845	0	2395.172
56	40.3	0	10602.39
60	36.005	0	13228.85
30	32.4	0	4149.736
18	34.1	0	1137.011
37	28.025	0	6203.902
59	27.72	0	14001.13
63	23.085	0	14451.84

55	32.775	0	12268.63
23	17.385	0	2775.192
18	26.315	0	2198.19
19	28.6	0	4687.797
63	28.31	0	13770.1
19	20.425	0	1625.434
62	32.965	0	15612.19
26	20.8	0	2302.3
24	26.6	0	3046.062
31	36.63	0	4949.759
41	21.78	0	6272.477
37	30.8	0	6313.759
38	37.05	0	6079.672
55	37.3	0	20630.28
18	38.665	0	3393.356
28	34.77	0	3556.922
60	24.53	0	12629.9
18	35.625	0	2211.131
21	33.63	0	3579.829
40	28.69	0	8059.679
58	31.825	0	13607.37
34	37.335	0	5989.524
43	27.36	0	8606.217
25	33.66	0	4504.662
64	24.7	0	30166.62
28	25.935	0	4133.642
19	28.9	0	1743.214
61	39.1	0	14235.07
19	27.9	1	16884.92
62	26.29	1	27808.73
27	42.13	1	39611.76
30	35.3	1	36837.47
34	31.92	1	37701.88
31	36.3	1	38711
22	35.6	1	35585.58
28	36.4	1	51194.56
35	36.67	1	39774.28
60	39.9	1	48173.36
36	35.2	1	38709.18
48	28	1	23568.27
36	34.43	1	37742.58

58	36.955	1	47496.49
18	31.68	1	34303.17
53	22.88	1	23244.79
20	22.42	1	14711.74
28	23.98	1	17663.14
27	24.75	1	16577.78
22	37.62	1	37165.16
37	34.8	1	39836.52
45	22.895	1	21098.55
57	31.16	1	43578.94
59	29.83	1	30184.94
64	31.3	1	47291.06
56	19.95	1	22412.65
38	19.3	1	15820.7
61	29.92	1	30942.19
20	28.025	1	17560.38
63	35.09	1	47055.53
29	27.94	1	19107.78
44	31.35	1	39556.49
19	28.3	1	17081.08
32	17.765	1	32734.19
34	25.3	1	18972.5
30	28.69	1	20745.99
46	30.495	1	40720.55
42	23.37	1	19964.75
48	24.42	1	21223.68
18	25.175	1	15518.18
30	35.53	1	36950.26
42	26.6	1	21348.71
18	36.85	1	36149.48
63	37.7	1	48824.45
36	41.895	1	43753.34
27	36.08	1	37133.9
35	27.74	1	20984.09
19	34.8	1	34779.62
42	24.64	1	19515.54
40	22.22	1	19444.27

REGRESSIONS

WITH ONE INDEPENDENT VARIABLE

1) Charges Vs Age

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
13999.2	10.10%	9.18%	6.43%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7813	3773	2.07	0.041	
age	309.2	93.2	3.32	0.001	1.00

Regression Equation

charges = 7813 + 309.2 age

Fits and Diagnostics for Unusual Observations

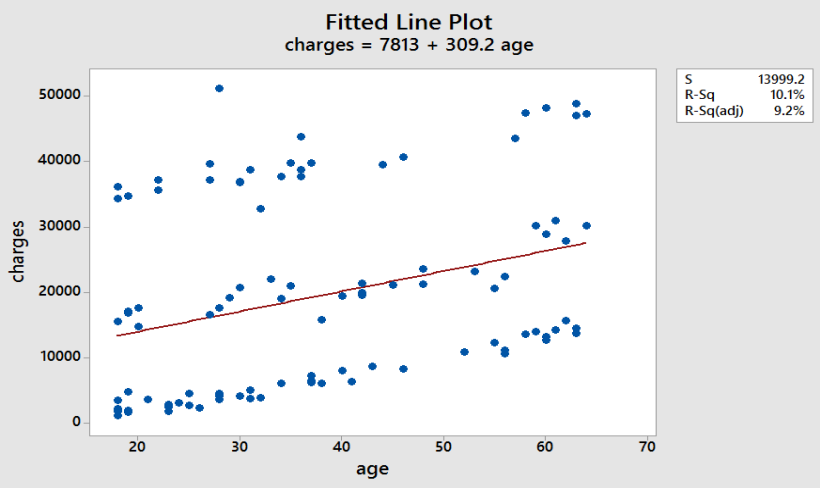
Obs	charges	Fit	Resid	Std Resid	
58	51195	16472	34722	2.50	R

R Large residual

The constant gives the charges when (“Age” =0) which does not occur.

P value of the age is 0.001 which is less than 0.05, so we can reject NULL hypothesis (“Age” does not have impact on charges (or) slope is 0 because of this variable)

Regression Equation:
Charges = 7813 + 309.2 * Age
Intercept = 7813
Slope = 309.2
For a unit increase in Age, the charges would increase by 309.2



R squared= 10.1 % indicates the amount of variance explained by the model

2) Charges vs BMI

Regression Equation:

Charges = -5432 + 832*BMI
Intercept = -5432
Slope (BMI) = 823
For a unit increase in BMI, the charges would increase by 823

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
13956.5	10.64%	9.73%	6.87%

Coefficients

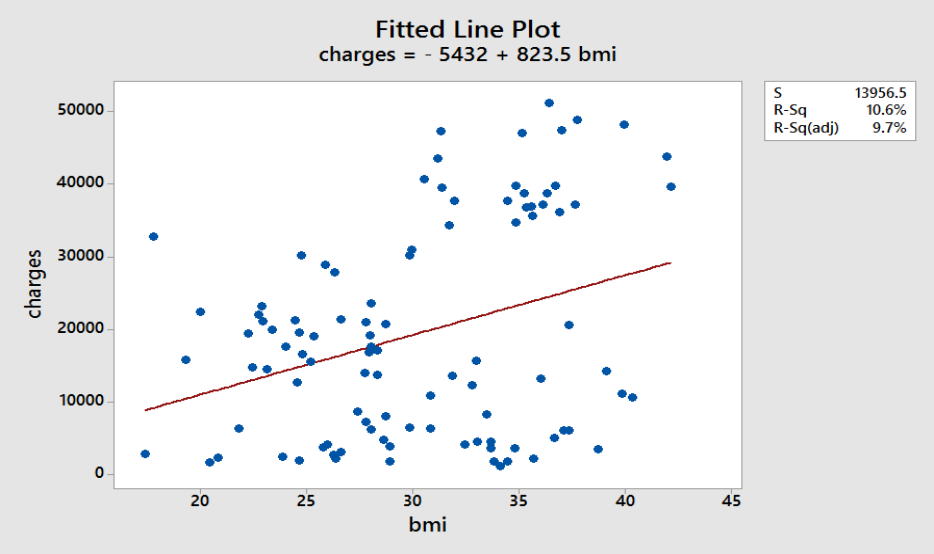
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-5432	7411	-0.73	0.465	
bmi	823	241	3.42	0.001	1.00

Regression Equation

charges = -5432 + 823 bmi

P value of the BMI is 0.001 which is less than 0.05 , so we can reject NULL hypothesis (“BMI” does not have impact on charges (or) slope is 0 because of this variable)

The constant gives the charges when (“BMI” =0) which does not occur.



R squared= 10.6 % indicates the amount of variance explained by the model

3) Charges vs Smoker

Regression Equation:

Charges = 8001 + 22875 * Smoker_1
Intercept = 8001
Slope (Smoker) =
For a smoker_1 which means if the person is a smoker, the charges would increase by 22875
For a non-smoker, the charges would be 22875 less

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
9192.77	61.23%	60.84%	59.63%

Coefficients

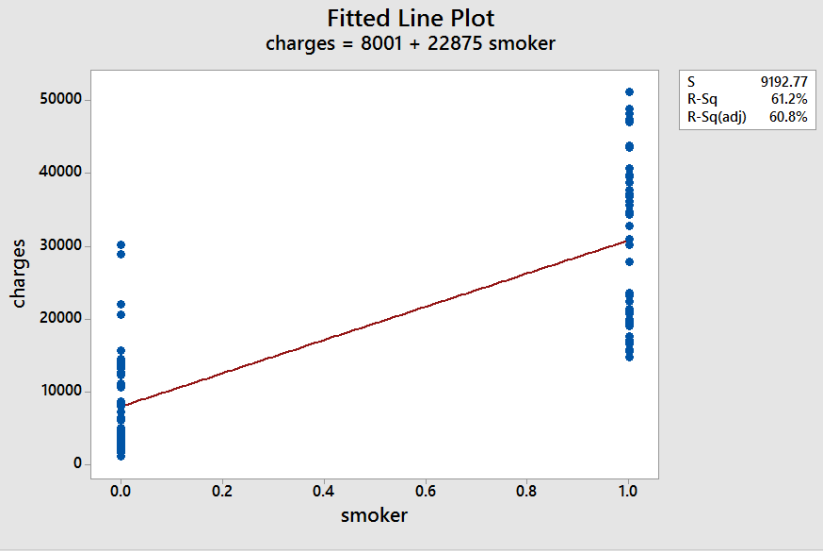
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	8001	1300	6.15	0.000	
smoker					
1	22875	1839	12.44	0.000	1.00

Regression Equation

charges = 8001 + 0.0 smoker_0 + 22875 smoker_1

P value of the “Smoker” is 0.000 which is less than 0.05 , so we can reject NULL hypothesis (“Smoker” does not have impact on charges (or) slope is 0 because of this variable)

The constant gives the charges when (“Smoker” =0).



R squared= 61.2 % indicates the amount of variance explained by the model

MULTIPLE REGRESSION

1) Charges Vs Age and BMI

Regression Equation = -15471 + 295.3 *Age + 788*BMI
Intercept = -15471
Slope (Age) = 295.3
Slope (BMI) = 788

When the model is built with Age and BMI as independent variables, the charges would increase by 295.3 for a unit increase Age and 788 for a unit increase in BMI

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
13287.6	19.83%	18.18%	14.54%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-15471	7672	-2.02	0.047	
age	295.3	88.6	3.33	0.001	1.00
bmi	788	230	3.43	0.001	1.00

Regression Equation

charges = -15471 + 295.3 age + 788 bmi

Adjusted R squared = 18.18 % - The amount of variance explained by this model
The constant gives the charges when (“Age” =0 and BMI = “0”) which does not occur.P

value (Age) = 0.001 which is less than 0.05 -> the null hypothesis that the slope is 0 can be rejected

P value (BMI) = 0.001 which is less than 0.05 -> the null hypothesis that the slope is 0 can be rejected

2) Charges Vs Age and Smoker

Regression Equation = -3250 + 300.7 *Age + 22772 *Smoker
Intercept = -3250

Slope (Age) = 300.7
Slope (Smoker) = 22772

When the model is built with Age and Smoker as independent variables, the charges would increase by 300.7 for a unit increase Age and 22772 if the person is a smoker

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-3250	2299	-1.41	0.161	
age	300.7	53.4	5.63	0.000	1.00
smoker					
1	22772	1605	14.19	0.000	1.00

Regression Equation

smoker			
0	charges	=	-3250 + 300.7 age
1	charges	=	19522 + 300.7 age

Fits and Diagnostics for Unusual Observations

Obs	charges	Fit	Resid	Std Resid	
58	51195	27941	23254	2.93	R

R Large residual

When the model is built with Age and BMI as independent variables, the charges would increase by 295.3 for a unit increase Age and 788 for a unit increase in BMI

For a smoker
Charges = 19522 + 300.7 * Age

For a non-smoker
Charges= -3250 + 300.7*Age

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
8022.57	70.78%	70.17%	68.99%

Adjusted R squared = 70.17 % - The amount of variance explained by this model
P value (Age) = 0.000 which is less than 0.05 -> the null hypothesis that the slope is 0 can be rejected
P value (BMI) = 0.000 which is less than 0.05 -> the null hypothesis that the slope is 0 can be rejected

3) Charges Vs BMI and Smoker

Regression Equation = -15685 + 787 * BMI + 22705 * Smoker
Intercept = -15685

Slope (BMI) = 787
Slope (Smoker) = 22705

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-15685	4308	-3.64	0.000	
bmi	787	138	5.70	0.000	1.00
smoker					
1	22705	1600	14.19	0.000	1.00

Regression Equation

smoker			
0	charges	=	-15685 + 787 bmi
1	charges	=	7020 + 787 bmi

Fits and Diagnostics for Unusual Observations

Obs	charges	Fit	Resid	Std Resid	
3	21984	2185	19800	2.52	R
9	28923	4652	24271	3.07	R
47	30167	3755	26412	3.35	R

R Large residual

When the model is built with BMI and Smoker as independent variables, the charges would increase by 787 for a unit increase Age and 22705 if the person is a smoker

For a smoker
Charges = -15685 + 787 * BMI

For a non-smoker
Charges= 7020 + 787 * BMI

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
7998.04	70.95%	70.36%	69.15%

Adjusted R squared = 70.36 % - The amount of variance explained by this model
P value (BMI) = 0.000 which is less than 0.05 -> the null hypothesis that the slope is 0 can be rejected
P value (Smoker) = 0.000 which is less than 0.05 -> the null hypothesis that the slope is 0 can be rejected

FULL MULTIPLE REGRESSION

Charges Vs Age, BMI and Smoker

Regression Equation = -25415 + 753 *BMI + 287.4 * Age + 22615 * Smoker_1
Intercept = -25415
Slope (Age) = 287.4
Slope (BMI) = 753
Slope (Smoker) = 22615

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-25415	3930	-6.47	0.000	
bmi	753	116	6.47	0.000	1.00
age	287.4	44.9	6.41	0.000	1.00
smoker					
1	22615	1346	16.80	0.000	1.00

Regression Equation

smoker			
0	charges	=	-25415 + 753 bmi + 287.4 age
1	charges	=	-2800 + 753 bmi + 287.4 age

When the model is built with Age, BMI and Smoker as independent variables, the charges would increase by 753 for unit increase in BMI, by 287.4 for unit increase in Age, by 22615 if the person is a smoker

For a smoker
Charges = -25415+753* BMI+287.4 * Age

For a non-smoker
Charges= -2800+753*BMi+287.4 *Age

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
6728.48	79.66%	79.02%	77.73%

Adjusted R squared = 79.02 % - The amount of variance explained by this model
P value (BMI) = 0.000 which is less than 0.05 -> the null hypothesis that the slope is 0 can be rejected
P value (Age) = 0.000 which is less than 0.05 -> the null hypothesis that the slope is 0 can be rejected
P value (Smoker) = 0.000 which is less than 0.05 -> the null hypothesis that the slope is 0 can be rejected

REGRESSION WITH INTERACTION

Charges Vs Age, BMI and Age*BMI

Regression Equation = -34563 + 1413 * BMI + 806 * Age – 16.6 * Age*BMI
Intercept = -34563
Slope (Age) = 806
Slope (BMI) = 1413
Slope (Age * BMI) = -16.6

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-34563	19590	-1.76	0.081	
bmi	1413	633	2.23	0.028	7.61
age	806	490	1.64	0.103	30.76
Age*BMI	-16.6	15.7	-1.06	0.292	38.65

Regression Equation

charges = -34563 + 1413 bmi + 806 age - 16.6 Age*BMI

Fits and Diagnostics for Unusual Observations

Obs	charges	Fit	Resid	Std Resid	
58	51195	22461	28733	2.20	R
84	32734	6865	25869	2.01	R

R Large residual

P value (BMI) = 0.028 which is less than 0.05 -> the null hypothesis that the slope is 0 can be rejected
P value (Age) = 0.103 which is more than 0.05 -> the null hypothesis that the slope is 0 cannot be rejected
P value (Age * BMI) = 0.292 which is more than 0.05 -> the null hypothesis that the slope is 0 cannot be rejected

When the model is built with Age, BMI and Age*BMI as independent variables, the charges would increase by 1413 for unit increase in BMI. The other variables are not significant.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
13279.2	20.76%	18.28%	13.59%

Adjusted R squared = 18.28 % - The amount of variance explained by this model

REGRESSIONS USING QUADRATIC TERMS

Charges Vs Age and Age^2

Regression Equation = -4690 + 489 * Age – 2.23 *(Age^2)
Intercept = 4690

Slope (Age) = 489
Slope (Age * Age) = -2.23

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4690	11274	0.42	0.678	
age	489	619	0.79	0.431	43.75
Age*Age	-2.23	7.57	-0.29	0.769	43.75

Regression Equation

charges = 4690 + 489 age - 2.23 Age*Age

P value (Age) = 0.431 which is more than 0.05 -> the null hypothesis that the slope is 0 cannot be rejected
P value (Age^2) = 0.769 which is more than 0.05 -> the null hypothesis that the slope is 0 cannot be rejected
P value of constant is also higher than 0.05, we cannot reject the null hypothesis

When the model is built with Age and (Age^2) as the independent variables, there is no significant variable.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
14064.9	10.18%	8.32%	4.59%

Adjusted R squared = 8.32 % - The amount of variance explained by this model

Charges Vs BMI and BMI^2

Regression Equation = 36117 -2051 * BMI + 47.9 *(BMI^2)
Intercept = 36117

Slope (BMI) = -2051
Slope (BMI * BMI) = 47.9

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	36117	34187	1.06	0.293	
bmi	-2051	2322	-0.88	0.379	93.35
BMI*BMI	47.9	38.5	1.24	0.216	93.35

Regression Equation

charges = 36117 - 2051 bmi + 47.9 BMI*BMI

P value (BMI) = 0.379 which is more than 0.05 -> the null hypothesis that the slope is 0 cannot be rejected
P value (BMI^2) = 0.7216 which is more than 0.05 -> the null hypothesis that the slope is 0 cannot be rejected
P value of constant is also higher than 0.05, so we cannot reject the null hypothesis

When the model is built with BMI and (BMI^2) as the independent variables, there is no significant variable in the model

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
13917.6	12.05%	10.24%	6.36%

Adjusted R squared = 10.24 % - The amount of variance explained by this model

FINDING THE BEST FIT MODEL

The best fit model is the one which as the highest R^2 value

Model	**Adjusted R^2**	**F Statistic**	Residual Standard Error	**RMSE**

Full Effect Model (Age+ BMI+ Smoker)	79.02	125.3	6728	6592.542
Age+ Smoker	70.17	117.5	8023	7901.3
BMI+ Smoker	70.36	118.5	7998	7877.16

From the above models, the best model is identified as the model with all independent variables. Full Main effect Multi Regression is the best model with an Adjusted R squared value of 79.02 and high F statistic and Residual Standard Error and RMSE values are the lowest for this model

```
Call:
lm(formula = data$charges ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-10423.2  -6019.4   -178.3   4657.7  20818.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -25415.03    3929.56  -6.468 4.16e-09 ***
i..age         287.40      44.85   6.408 5.49e-09 ***
bmi            753.01     116.33   6.473 4.06e-09 ***
smoker       22614.97    1346.00  16.802  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6728 on 96 degrees of freedom
Multiple R-squared:  0.7966,    Adjusted R-squared:  0.7902
F-statistic: 125.3 on 3 and 96 DF,  p-value: < 2.2e-16
```

The p value for all variables is less than 0.05 indicating that the null hypothesis (slope is 0 i.e..., there is no impact by the independent variable) can be rejected.

The F statistic is 125.3 which is and indicator of the amount of variance explained by the model in comparison to the variance due to errors and randomness. The F statistic for this model is also high.

Regression Equation

Charges = -25415 + 753 *BMI + 287.4 * Age + 22615 * Smoker_1

For a smoker
Charges = -25415+753* BMI+287.4 * Age

For a non-smoker
Charges= -2800+753*BMi+287.4 *Age

Intercept = -25415
Slope (Age) = 287.4
Slope (BMI) = 753
Slope (Smoker) = 22615

When the model is built with Age, BMI and Smoker as independent variables, the charges would increase by 753 for unit increase in BMI, by 287.4 for unit increase in Age, by 22615 if the person is a smoker

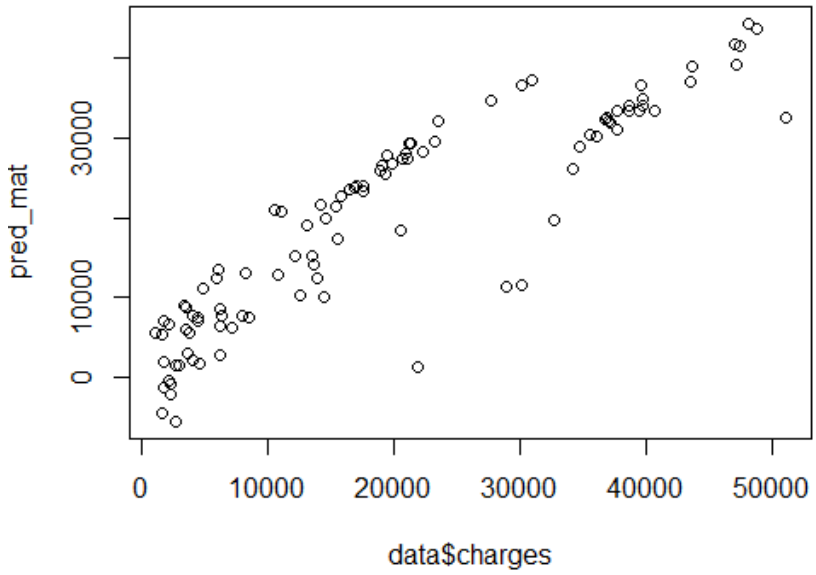
PREDICTIONS USING THE MODEL

```
> predictions =predict(reg)
> predictions
        1         2         3         4         5         6         7         8
 5187.2389  7481.4109  1166.1851  5528.6122  2876.7675 12985.9135  6107.1770  7680.9642
        9        10        11        12        13        14        15        16
11286.6345  1513.8184  7098.6283 20664.0945 -1430.4480 12707.3046  -849.3739 21025.5385
       17        18        19        20        21        22        23        24
18940.9634  7604.4036  5435.7316  6321.7843 12414.8912 10074.2932 15071.7525 -5713.8072
       25        26        27        28        29        30        31        32
 -426.4376  1581.5851 14008.7614 -4574.2574 17226.6160 -2280.0875  1512.5627 11077.0273
       33        34        35        36        37        38        39        40
 2768.8431  8411.3822 13405.0826 18479.1149  8873.2143  8814.2355 10300.1937  6584.0692
       41        42        43        44        45        46        47        48
 5944.0142  7684.7313 15218.5912 12470.0947  7545.4268  7116.1998 11577.8004  2161.4076
       49        50        51        52        53        54        55        56
 1807.4875 21558.9228 23669.4479 34815.2545 36683.9460 32403.0962 31007.5236 33443.5032
       57        58        59        60        61        62        63        64
30329.8080 32656.6076 34871.7116 44488.8991 34052.1883 32079.3149 33472.3720 41696.4922
       65        66        67        68        69        70        71        72
26228.4203 29660.9068 19830.3615 23304.2450 23596.6626 31850.8846 34038.3839 27373.0113
       73        74        75        76        77        78        79        80
37045.4105 36618.7072 39162.6235 28316.7891 22654.1548 37261.2756 24050.9728 41729.1259
       81        82        83        84        85        86        87        88
26573.5566 33452.2972 23970.6512 19773.8941 26022.6089 27425.7116 33383.2728 26868.4937
       89        90        91        92        93        94        95        96
29383.5454 21330.1016 32576.2880 29300.7104 30121.4729 43694.4775 39093.5786 32128.2461
       97        98        99       100
28147.3479 28865.2048 27824.8142 25427.7365
> |
```

The above predictions are the predictions for “Charges” made by the model.

S. No	Actual Charges	Predicted Charges						
1	1725.552	5187.239						
2	4449.462	7481.411						
3	21984.47	1166.185						
4	3866.855	5528.612						
5	3756.622	2876.767						
6	8240.59	12985.91						
7	7281.506	6107.177						
8	6406.411	7680.964						
9	28923.14	11286.63						
10	2721.321	1513.818						
11	1826.843	7098.628						
12	11090.72	20664.09	53	39611.76	36683.95	94	48824.45	43694.48
13	1837.237	-1430.45	54	36837.47	32403.1	95	43753.34	39093.58
14	10797.34	12707.3	55	37701.88	31007.52	96	37133.9	32128.25
15	2395.172	-849.374	56	38711	33443.5	97	20984.09	28147.35
16	10602.39	21025.54	57	35585.58	30329.81	98	34779.62	28865.2

17	13228.85	18940.96	58	51194.56	32656.61	99	19515.54	27824.81
18	4149.736	7604.404	59	39774.28	34871.71	100	19444.27	25427.74
19	1137.011	5435.732	60	48173.36	44488.9			
20	6203.902	6321.784	61	38709.18	34052.19			
21	14001.13	12414.89	62	23568.27	32079.31			
22	14451.84	10074.29	63	37742.58	33472.37			
23	12268.63	15071.75	64	47496.49	41696.49			
24	2775.192	-5713.81	65	34303.17	26228.42			
25	2198.19	-426.438	66	23244.79	29660.91			
26	4687.797	1581.585	67	14711.74	19830.36			
27	13770.1	14008.76	68	17663.14	23304.25			
28	1625.434	-4574.26	69	16577.78	23596.66			
29	15612.19	17226.62	70	37165.16	31850.88			
30	2302.3	-2280.09	71	39836.52	34038.38			
31	3046.062	1512.563	72	21098.55	27373.01			
32	4949.759	11077.03	73	43578.94	37045.41			
33	6272.477	2768.843	74	30184.94	36618.71			
34	6313.759	8411.382	75	47291.06	39162.62			
35	6079.672	13405.08	76	22412.65	28316.79			
36	20630.28	18479.11	77	15820.7	22654.15			
37	3393.356	8873.214	78	30942.19	37261.28			
38	3556.922	8814.235	79	17560.38	24050.97			
39	12629.9	10300.19	80	47055.53	41729.13			
40	2211.131	6584.069	81	19107.78	26573.56			
41	3579.829	5944.014	82	39556.49	33452.3			
42	8059.679	7684.731	83	17081.08	23970.65			
43	13607.37	15218.59	84	32734.19	19773.89			
44	5989.524	12470.09	85	18972.5	26022.61			
45	8606.217	7545.427	86	20745.99	27425.71			
46	4504.662	7116.2	87	40720.55	33383.27			
47	30166.62	11577.8	88	19964.75	26868.49			
48	4133.642	2161.408	89	21223.68	29383.55			
49	1743.214	1807.488	90	15518.18	21330.1			
50	14235.07	21558.92	91	36950.26	32576.29			
51	16884.92	23669.45	92	21348.71	29300.71			
52	27808.73	34815.25	93	36149.48	30121.47			



The above plot is the plot of actual values (x axis) and the predictions (y axis) made by the model

Fits and Diagnostics for All Observations

Obs	charges	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI
1	1726	5187	1372	(2465, 7910)	-3462	-0.53	-0.52	0.0415493
2	4449	7481	1101	(5297, 9666)	-3032	-0.46	-0.45	0.0267555
3	21984	1166	1292	(-1397, 3730)	20818	3.15	3.31	0.0368447
4	3867	5529	991	(3562, 7495)	-1662	-0.25	-0.25	0.0216767
5	3757	2877	1110	(674, 5080)	880	0.13	0.13	0.0272040
6	8241	12986	1091	(10819, 15153)	-4745	-0.71	-0.71	0.0263153
7	7282	6107	990	(4142, 8073)	1174	0.18	0.18	0.0216540
8	6406	7681	952	(5791, 9571)	-1275	-0.19	-0.19	0.0200275
9	28923	11287	1491	(8328, 14245)	17637	2.69	2.78	0.0490773
10	2721	1514	1181	(-831, 3859)	1208	0.18	0.18	0.0308342
11	1827	7099	1267	(4584, 9613)	-5272	-0.80	-0.80	0.0354360
12	11091	20664	1671	(17346, 23982)	-9573	-1.47	-1.48	0.0617125
13	1837	-1430	1396	(-4201, 1340)	3268	0.50	0.49	0.0430341
14	10797	12707	1155	(10414, 15001)	-1910	-0.29	-0.29	0.0294814
15	2395	-849	1345	(-3519, 1821)	3245	0.49	0.49	0.0399640
16	10602	21026	1709	(17634, 24417)	-10423	-1.60	-1.61	0.0644784
17	13229	18941	1530	(15904, 21978)	-5712	-0.87	-0.87	0.0516942
18	4150	7604	1047	(5526, 9683)	-3455	-0.52	-0.52	0.0242161
19	1137	5436	1385	(2686, 8185)	-4299	-0.65	-0.65	0.0423746
20	6204	6322	981	(4374, 8270)	-118	-0.02	-0.02	0.0212784
21	14001	12415	1394	(9648, 15182)	1586	0.24	0.24	0.0429181
22	14452	10074	1724	(6652, 13496)	4378	0.67	0.67	0.0656495
23	12269	15072	1266	(12559, 17584)	-2803	-0.42	-0.42	0.0353853
24	2775	-5714	1850	(-9386, -2042)	8489	1.31	1.32	0.0755913
25	2198	-426	1350	(-3106, 2253)	2625	0.40	0.40	0.0402546
26	4688	1582	1267	(-933, 4096)	3106	0.47	0.47	0.0354545
27	13770	14009	1512	(11007, 17010)	-239	-0.04	-0.04	0.0505083
28	1625	-4574	1664	(-7877, -1272)	6200	0.95	0.95	0.0611475
29	15612	17227	1483	(14283, 20170)	-1614	-0.25	-0.24	0.0485678

30	2302	-2280	1512	(-5281, 721)	4582	0.70	0.70	0.0504980
31	3046	1513	1188	(-845, 3870)	1533	0.23	0.23	0.0311584
32	4950	11077	1260	(8577, 13577)	-6127	-0.93	-0.93	0.0350426
33	6272	2769	1371	(47, 5491)	3504	0.53	0.53	0.0415454
34	6314	8411	955	(6515, 10308)	-2098	-0.31	-0.31	0.0201599
35	6080	13405	1249	(10927, 15884)	-7325	-1.11	-1.11	0.0344354
36	20630	18479	1473	(15555, 21403)	2151	0.33	0.33	0.0479204
37	3393	8873	1655	(5589, 12158)	-5480	-0.84	-0.84	0.0604700
38	3557	8814	1184	(6465, 11164)	-5257	-0.79	-0.79	0.0309425
39	12630	10300	1552	(7219, 13382)	2330	0.36	0.35	0.0532334
40	2211	6584	1459	(3688, 9481)	-4373	-0.67	-0.66	0.0470338
41	3580	5944	1282	(3398, 8490)	-2364	-0.36	-0.36	0.0363294
42	8060	7685	973	(5753, 9617)	375	0.06	0.06	0.0209233
43	13607	15219	1335	(12570, 17868)	-1611	-0.24	-0.24	0.0393407
44	5990	12470	1285	(9920, 15020)	-6481	-0.98	-0.98	0.0364530
45	8606	7545	1038	(5486, 9605)	1061	0.16	0.16	0.0237788
46	4505	7116	1187	(4760, 9472)	-2612	-0.39	-0.39	0.0311219
47	30167	11578	1670	(8263, 14893)	18589	2.85	2.97	0.0616013
48	4134	2161	1140	(-101, 4424)	1972	0.30	0.30	0.0287012
49	1743	1807	1264	(-701, 4316)	-64	-0.01	-0.01	0.0352716
50	14235	21559	1738	(18109, 25009)	-7324	-1.13	-1.13	0.0667148
51	16885	23669	1292	(21104, 26235)	-6785	-1.03	-1.03	0.0368976
52	27809	34815	1534	(31771, 37860)	-7007	-1.07	-1.07	0.0519641
53	39612	36684	1758	(33195, 40173)	2928	0.45	0.45	0.0682531
54	36837	32403	1176	(30069, 34737)	4434	0.67	0.67	0.0305295
55	37702	31008	986	(29051, 32964)	6694	1.01	1.01	0.0214677
56	38711	33444	1226	(31010, 35877)	5267	0.80	0.79	0.0331856
57	35586	30330	1350	(27649, 33010)	5256	0.80	0.80	0.0402829
58	51195	32657	1276	(30124, 35189)	18538	2.81	2.91	0.0359485
59	39774	34872	1215	(32460, 37284)	4903	0.74	0.74	0.0326179
60	48173	44489	1745	(41026, 47952)	3684	0.57	0.56	0.0672285
61	38709	34052	1113	(31842, 36262)	4657	0.70	0.70	0.0273787
62	23568	32079	1095	(29905, 34254)	-8511	-1.28	-1.29	0.0265023
63	37743	33472	1070	(31348, 35596)	4270	0.64	0.64	0.0252902
64	47496	41696	1504	(38710, 44683)	5800	0.88	0.88	0.0499895
65	34303	26228	1315	(23618, 28839)	8075	1.22	1.23	0.0381975
66	23245	29661	1474	(26734, 32587)	-6416	-0.98	-0.98	0.0480107
67	14712	19830	1522	(16810, 22851)	-5119	-0.78	-0.78	0.0511443
68	17663	23304	1269	(20786, 25823)	-5641	-0.85	-0.85	0.0355562
69	16578	23597	1236	(21143, 26050)	-7019	-1.06	-1.06	0.0337524
70	37165	31851	1478	(28918, 34784)	5314	0.81	0.81	0.0482273
71	39837	34038	1087	(31881, 36196)	5798	0.87	0.87	0.0260902
72	21099	27373	1334	(24724, 30022)	-6274	-0.95	-0.95	0.0393257
73	43579	37045	1285	(34494, 39597)	6534	0.99	0.99	0.0364933
74	30185	36619	1349	(33940, 39297)	-6434	-0.98	-0.98	0.0402226
75	47291	39163	1514	(36158, 42167)	8128	1.24	1.24	0.0506156
76	22413	28317	1766	(24812, 31821)	-5904	-0.91	-0.91	0.0688509
77	15821	22654	1596	(19486, 25822)	-6833	-1.05	-1.05	0.0562593
78	30942	37261	1414	(34455, 40067)	-6319	-0.96	-0.96	0.0441403
79	17560	24051	1261	(21547, 26555)	-6491	-0.98	-0.98	0.0351482
80	47056	41729	1562	(38629, 44829)	5326	0.81	0.81	0.0538699
81	19108	26574	1061	(24467, 28680)	-7466	-1.12	-1.13	0.0248691
82	39556	33452	998	(31472, 35433)	6104	0.92	0.92	0.0219857
83	17081	23971	1285	(21421, 26520)	-6890	-1.04	-1.04	0.0364484
84	32734	19774	1751	(16298, 23250)	12960	1.99	2.03	0.0677493
85	18972	26023	1124	(23791, 28255)	-7050	-1.06	-1.06	0.0279300
86	20746	27426	1028	(25386, 29466)	-6680	-1.00	-1.00	0.0233271
87	40721	33383	1021	(31357, 35409)	7337	1.10	1.10	0.0230113
88	19965	26868	1268	(24352, 29385)	-6904	-1.04	-1.05	0.0355019
89	21224	29384	1271	(26861, 31906)	-8160	-1.23	-1.24	0.0356615
90	15518	21330	1414	(18523, 24137)	-5812	-0.88	-0.88	0.0441612
91	36950	32576	1189	(30215, 34937)	4374	0.66	0.66	0.0312503
92	21349	29301	1065	(27186, 31416)	-7952	-1.20	-1.20	0.0250768
91	36950	32576	1189	(30215, 34937)	4374	0.66	0.66	0.0312503
92	21349	29301	1065	(27186, 31416)	-7952	-1.20	-1.20	0.0250768
93	36149	30121	1527	(27091, 33152)	6028	0.92	0.92	0.0515003
94	48824	43694	1684	(40351, 47038)	5130	0.79	0.79	0.0626722
95	43753	39094	1655	(35809, 42378)	4660	0.71	0.71	0.0604770
96	37134	32128	1272	(29603, 34654)	5006	0.76	0.76	0.0357536
97	20984	28147	1003	(26156, 30139)	-7163	-1.08	-1.08	0.0222368
98	34780	28865	1388	(26110, 31620)	5914	0.90	0.90	0.0425546
99	19516	27825	1178	(25486, 30163)	-8309	-1.25	-1.26	0.0306588
100	19444	25428	1345	(22757, 28098)	-5983	-0.91	-0.91	0.0399730

Fits and Diagnostics for Unusual Observations

Obs	charges	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI	Cook's D
3	21984	1166	1292	(-1397, 3730)	20818	3.15	3.31	0.0368447	0.10
9	28923	11287	1491	(8328, 14245)	17637	2.69	2.78	0.0490773	0.09
47	30167	11578	1670	(8263, 14893)	18589	2.85	2.97	0.0616013	0.13
58	51195	32657	1276	(30124, 35189)	18538	2.81	2.91	0.0359485	0.07

Obs	DFITS	
3	0.647856	R
9	0.631694	R
47	0.759788	R
58	0.562593	R

R Large residual

PREDICTIONS for the dependent variables based on own values:

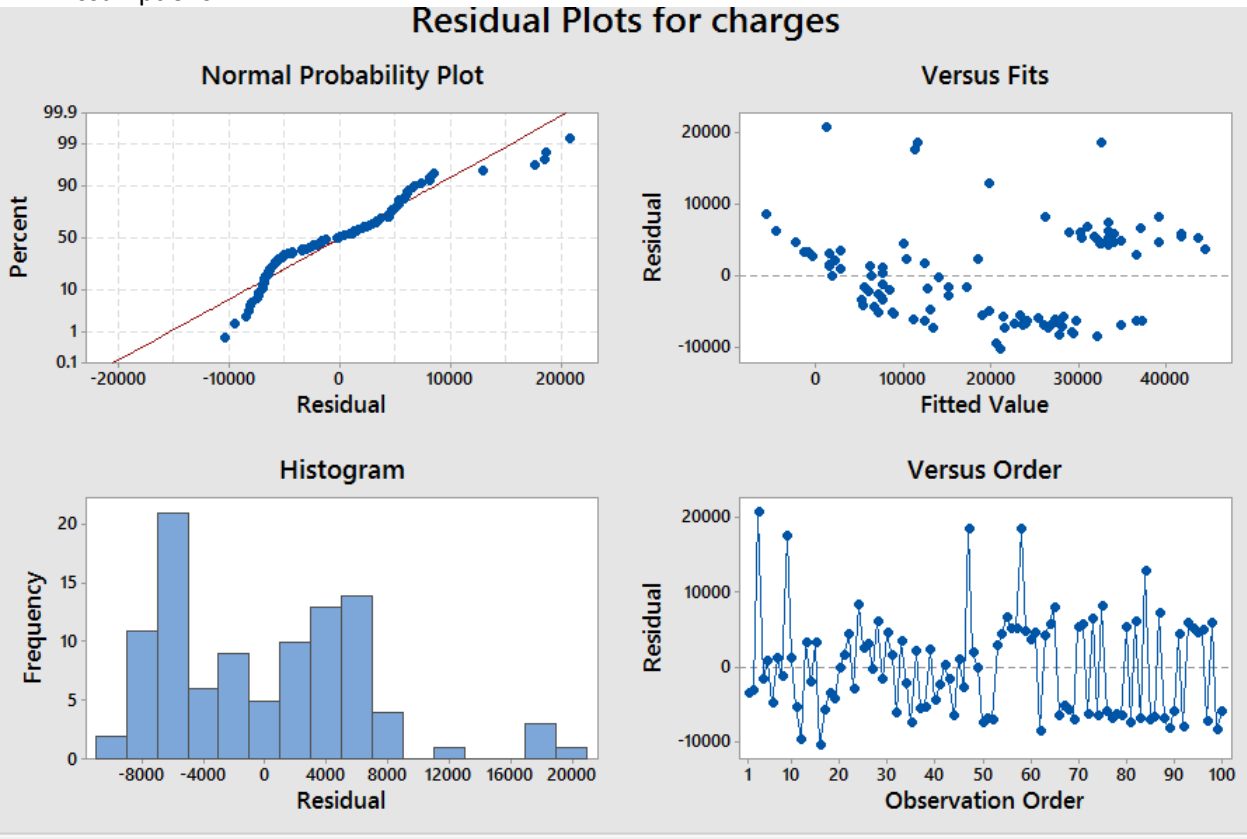
Charges = -25415 + 753 *BMI + 287.4 * Age + 22615 * Smoker_1

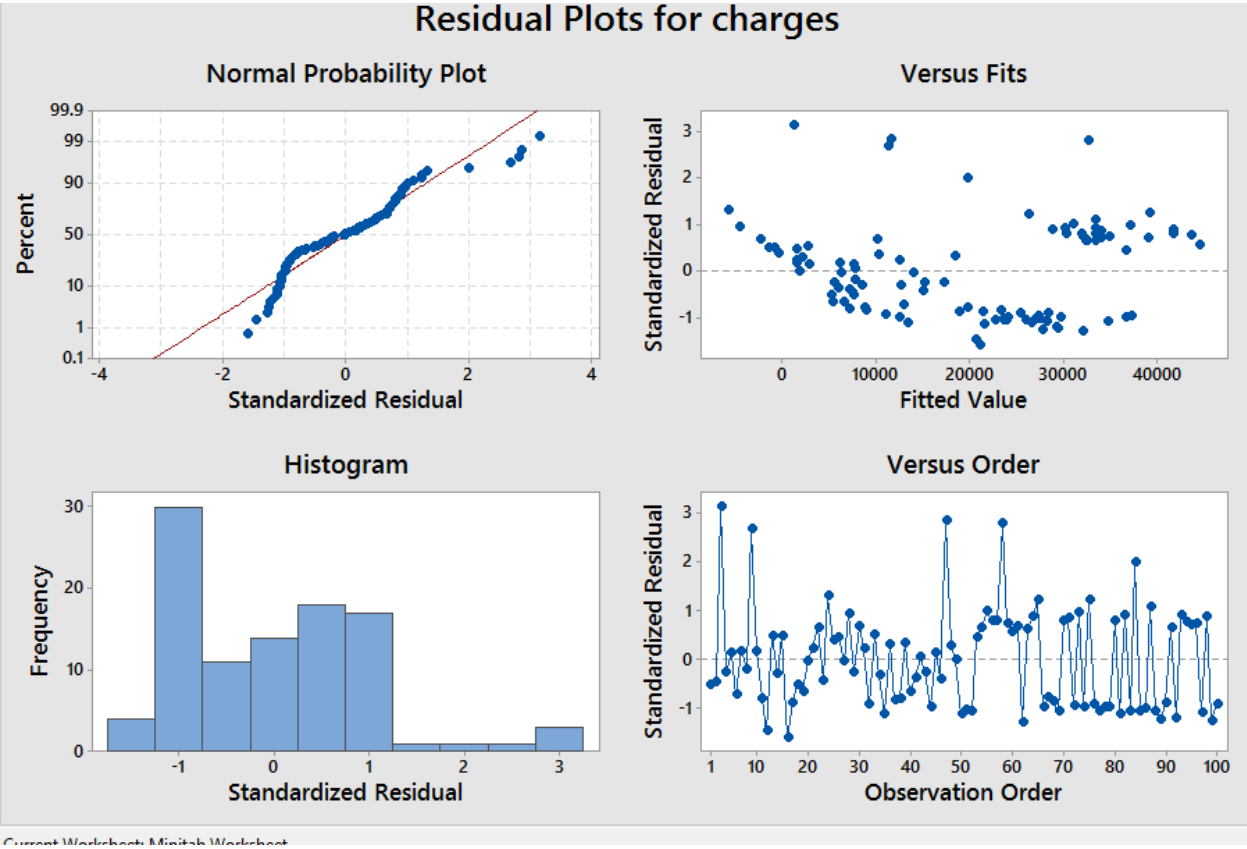
```
> predict(reg,data.frame(i..age=30,bmi=25,smoker=1),interval="confidence")
       fit      lwr      upr
1 24647.11 22308.2 26986.02
```

So, for a person with age 30, bmi 25 and who is a smoker, the predicted charges are 24647.11 with a confidence interval of 22308.2 and 26986.02

ASSUMPTIONS ASSESSMENT

LINE Assumptions





Test for

Normality

From the histogram plot of the residuals, the residuals seem to have a right skewed normal plot. There are few errors which are highly positive on the right end of the distribution. Its probably close to a right skewed normal distribution.

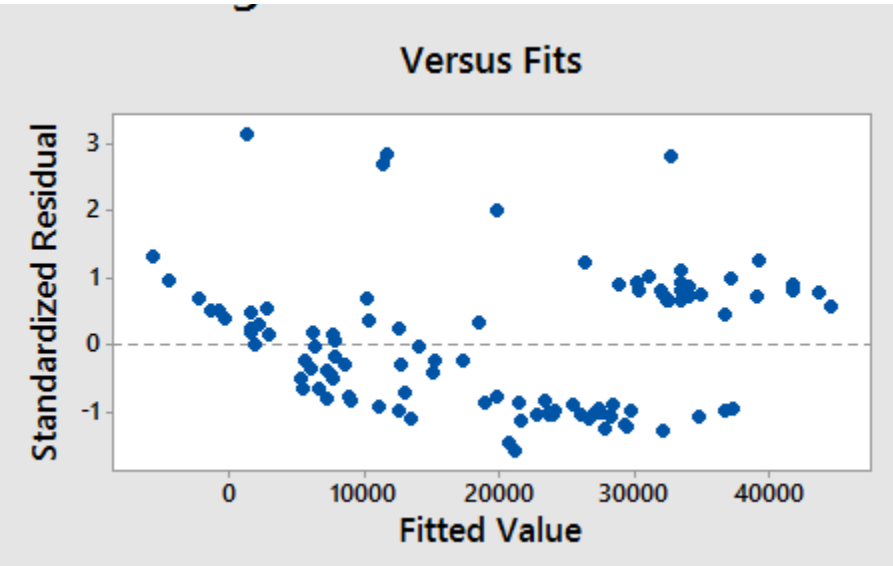
Normality Plot – or Q - Q plot

From the normal probability plot, Residuals are fairly normal. Though some parts of the head and tail seem as outliers. The rest of the residual plot is fairly normal.

Linearity

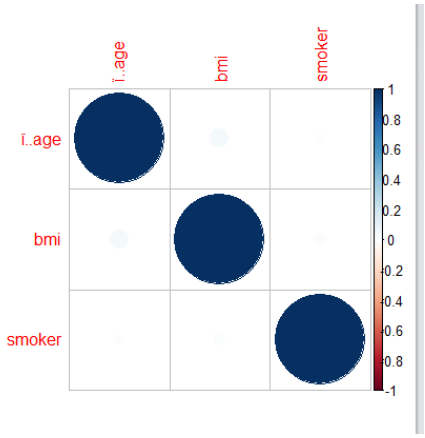
From the plot and the model which gave a good R squared which implies that there is a linear relationship. The residuals are fairly equally distributed on both the sides which also indicate that there is linearity

Equality of Variances



From the graph its evident that the residuals are randomly scattered around the horizontal line at 0. This suggests that the assumption the relationship has constant variance is reasonable. Also, there no specific pattern in the residuals. Hence Equality of Variance assumption is fulfilled.

Independence




```
                ï..age        bmi      smoker
ï..age  1.00000000 0.04588919 0.01131971
bmi     0.04588919 1.00000000 0.01856332
smoker  0.01131971 0.01856332 1.00000000
```

The correlation matrix shows that the correlation between the independent variables is close to zero. So the variables are independent and there is no multi-collinearity.

```
> vif(reg)
  ï..age      bmi    smoker
1.002220 1.002437 1.000455
```

PARTIAL TEST

Partial F test

Analysis to isolate the impact of one independent variable in the model fit.

DFR – Degrees of freedom of reduction =1

SSR (Full model) – SSR (Reduced Model) /MSE (Full model)

FULL MODEL

SUMMARY OUTPUT								
Regression Statistics								
Multiple F	0.892499							
R Square	0.796555							
Adjusted I	0.790198							
Standard E	6728.484							
Observati	100							
ANOVA								
	df	*SS*	*MS*	*F*	*gnificance F*			
Regressio	3	1.7E+10	5.67E+09	125.2908	4.5E-33			
Residual	96	4.35E+09	45272504					
Total	99	2.14E+10						
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-25415	3929.563	-6.46765	4.16E-09	-33215.1	-17614.9	-33215.1	-17614.9
age	287.3988	44.85235	6.407664	5.49E-09	198.3676	376.4301	198.3676	376.4301
bmi	753.0083	116.3303	6.473018	4.06E-09	522.0944	983.9221	522.0944	983.9221
smoker	22614.97	1346.003	16.80158	2.43E-30	19943.17	25286.76	19943.17	25286.76

REDUCED MODEL

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.842344
R Square	0.709544
Adjusted R Squ	0.703555
Standard Error	7998.045
Observations	100

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	1.52E+10	7578941280	118.4789	9.10735E-27
Residual	97	6.2E+09	63968723.28		
Total	99	2.14E+10			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-15685.5	4308.2	-3.640846197	0.000438	-24236.0782	-7134.91	-24236.1	-7134.91
bmi	787.0657	138.1356	5.697774443	1.3E-07	512.9047217	1061.227	512.9047	1061.227
smoker	22705.36	1599.885	14.1918741	2.09E-25	19530.03366	25880.69	19530.03	25880.69

SSE - FULL	4346160344							
SSE - Reduced	6204966158							
Diff	1858805813		H0	Age has no impact				
MSE - Full	45272503.59		HA	Age has impace				
Fstat	41.0581626							
Critical value	3.937116911							
SSR - Full	17016688373		Numeratc	1858805813	21362848717	SST - Full	Partial R2	0.299567
SSR - Reduced	15157882560		Denomina	6204966158	15157882560	SSR - Reduced		

The F-stat is greater than the critical value

The partial ^2 is 0.29 indicates that keeping the other two independent variables constant, the “AGE” variable explains 29% variance in the model.

```
Model 1: data$charges ~ data$bmi + data$smoker
Model 2: data$charges ~ ï..age + bmi + smoker
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1     97 6204966158
2     96 4346160344  1 1858805813 41.058 5.485e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

NULL – Age has no impact

Alternate- Age has impact

Since F = 41.058 and p value is less than 0.05, we can reject null hypothesis. And we have evidence to suggest that AGE has significant contribution towards the “Charges”.