**Can we predict the medical charges of our customers?**

Prepared by Jennifer Tatum

March 21, 2017

To determine if our customer’s medical charges are predictable, I’ve used linear regression for analysis of six demographic and geographic variables. My research shows that medical charges may be explained by various forms of these six variables however there are statistical problems with the data.

**Key Takeaways**

* Age, BMI, children and smoker status are significant variables in predicting medical charges.
* BMI values greater than 30 (obesity) have a larger effect on medical charges than BMI does as a whole.
* Obese smokers have the largest effect on medical charges.

**Analysis Summary**

To answer whether we can predict the medical charges of our customers, I considered 1338 observations of six demographic and geographic variables: age, sex, BMI, children, smoker status, and region. Age ranges from 18 to 64 with a fairly equal distribution of occurrences. Sex is fairly equal with 662 females and 676 males. BMI ranges from 15 to 53 with the mean and median both close to 30 (obese level). Children range from 0 to 5 with most observations having a 0 or 1 value. There are 274 smokers in the data. The region is fairly equally distributed with just slightly more observations from the Southeast. Medical charges range from $1,122 to $63,770 with a median and mean of around $9k and $13k respectively.

To analyze the data, I first conducted simple linear regression of each variable compared to medical charges. This provided me with the amount of charges that can be explained by each variable independently. It is also worth noting that children was not statistically significantly on its own, but is when used with other variables.

**Full Model**

I conducted a multiple linear regression with all six variables against medical charges. Four variables proved to be more significant than the others: age, BMI, children and smoker status.

Age has a strong statistical significance with a t-value of 21.6, a very low p-value and a .001 significance code. Age can explain almost 9% of charges when looked at alone. For each year increase in age, medical charges increase by $256.90 in the full model.

BMI has a strong statistical significance with a t-value of 11.9, a very low p-value and a .001 significance code. BMI can explain almost 4% of charges independently. For each increase in BMI, charges increase by $339.20.

Smoker has a strong statistical significance with a t-value of 57.7, a very low p-value and a .001 significance code. Smoker can explain almost 62% of charges when looked at alone. A smoker’s medical charges are $23,848.50 more than non-smokers holding everything else constant.

**Regression Analysis of Full Model**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | t-value | p-value | R^2 of simple models | Change in charges due to increase in variable | statistical significance |
| Age | 21.6 | Very small | 9% | 257.70 | \*\*\* |
| SexMale | -0.4 | .036 | .3% | 1387.20 | \* |
| BMI | 11.9 | Very small | 4% | 393.87 | \*\*\* |
| Children | 3.5 | .013 | .4% | 683.10 | \* |
| SmokerYes | 57.7 | Very small | 62% | 23,616.00 | \*\*\* |
| Region |  |  |  |  |  |
| NW  NE  SW  SE | -0.7  0.7  -1.3  -1.4 | 0.5  0.5  0.2  0.2 | .6% total | -353.00  353.00  -607.10  -682.10 | -  -  -  - |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Reduced Model**

The full model can explain 75.1% of medical charges. When the insignificant variables (sex and region) are removed, the reduced model can explain 74.9% of medical charges. Because there is such a small decrease in the amount of charges that can be explained by the model, I will use the reduced model going forward as it is simpler and more streamlined. However it is not greatly more significant than the full model when looking at a partial f-test (f-value 1.63 and p-value .17)

**Reduced Model plus Obesity**

Next I added obesity (BMI>30) to the reduced model and the percent of medical charges explained by the model increase very slightly from 74.9% to 75.5%. Obesity provides a statistically significant explanation of charges. Obese individuals experience on average $2,772.76 more in medical expenses than those who are not obese (2904.47-131.71).

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7816.28 1233.73 -6.34 3.23e-10 \*\*\*

age 258.03 11.78 21.91 < 2e-16 \*\*\*

bmi 131.71 44.93 2.93 0.003428 \*\*

children 473.87 136.41 3.474 0.000529 \*\*\*

smokeryes 23819.41 407.10 58.51 < 2e-16 \*\*\*

obesityTrue 2904.47 547.33 5.31 1.31e-07 \*\*\*

This model is statistically more significant than the reduced model based on a partial f-test resulting in an f-value of 28.16 and p-value that is very small.

**Reduced Model plus Obesity and Obese Smoker**

Next I added smokers who are obesity (smokerYes and BMI>30) to the model and the percent of medical charges explained by the model significantly increased from 75.5% to 86.2%. Obese smoker provides a statistically significant explanation of charges. Obese smokers experience $7,082.27 more in medical expenses than those who are not obese smokers (19,684+809-13,412).

Estimate Std. Error tvalue Pr(>|t|)

(Intercept) -5097.88 929.89 -5.48 5.02e-08 \*\*\*

age 264.23 8.84 29.88 < 2e-16 \*\*\*

bmi 97.70 33.74 2.90 0.00384 \*\*

children 512.42 102.40 5.00 6.35e-07 \*\*\*

smokeryes 13412.40 445.19 30.13 < 2e-16 \*\*\*

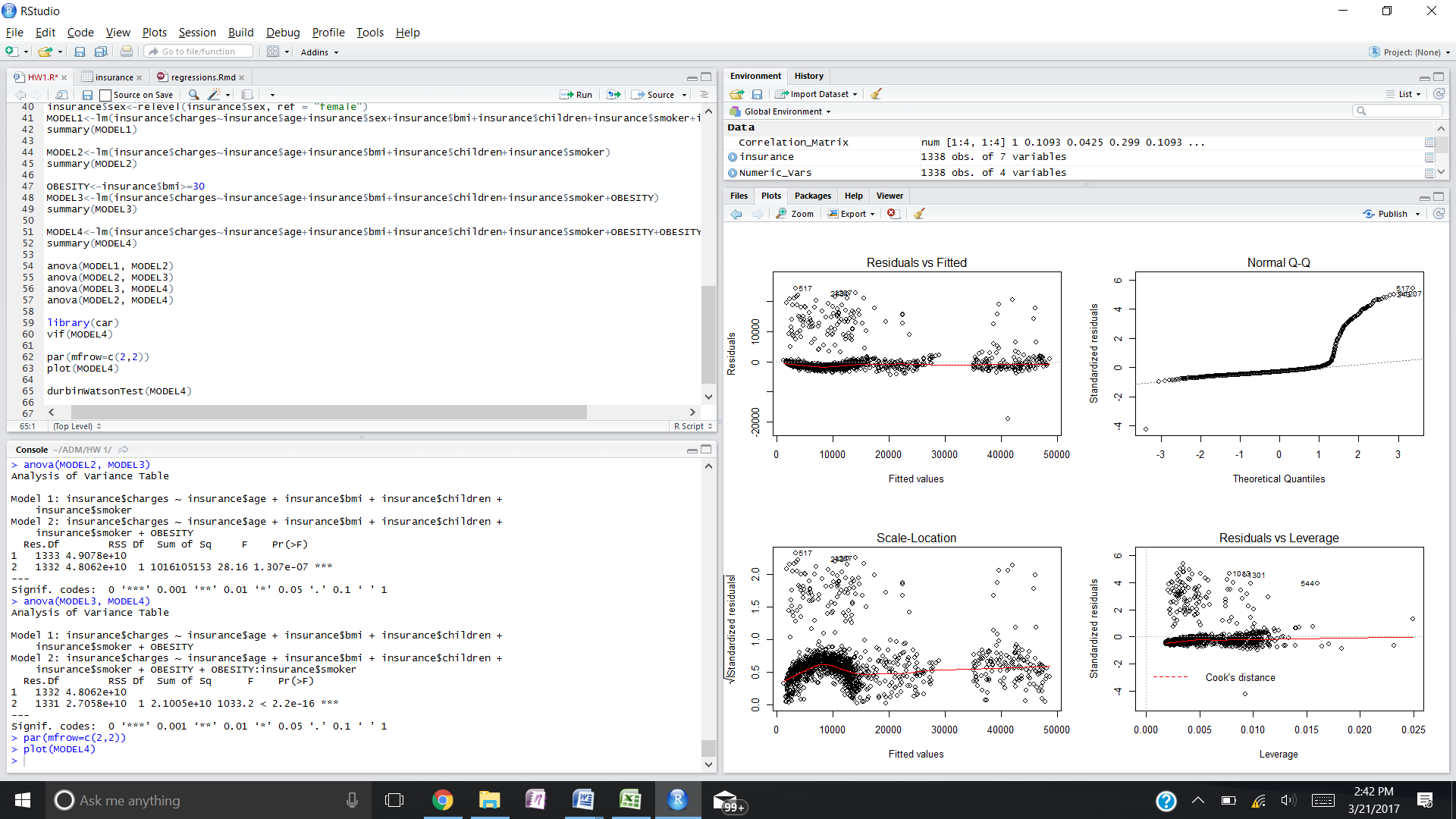
obesityTrue -809.79 426.76 -1.90 0.05798 .

smoker:obese 19684.87 612.39 32.14 < 2e-16 \*\*\*

The model with Obesity and Obese Smokers is the preferred model. 86% of the change in medical charges is explained by this model compared to 74% and 75% for the first two models respectively. And this model is statistically more significant than the previous model based on a partial f-test resulting in an f-value of 1033.2 and p-value that is very small.

**Statistical problems with the chosen model**

Finally we will review the assumptions of linear regression to see if there are any problems with the chosen model.



The upper left graph displays the models linear and additive characteristics. Because the result is mostly linear I will assume that the model does pass this assumption noting that there are several observations outside of this line.

The upper right graph displays the models distribution of errors. This graph should show a straight line. Because it is an S shape, this implies that there is error that is not explained by the model. The model does not pass this assumption.

The lower left graph is testing homoskedasticy. The variance in error should not change and should be equally distributed across the graph left to right. There appears to be a much larger amount of observations when medical charges are lower than when they are higher.

The lower right graph is used to observe outliers in the data. The majority of observations are fairly close to the line. However there are several outliers when charges are lower.

The Durbin-Watson value is 2.06 verifying that there is no autocorrelation in the variables.

**Commentary**

The statistical strength of the recommended model is great. It is hard to get better that explaining 86% of the target. However given there are so many issues with the assumptions, I do not recommend using this model for predicting medical charges.