Regression on Insurance Charges in America

Cepehr Alizadeh, Seth Harrison, Said Mrad, Max Ryoo

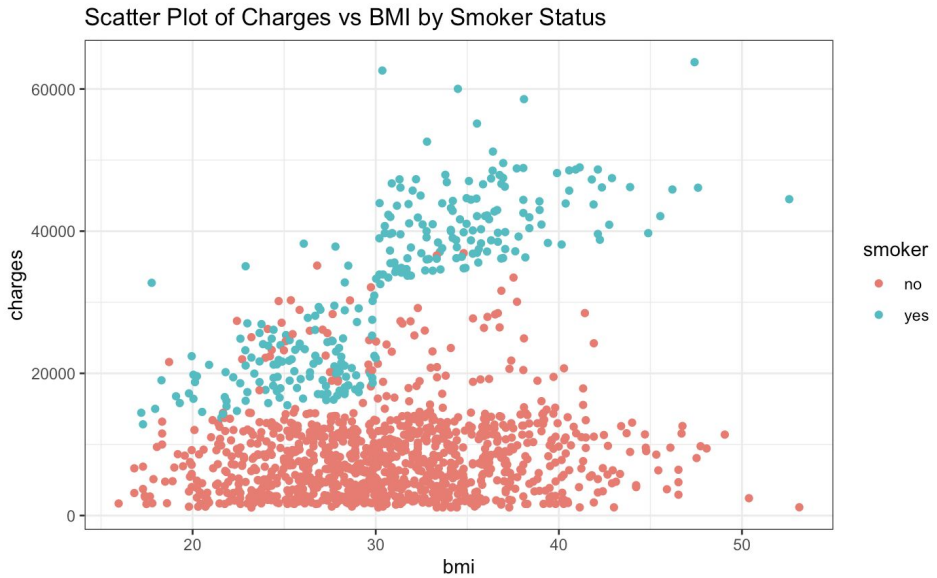
Motivation

- Medical Insurance
  - Over 92% of Americans have it
  - Prices are at times unclear
- How can regression be used to help?
  - Predicting costs of insurance bills
  - Comparing an individual to the average

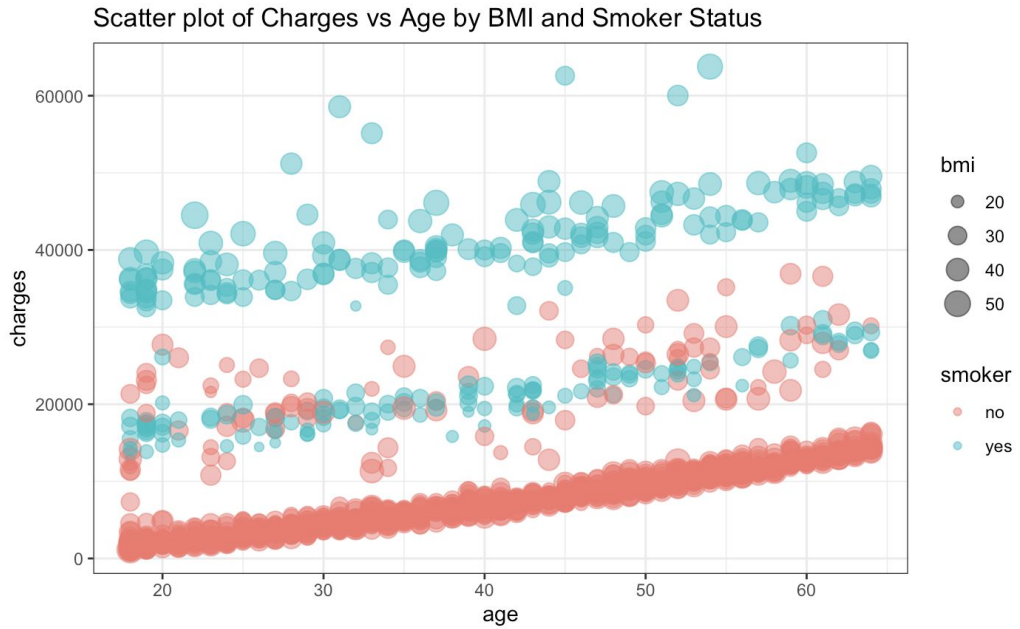
Dataset Introduction

- Dataset
  - Provided by kaggle for analysis
- Predictors for regression
  - Our dataset included six predictor variables
  - Age, Sex, BMI, Children, Smoker, Region
- Response
  - Goal is to predict charges for a beneficiary

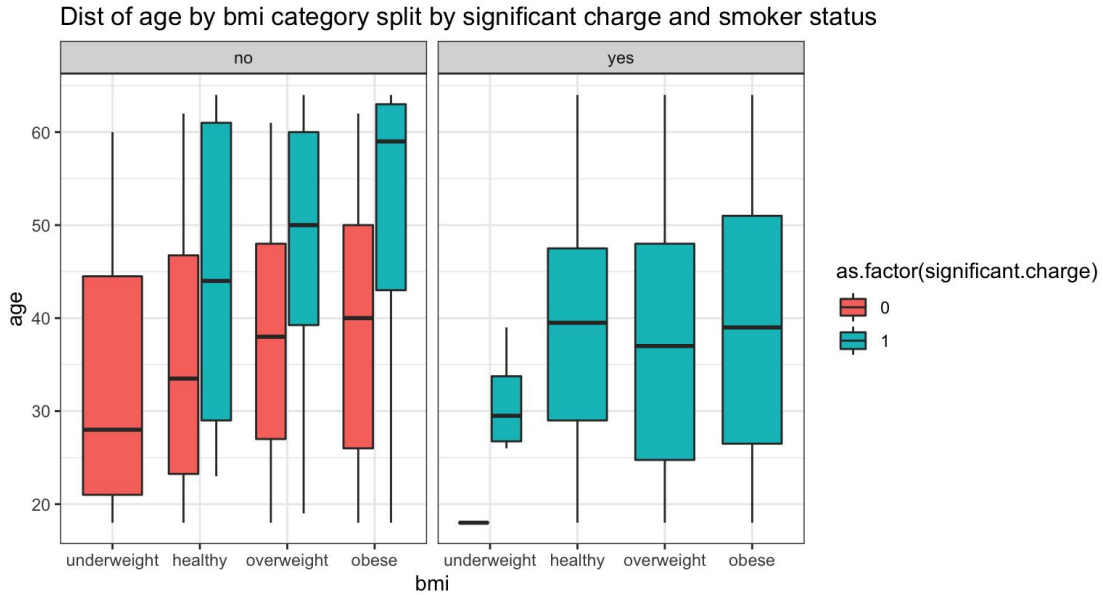
Exploratory Data Analysis



EDA

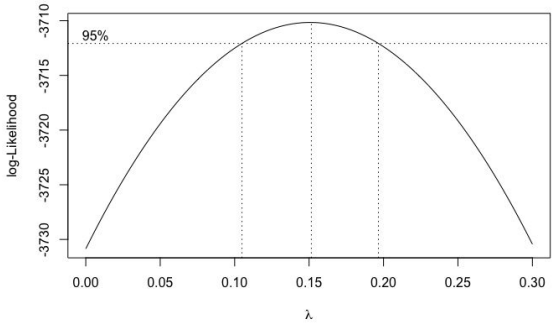

Scatter plot of Charges vs Age by BMI and Smoker Status
charges
60000
40000
20000
0
20
30
40
50
60
age
bmi
20
30
40
50
smoker
no
yes

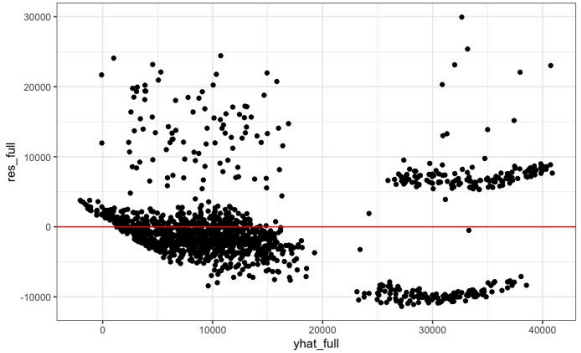
EDA

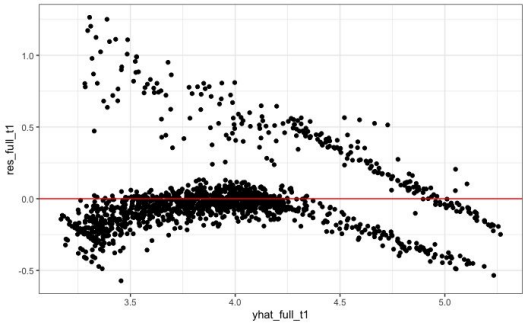

Dist of age by bmi category split by significant charge and smoker status
no
yes
60
50
40
30
20
age
underweight
healthy
overweight
obese
bmi
as.factor(significant.charge)
0
1

Multiple Linear Regression - Full Model

- All predictors
- Residual plots showed two distinct groups
  - Charges < 20,000 vs Charges > 20,000
- Transformation / adding interactions did not help

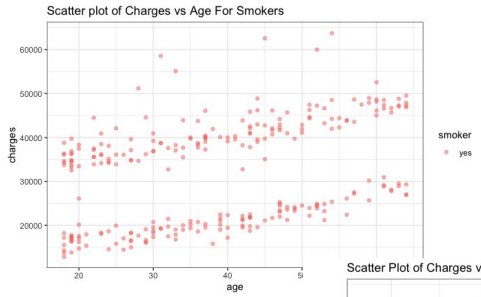


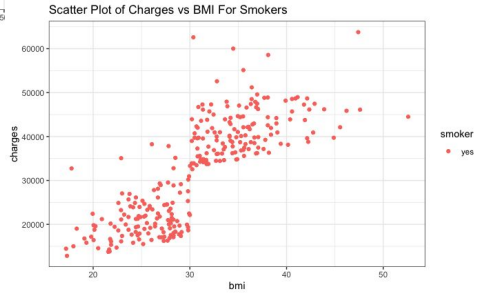




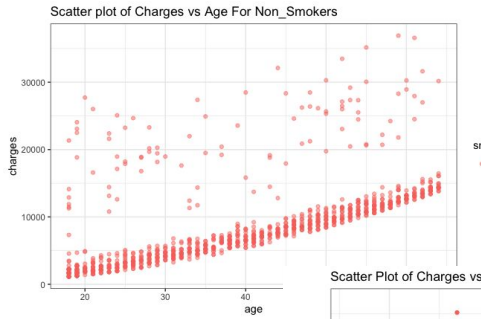
Multiple Linear Regression - Separation of smokers

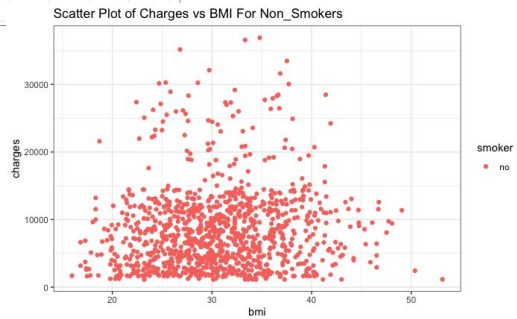
Smokers





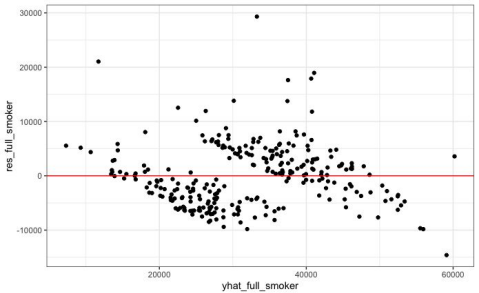
Non Smokers

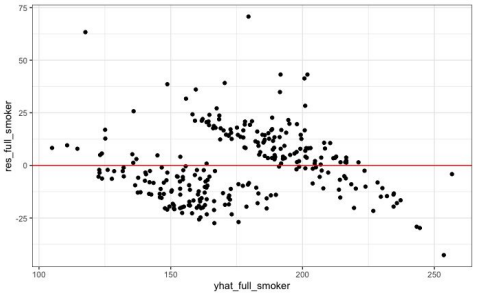


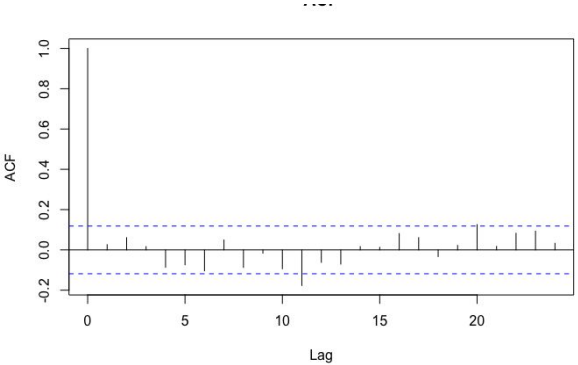


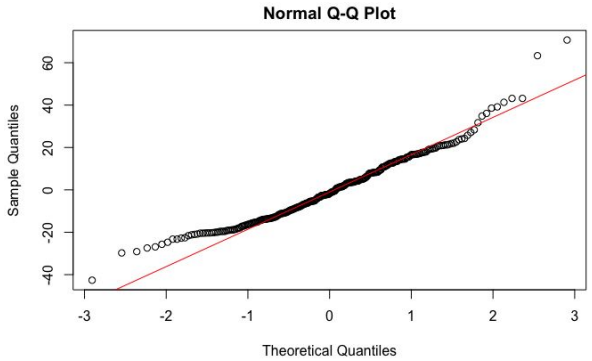
Multiple Linear Regression - Smokers

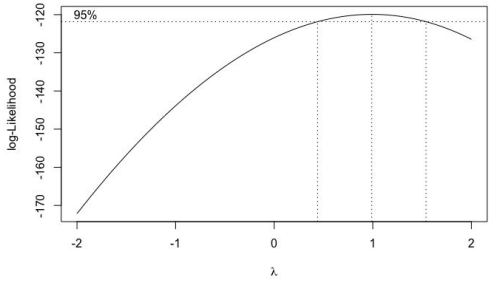
- Stepwise forward selection
  - Predictors Selected (BMI and Age)
  - AIC=4747.41
- Partial F - Test showed we can drop all other predictors
- 11 outliers (leverage & DFFITS)







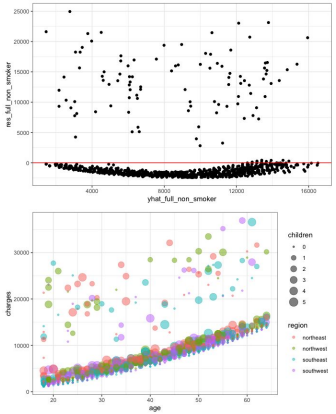




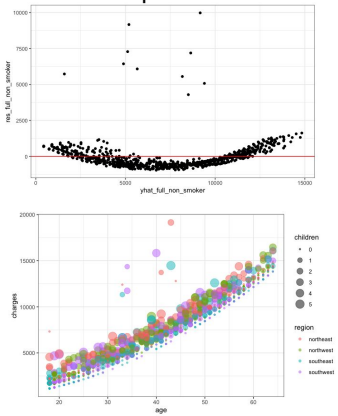
Multiple Linear Regression non-smokers

- Stepwise forward selection
  - Predictors Selected (Age, Region, Children, and Sex)
    - Positive Predictors (Age, Children)
    - Negative (region and sex where northeast and female were reference classes)
  - AIC=17949.26
- Total Dropped = 9.77%

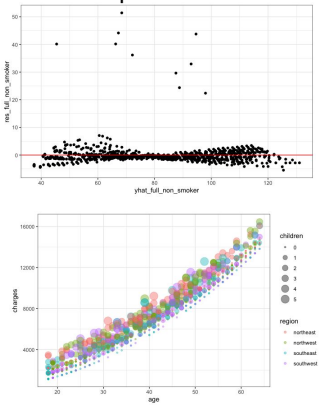
Model 1



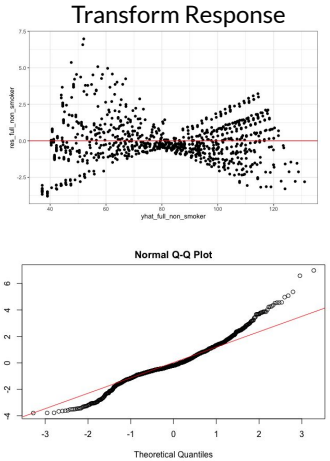
Drop Outliers



Drop Outliers x2

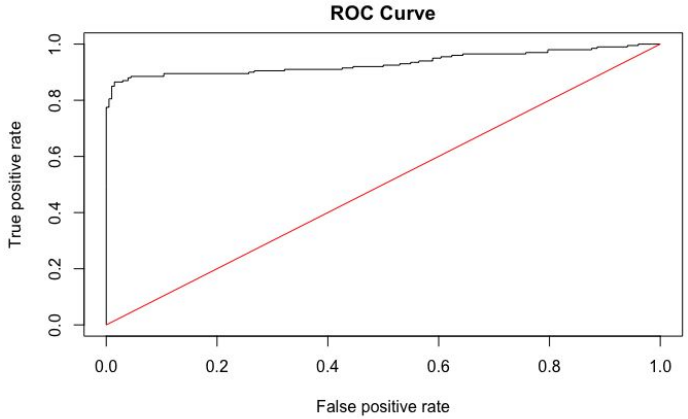


Transform Response



Logistic Regression

- AUC : 0.9335396
- Decrease threshold for context of problem

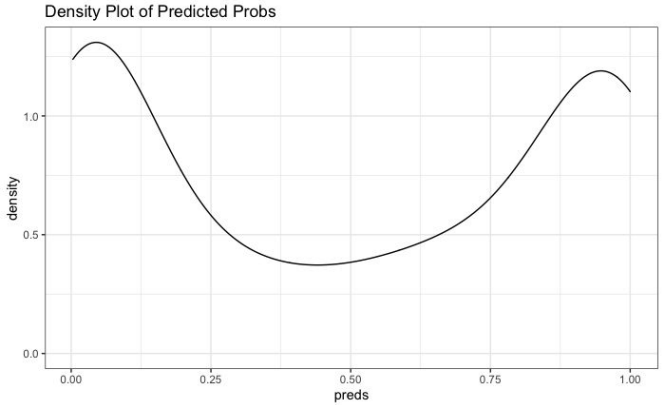


Threshold = 0.5

	FALSE	TRUE
FALSE	181	21
TRUE	21	179

Threshold = 0.25

	FALSE	TRUE
FALSE	147	55
TRUE	20	180



Regression Summary

- Smoker Multiple Linear Regression
  - y^(0.5) = 19.9145 + 4.1245 bmi + 0.7634 age
  - R-squared = 0.7587, Adjusted R - Squared = 0.7569
- Non Smoker Multiple Linear Regression*
  - y^(0.5) = 14.284 - 1.388I1 - 4.108I2 - 3.919I3 + 1.679age + 3.585children - 3.230sex
  - R-squared = 0.9962, Adjusted R - Squared = 0.9962
- Logistic Regression
  - Y = -8.397 + 0.185age + 0.010 bmi + 0.180 children + 22.858 smoker - 0.279I1 - 0.813I2 - 0.887I3 - 0.278sex
    - Null Deviance 1297.57 on 935 DF ; Residual Deviance 503.54 on 927 DF
    - AIC: 521.54

* Multiple Linear Regression assumption not met

Future Work

- Missing predictors in the dataset
  - Socioeconomic factors
  - Pre-existing conditions
- Other datasets
  - >900 medical insurance companies
  - Different regions

Conclusion

- Splitting the dataset for MLR
  - Smokers' model
  - Non-smokers' model
- Logistic regression model satisfactory
  - Opinion of experts on threshold