Multiple Regression Analysis - Health Insurance Data

Liam Daly
Nicholas Khemvisay
Luis Colin Cornejo

Introducing Our Topic

Our Main Objective

We want to determine which of our six independent variables are best to include in a model that predicts insurance charges, our response variable.

Motivation

Using our analysis, Hospitals and healthcare providers may be able to….

- leverage our results to allocate resources efficiently,
- anticipate demand for specific services
- tailor preventive care programs for different groups

**((1) US Institute of Medicine)*

The Data

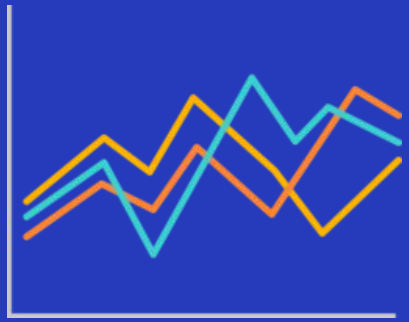
	age <int>	sex <chr>	bmi <dbl>	children <int>	smoker <chr>	region <chr>	charges <dbl>
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855
6	31	female	25.740	0	no	southeast	3756.622

Predictor Variables:

Age (X_1): numerical continuous
Sex (X_2): categorical
BMI (X_3): numerical continuous
Number of Children (X_4): numerical discrete
Smoker (X_5): categorical
Region (X_6): categorical

Response Variable:

Insurance Charges (Y): numerical continuous

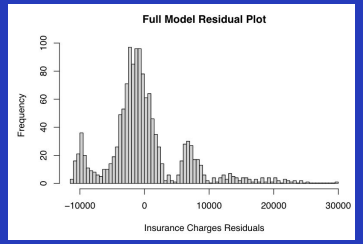


Hypothesis

We believe that Age (X_1), BMI (X_3), Number of Children (X_4), and Smoker (X_5) are the best predictors for insurance charges.

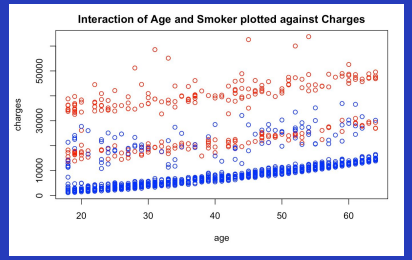
Highlights from EDA

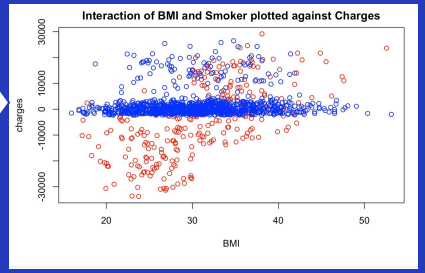
1. Non-Normal Histogram of charges

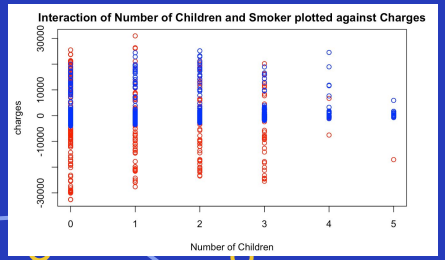


$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon$$

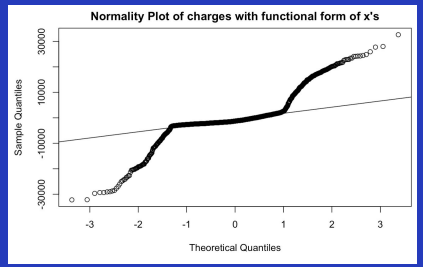
2. Observed Linear Relationships with smoker

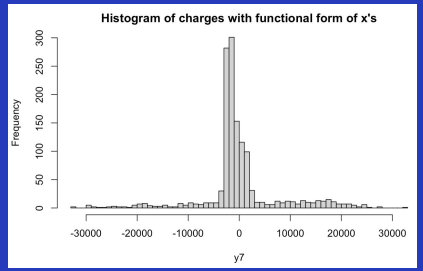






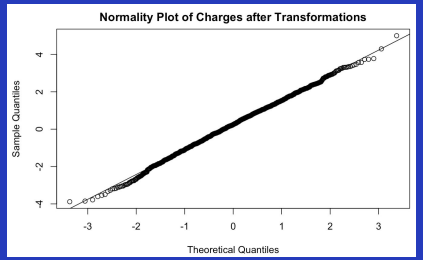
3. Functional form of x's without transformations.

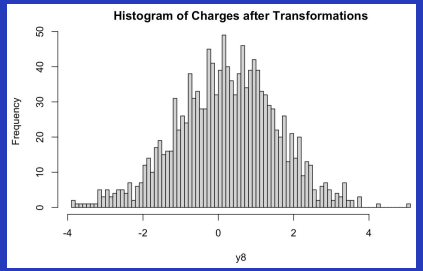




$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_4 + \beta_4 (x_1)(x_5) + \beta_5 (x_3)(x_5) + \beta_6 (x_4)(x_5) + \epsilon$$

4. Functional form of x's with transformations.





$$log(y) = 0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_4 + \beta_4 (x_1)(x_5) + \beta_5 (x_3)(x_5) + \beta_6 (x_4)(x_5) + \epsilon$$

We decided to keep all numerical x variables as well as their interactions with smoker.

Predictor Variables:

- x_1 (age)
- x_3 (BMI)
- x_4 (children)
- $(x_1)(x_5)$ (age * smoker)
- $(x_3)(x_5)$ (BMI * smoker)

Final Model

$$log(y) = 0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_4 + \beta_4 (x_1)(x_5) + \beta_5 (x_3)(x_5) + \epsilon$$

```
Call:
lm(formula = log(y) ~ 0 + x1 + x3 + x4 + z1 + z2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8438 -0.6565  0.2469  1.1400  4.9933

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
x1  0.075854   0.002758  27.501  < 2e-16 ***
x3  0.173248   0.003763  46.041  < 2e-16 ***
x4  0.248180   0.030639   8.100 1.23e-15 ***
z1 -0.020711   0.005973  -3.467 0.000543 ***
z2  0.075454   0.007843   9.621  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.361 on 1333 degrees of freedom
Multiple R-squared:  0.9779,    Adjusted R-squared:  0.9779
F-statistic: 1.182e+04 on 5 and 1333 DF,  p-value: < 2.2e-16
```

All coefficients are statistically significant

Model explains 97.79% of variance in log(charges)

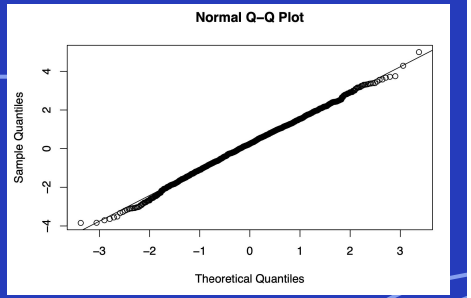
The residuals' spread and the low residual standard error suggest an excellent model fit

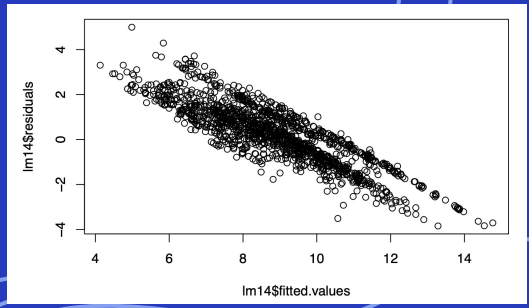
Assumptions on residuals are satisfied and multicollinearity is not present

```
dwt(lm14)


##  lag Autocorrelation D-W Statistic p-value
##    1       0.0148239      1.969075   0.534
##  Alternative hypothesis: rho != 0
```

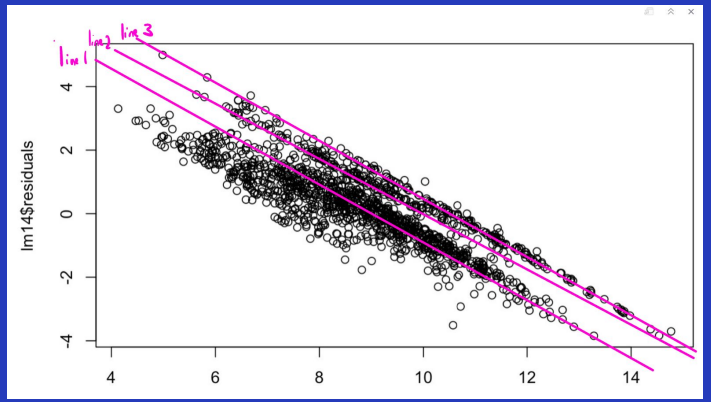
```
       x1       x3       x4       z1       z2
9.533344 9.998561 1.797989 8.850391 8.943583
```



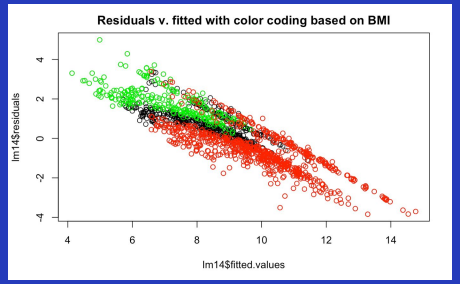


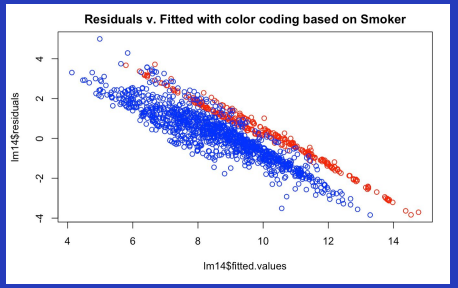
Findings

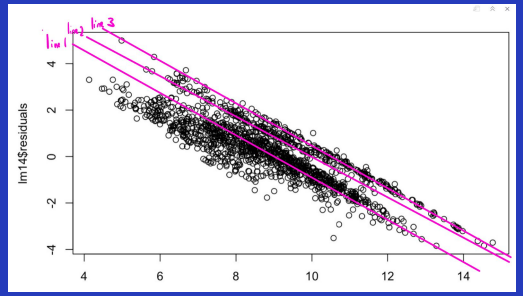
- Age, BMI and number of children **positively influence** and predict log(charges).
- The interaction between age and smoker has a **slight negative** coefficient.
- Number of children has the **most influence** on log(charges), followed by BMI and Age.
- Three **distinct downward linear patterns** in our residuals v. fitted plot.



Conclusions, Limitations and Future Work







Conclusion:

The best predictors for the natural log of insurance charges are age, BMI, number of children, the interaction between age and smoker, and the interaction between BMI and smoker.

Healthcare providers can anticipate higher demand for services involving…

- Elderly individuals
- Overweight individuals
- Individuals with more children
- Elderly or overweight smokers

Limitations:

- Transformed Charges variable
- Possible existence of a hidden categorical variable.
- Our analysis is limited to the data that's present in the csv file.

Works Cited

(1) National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Care Services; Committee on Health Care Utilization and Adults with Disabilities. Health-Care Utilization as a Proxy in Disability Determination. Washington (DC): National Academies Press (US); 2018 Mar 1. 2, Factors That Affect Health-Care Utilization. Available from: https://www.ncbi.nlm.nih.gov/books/NBK500097/

(We referenced all works below when formulating our hypothesis)

(2) Bui AL, Dieleman JL, Hamavid H, Birger M, Chapin A, Duber HC, Horst C, Reynolds A, Squires E, Chung PJ, Murray CJ. Spending on Children's Personal Health Care in the United States, 1996-2013. JAMA Pediatr. 2017 Feb 1;171(2):181-189. doi: 10.1001/jamapediatrics.2016.4086. PMID: 28027344; PMCID: PMC5546095. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5546095/

(3) Lantz, Brett. Machine Learning with R - Third Edition. Packt Publishing, 2019.
Ward ZJ, Bleich SN, Long MW, Gortmaker SL. Association of body mass index with health care expenditures in the United States by age and sex. PLoS One. 2021 Mar 24;16(3)e0247307 doi: 10.1371/journal.pone.0247307.PMID: 33760880; PMCID: PMC7990296 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7990296/

(4) Wilcoxon, Nicole. Older Adults Sacrificing Basic Needs Due to Healthcare Costs, June 15 2022 https://news.gallup.com/poll/393494/older-adults-sacrificing-basic-needs-due-healthcare-costs.aspx

(5) Xu X, Bishop EE, Kennedy SM, Simpson SA, Pechacek TF. Annual healthcare spending attributable to cigarette smoking: an update. Am J Prev Med. 2015 Mar;48(3):326-33. doi: 10.1016/j.amepre.2014.10.012. Epub 2014 Dec 10. PMID: 25498551; PMCID: PMC4603661. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4603661/6