Final project

Cameron Pick

3/07/2021

# Medical expenses and insurance premiums

## 1.) Tell us why did I choose this research idea and the dataset.

My research questions were: i. What factors play a role in creating medical expenses?

1. What role does geographic region play in medical expenses?
2. Is it reasonable to conclude that current medical expenses can predict future medical expenses?

The final model shows which are the most important factors:

(Intercept) 26.516 < 2e-16 \*\*\*

age 39.908 < 2e-16 \*\*\*

bmi 9.008 < 2e-16 \*\*\*

children 8.088 1.83e-15 \*\*\*

smokeryes 64.246 < 2e-16 \*\*\*

regionsoutheast -4.602 4.75e-06 \*\*\*

regionsouthwest -3.593 0.000344 \*\*\*

Geographic plays a role, but mostly for the southern parts of the country. I choose this dataset because it is interesting to know why insurance premiums are so expensive. If many people have high medical expenses, insurance companies must charge higher premiums to cover their costs along with claim payouts.

## 2.)Summarize the problem statement I addressed.

I addressed what variables lead to higher expenses thus possibly higher premiums.

## 3.)Summarize the methodology employed. Why do I think your method is appropriate? Refer to the literature.

My methodology was first create a simple linear model with expenses as dependent and age, sex, bmi, children,smoker, and region as independent. Then, I ran diagnostic plots to see if linear regression assumptions were met. The plots showed that normality of residuals and homoscedasticity assumptions may have been violated. The others: Linearity, independence, and multicolinearity seem to be satisfied. The normality assumption:

Shapiro-Wilk normality test

data: sresid

W = 0.89839, p-value < 2.2e-16

The homoscedasticity assumption:

studentized Breusch-Pagan test

data: model

BP = 121.59, df = 8,

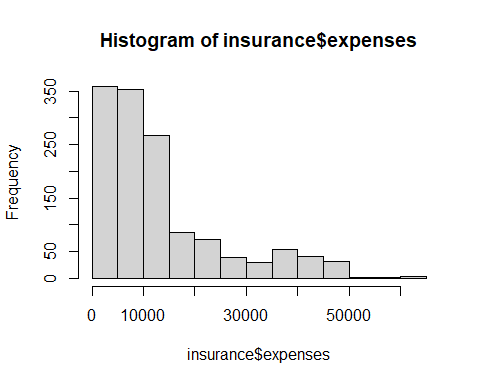
p-value < 2.2e-16

Non-constant Variance Score Test Variance formula: ~ fitted.values

Chisquare = 235.802, Df = 1, p = < 2.22e-16

A histogram of my dependent variable shows that it is skewed which is expected since we don’t expect too many expenses on the right side.

hist(insurance$expenses)



Next, I wanted to identify outliers to see if this helps with the violations. I do this by finding outliers using Cook’s Distance and saving those values and dropping those. There were 68. Next, I dropped the region of northwest and dropped sex.

Still, previously addressed assumptions are violated. To tackle this further, I do a Box-Cox transformation to see what value is best for transformation. This showed .26 was appropriate. So,I created a model with expenses raised to .26 along with no changes to independent variables. Assumptions may still be violated but confidence and prediction intervals are not being considered so I will ignore it. Too, data set is large. I think this methodology is appropriate because I fitted the model and checked if assumptions were met.Then, tried various ways to met them, unsuccessfully due to nature of data.

summary(model1)

##   
## Call:  
## lm(formula = expenses^0.26 ~ ., data = insurance3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5337 -0.4452 -0.1033 0.2483 4.7138   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.787407 0.180548 26.516 < 2e-16 \*\*\*  
## age 0.089830 0.002251 39.908 < 2e-16 \*\*\*  
## bmi 0.047915 0.005319 9.008 < 2e-16 \*\*\*  
## children 0.208778 0.025813 8.088 1.83e-15 \*\*\*  
## smokeryes 5.005310 0.077908 64.246 < 2e-16 \*\*\*  
## regionsoutheast -0.366513 0.079646 -4.602 4.75e-06 \*\*\*  
## regionsouthwest -0.282901 0.078739 -3.593 0.000344 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9721 on 956 degrees of freedom  
## Multiple R-squared: 0.8603, Adjusted R-squared: 0.8595   
## F-statistic: 981.5 on 6 and 956 DF, p-value: < 2.2e-16

## 4.)Summarize the interesting insights that my analysis provided.

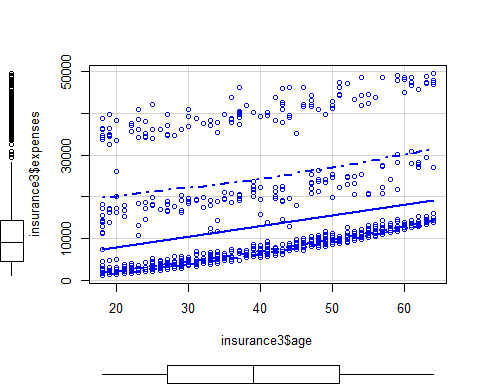
Some interesting insights is the model has good r-squared value. Predictions from the model seem to do well. However, a couple of the underlying assumptions were not satisfied. It seemed the older and if you were a smoker medical expenses were higher. For example, holding all other equal if you are a smoker, you expenses increase by $5. Might not seem much but when you transform back to orginial scale that is almost $500!

## 5.)Summarize the implications to the consumer (target audience) of my analysis.

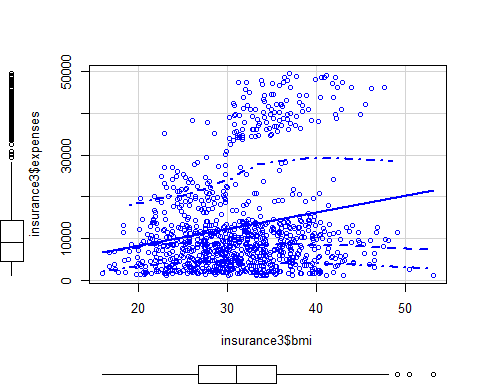
Implications are that there is way for insurance companies to predict what factors lead to higher medical expenses. Thus, they can adjust and calculate premium rates based on that.

## 6.)Overall, write a coherent narrative that tells a story with the data as I complete this section

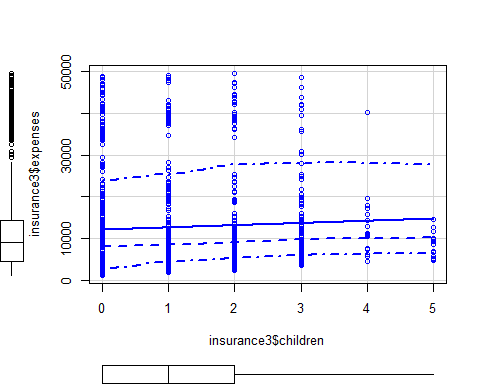
scatterplot(insurance3$age,insurance3$expenses)



scatterplot(insurance3$bmi,insurance3$expenses)



scatterplot(insurance3$children,insurance3$expenses)



As noted before, the older you get and the higher bmi the higher medical expenses tend to be.

## 7.)Discuss the limitations of my analysis and how I, or someone else, could improve or build on it.

Limitations are a couple of the assumptions were not met. Therefore, it can lead to incorrect interpretations about the data. Also, it can lead to bad confidence and prediction intervals. I or someone could improve it by maybe using nonparametic modeling or bootstrapping. Or finding a way to get the dependent variable more normal.

To answer the final research question, is it reasonable to conclude that current medical expenses can predict future medical expenses?

Let’s run some predictions. I’ll pick an actual data point and see how the model predicts it. Let’s pick value 32 from insurance3 dataset. The age is 18, bmi is 26.3, the have 0 kids, not a smoker, live in the northeast region. They have a $2198.19 for expenses.

pred4 <- predict(model1, newdata=data.frame(age=18, bmi=26.3, children=0, smoker="no", region="northeast"))  
pred4^3.85

## 1   
## 2542.516

This isn’t too far off. If a future 18 year came looking for insurance, we could expect them to have around $2542 for medical expenses and thus charge a premium based on that.

Next two expenses are $14313.85 and $46889.26, respectively.

pred5 <- predict(model1, newdata=data.frame(age=64, bmi=36, children=0, smoker="no", region="southeast"))  
pred5^3.85

## 1   
## 13808.5

pred6 <- predict(model1, newdata=data.frame(age=64, bmi=33.9, children=0, smoker="yes", region="southeast"))  
pred6^3.85

## 1   
## 52168.12

The other variables are constant except for bmi, which are close and smoker. The results are stunning that a smoker has way more medical costs than non-smoker. To answer, I believe past medical expenses can be used to predict future expenses. But limitations listed above must be considered.