Medical Insurance Costs

Molly MacKinnon, Richa Jain, Levon Haroutunian, Nihal Dhillon, Jaivardhan Singh

Introduction

- Medical Insurance in the U.S. is notorious for being corrupt, as providers try and capitalize off of clients as much as possible
- Private insurance companies rather than public healthcare for all is the root of this corruption

Parameters

- Data includes 1338 individuals ages 18-64, as 65+ individuals are usually covered by government
- Dependent variable: medical charges
- Independent variables: age, sex, BMI, number of children, smoker status, region of residence
- Quantitative variables: medical charges, age, BMI, number of children
- Qualitative variables: sex, smoker status, region of residency

Parameters Continued

- Charges: measured in USD
- Age: measured in years
- Sex: male or female
- BMI: measured by weight (kilograms) divided by square of height(meters)
- Children: measured by number of children, 0-5
- Smoker: yes or no
- Region: Southeast, Southwest, Northeast, Northwest

Goals

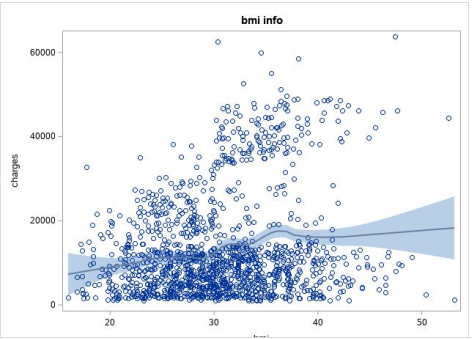
- Determine how much weight each variable has on individual charges
- Find correlations between the variables
- Find the direct weight and influence that the number of children as dependents cause on an individual's insurance plan
- Is there a linear relationship between number of dependents and overall charge?
- Do charges increase at different age stages in life due to certain societal standards?

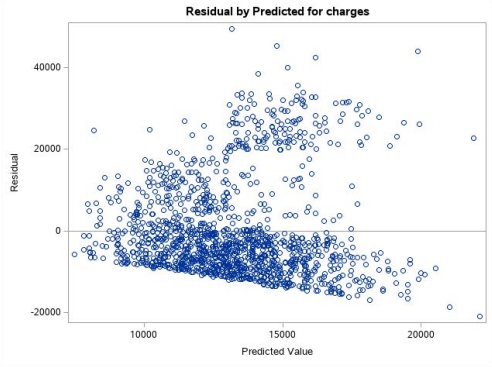
Transformation Disclaimer

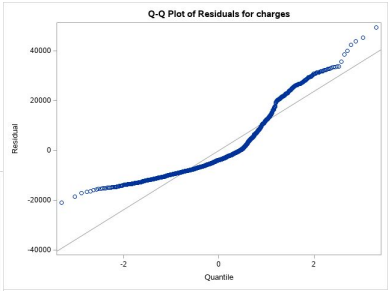
- Our data contains a transformation of charges to log(charges) and BMI to log(BMI)
- This was done in an effort to better fit the data, specifically the data of charges and BMI
    - Nonconstant variance and pattern within original scatterplot
- The duration of our data will contain these transformations
- Other attempted transformations are in the table to the right

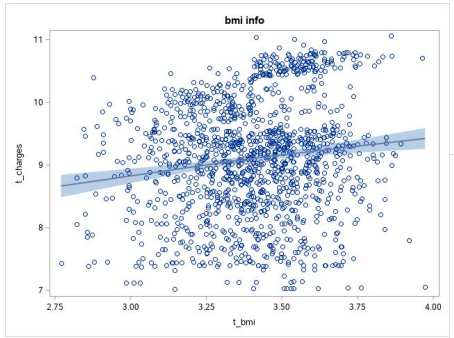
bmi	*	log(charges)
sqrt(bmi)	*	log10(charges)
sqrt(bmi)	*	log2(charges)
sqrt(bmi)	*	log(charges)
sqrt(sqrt(bmi))	*	log(charges)
(bmi)**-1	*	log(charges)
log10(bmi)	*	log2(charges)

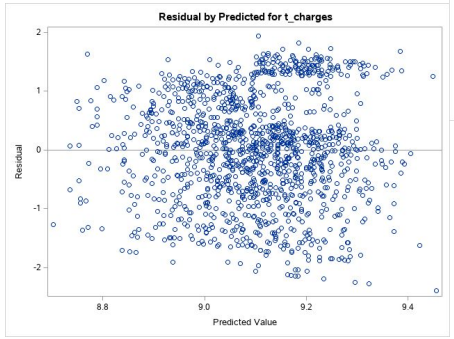
Transformation Disclaimer Continued

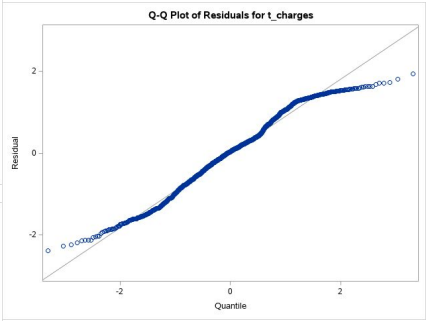










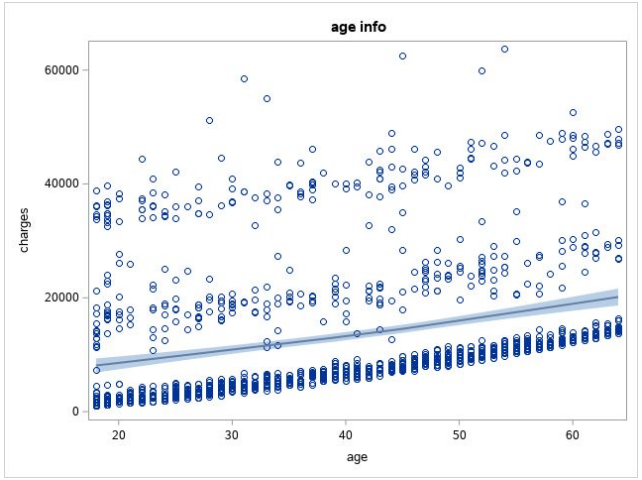


Exploratory Data Analysis: Boxplots

- Charges: strong right skew, many outliers with high log(charges)
- Age: symmetric, very slight right skew, no outliers
- Log(BMI): slight left skew, very few outliers with high log(BMI) and some outliers with low log(BMI)
- Sex, Children, Smoker: not applicable for qualitative parameters

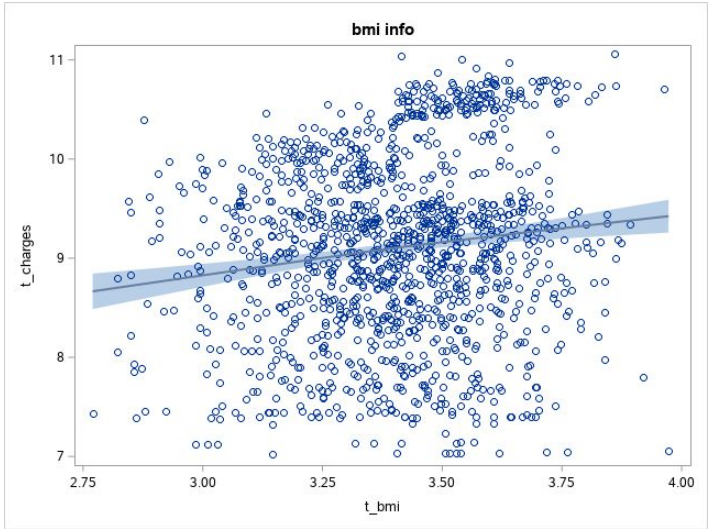
Exploratory Data Analysis: Scatterplots

- LOESS curve of log(charges) and age
- Interesting pattern displays three different linear configurations, with the bottom pattern the most concentrated and the top pattern the least concentrated
- Some outliers with high log(charges)



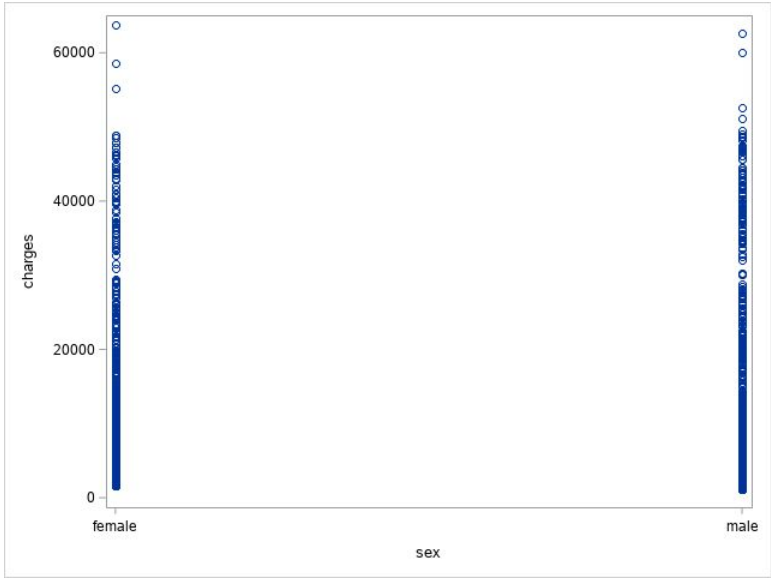
Exploratory Data Analysis: Scatterplots

- LOESS curve of log(charges) and log(BMI)
- Randomly scattered with possible pattern in the upper right quadrant, almost in a quadratic shape
- No significant outliers



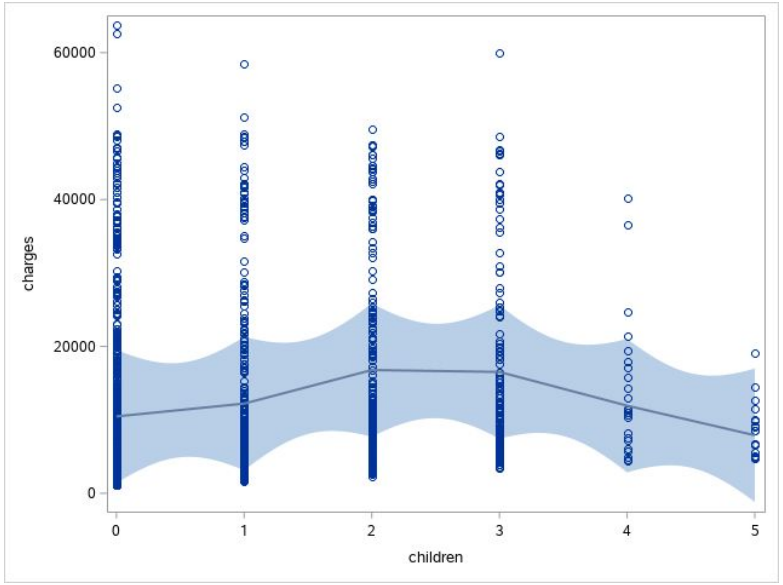
Exploratory Data Analysis: Scatterplots

- Scatterplot of log(charges) and sex
- Some outliers with very high log(charges) for both female and male
- Otherwise, female and male have similar distributions and there is not much of a distinct pattern
- Jump around $30000 for both female and male



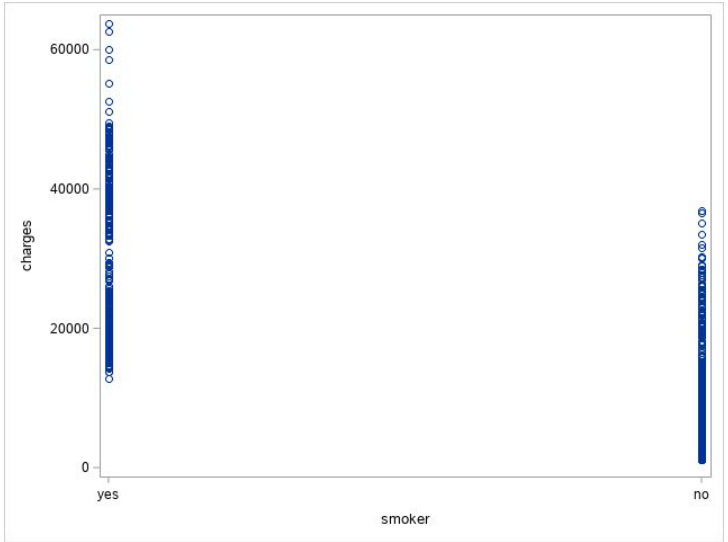
Exploratory Data Analysis: Scatterplots

- LOESS curve of log(charges) and children
  - LOESS curve typically not available for qualitative variables, but since children involves numerics, it is available for investigation
- Some outliers for each number of children, more concentrated towards the lower end
- Bell or hill shape with peak around 2 and 3 children
- More data towards lower end of number of children
- Jump around $30000 for 0-4, as 5 does not reach this high
- It may make sense to treat children as a continuous variable instead of a categorical, despite being discrete
  - Could be potential quadratic relationship between number children and log(charges)



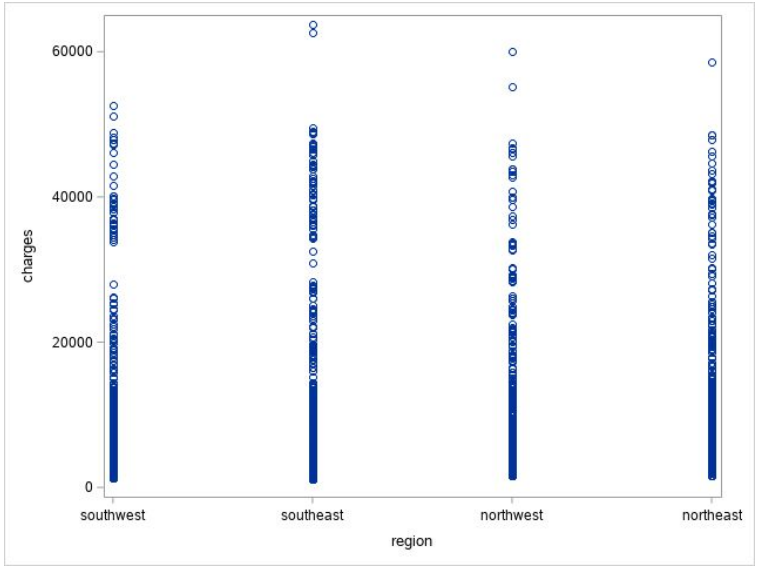
Exploratory Data Analysis: Scatterplots

- Scatterplot of log(charges) and smoker
- Pattern displays that, on average, smokers generate higher charges than non-smokers
- Some outliers with high log(charges) for smokers, non-smokers have very few insignificant outliers
- Jump around $30000 for both yes and no



Exploratory Data Analysis: Scatterplots

- Scatterplot of log(charges) and region
- Some high outliers for every region except Southwest
- Very similar distributions for each region, no indicative pattern
- Jump around $30000 for each region, but the jump is more significant from the left end and gets less significant going towards the right end



Simple Linear Regression Models

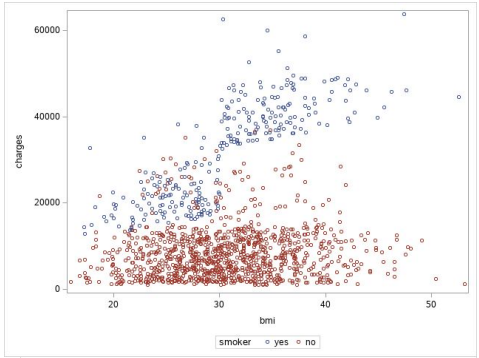
- Age
  - Fitted model: $\widehat{log(charges)} = 3165.88501 + 257.72262(age)$
  - P-value of age is <.0001, which indicates that age parameter is significant
- Log(BMI)
  - Fitted model: $\widehat{log(charges)} = 6.696732 + 0.62632(log(BMI))$
  - P-value of BMI is <.0001, which indicates that log(BMI) parameter is significant
- Sex
  - Fitted model: $\widehat{log(charges)} = 12570 + 1387.17233(isMale)$
  - P-value of isMale <.0361, which indicates that sex parameter might not be significant
  - X=0 for female and X=1 for male

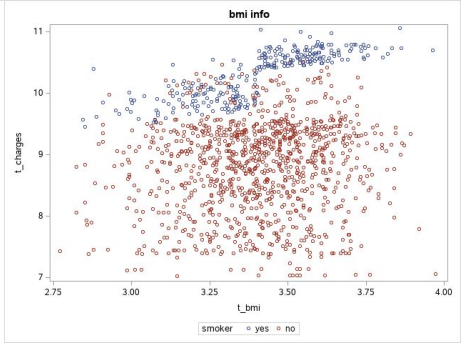
Simple Linear Regression Models Continued

- Children
  - Fitted model: $\widehat{log(charges)} = 12366 - 365.19623(children\_1) + 2707.58813(chidren\_2) + 2989.34277(children\_3) + 1484.68071(children\_4) - 3579.94035(children\_5)$
  - P-values of 1, 4-5 are high and thus insignificant, and p-values of 2,3 are .0035 and .0060, respectively, which indicates some possible significance
  - X=0 for having 0 children, and this logic holds up until X=5 for having 5 children
- Smoker
  - Fitted model: $\widehat{log(charges)} = 8434.26830 + 23616(yes\_smoker)$
  - P-value of yes_smoker is <.0001, which indicates that smoker parameter is significant
  - X=0 for non_smoker and X=1 for yes_smoker
- Region
  - Fitted model: $\widehat{log(charges)} = 12347 + 1059.44717(NE) + 70.63800(NW) + 2388.47406(SE)$.
  - P-value for NE, NW are high and thus insignificant, however p-value for SE is .0097, which could possibly be significant
  - X=0 for SW, X=1 for NE, X=2 for NW, X=3 for SE

Multiple Linear Regression Analysis: Grouped Plots

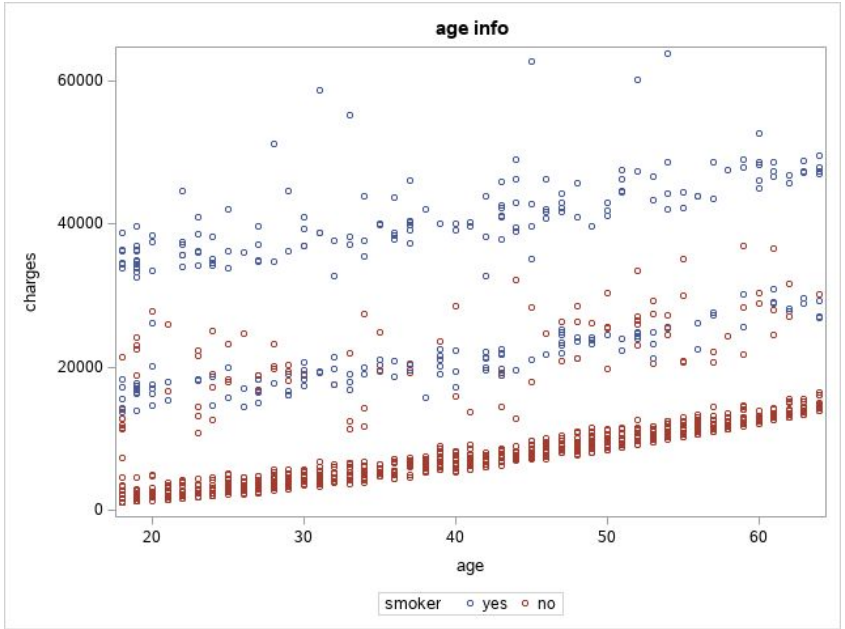
- Scatterplot of charges and BMI grouped by smoker (top) and scatterplot of log(charges) and log(BMI) grouped by smoker (bottom)
- Blue is yes_smoker and red is non_smoker
- Clear pattern where smokers tend to pay more than non-smokers regardless of their BMI, and also that smokers with high BMIs are charged more than smokers with low BMIs
- Indicates important relationship to be explored





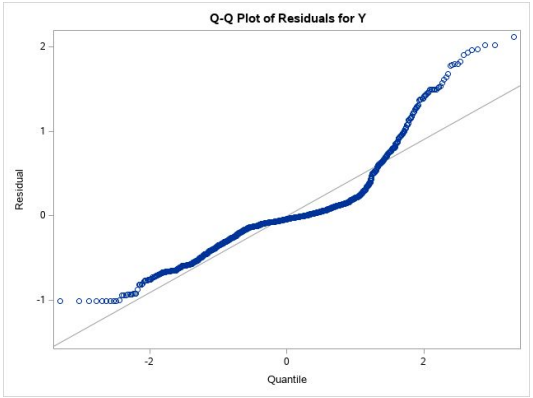
Multiple Linear Regression Analysis: Grouped Plots

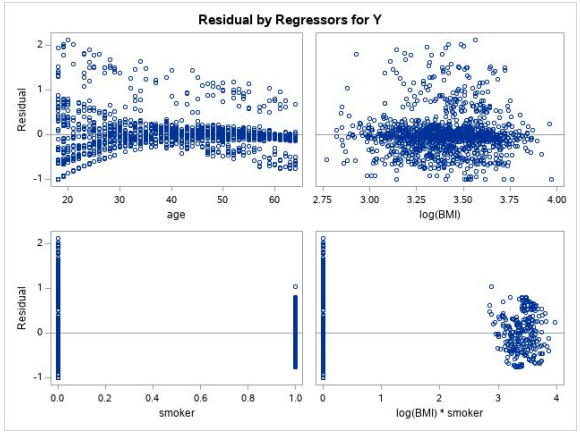
- Scatterplot of log(charges) and age grouped by smoker
- Blue is yes_smoker and red is non_smoker
- Entire bottom linear pattern represents non-smokers with low charges among all ages
- Smokers are charged much more than non-smokers of any age
- Intersection of highest-charged non-smoker and lowest-charged smoker in middle linear pattern
- Indicates important relationship to be explored

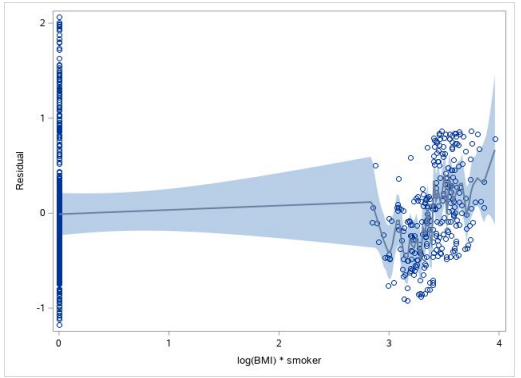


Multiple Linear Regression Models: Base Model

- $Y = b_0 + b_1(age) + b_2(log(BMI)) + b_3(smoker) + b_4(log(BMI) * smoker) + \varepsilon.$
- Assumptions of MLR not met, but there are indications to follow the relationship between log(charges) and log(BMI)*smoker



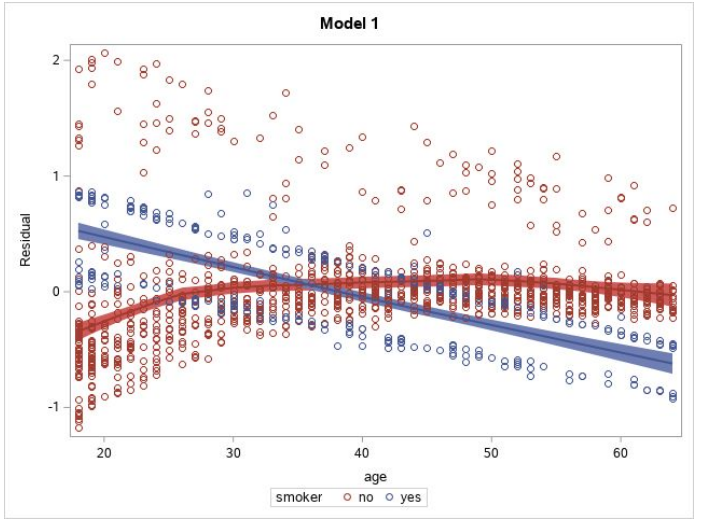




Multiple Linear Regression Models: Model 1

- $Y = b_0 + b_1(age) + b_2(log(BMI)) + b_3(smoker) + b_4(log(BMI) * smoker + b_5(sex) + b_6(region) + b_7(children) + \varepsilon.$

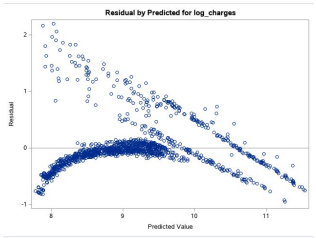
- Did not meet MLR assumptions but found the residual plot of log(charges) and age grouped by smoker interesting to explore, where blue=yes_smoker and red=non_smoker

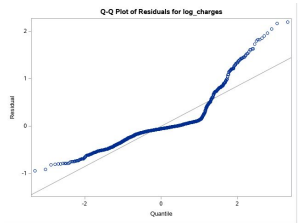


Multiple Linear Regression Models: Model 1.5

$$Y \sim age + log(BMI) + smoker + log(BMI) * smoker + sex + region + c\_children + (c\_children)^2$$

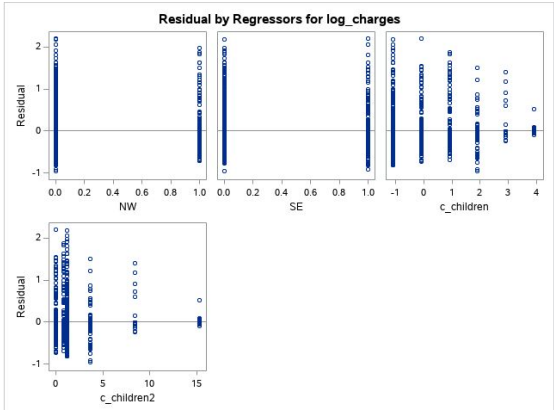
- The centered children parameter, has a low (<.0001) p-value, compared to the four categorical children parameter which were all statistically insignificant under 5% confidence level
- The squared children term however was only significant under 10% confidence level, but may be practically significant
    - Quadratic makes sense visually but may be low amounts of data at higher numbers of children
- However, violations were broken, with similar looking residual and qq-plot

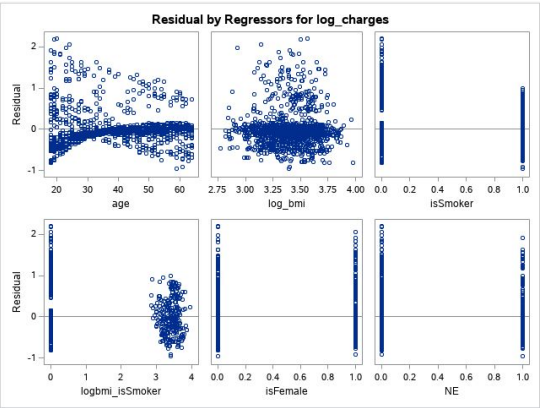




Multiple Linear Regression Models: Model 1.5

- The residuals pattern for c_children looks okay but looks like there is some non-constant variance and pattern for c_children2.
- VIF for all parameters besides smokers and smokers*log(BMI) were < 10

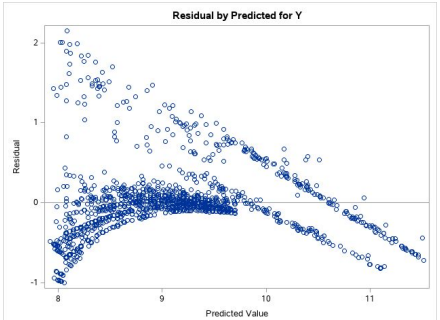


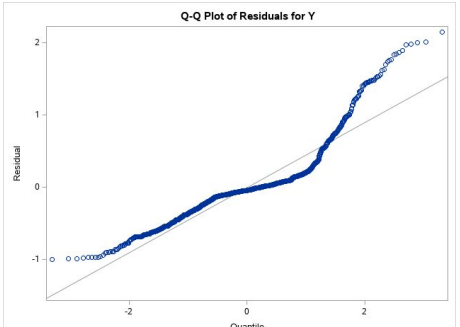


Multiple Linear Regression Models: Exploratory

- Models to the right were explored and ruled not the best fit for our MLR
- All had almost identical plots to the ones below, which do not fit assumptions of an MLR

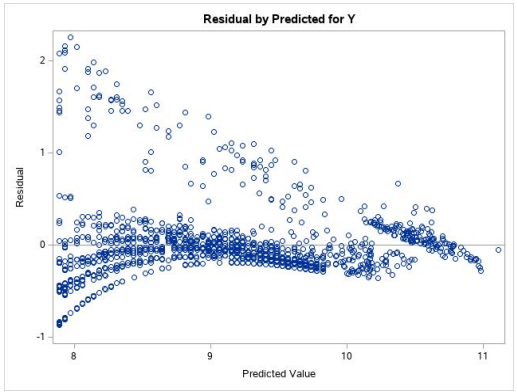
Model Numbers	Model
2	Base + log(BMI)*sex, sex
3	Base + log(BMI)*region, region
4	Base + log(BMI)*children, children
6	Base + age*sex, sex
7	Base + age*region, region
8	Base + age*children, children

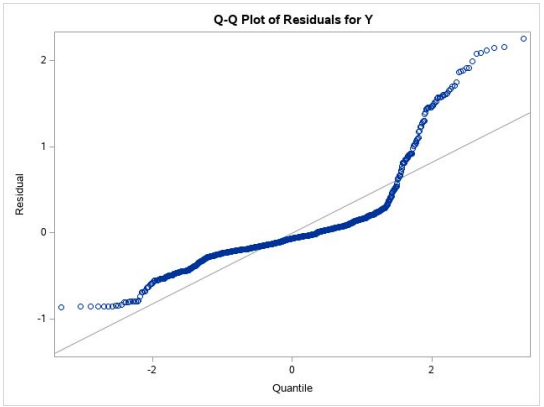


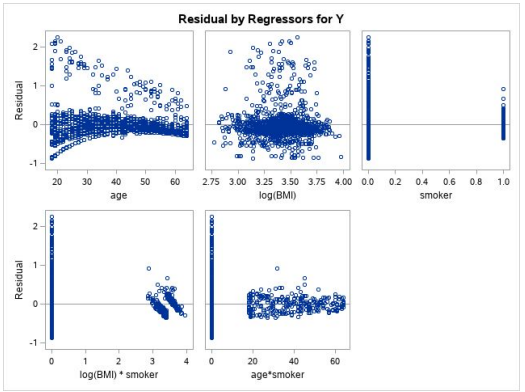


MLR: Model 5

- $Y = b_0 + b_1(age) + b_2(log(BMI)) + b_3(smoker) + b_4(log(BMI) * smoker) + b_5(age * smoker) + \varepsilon$
- First instance of variation in the model assumptions, where the model leans more towards suiting an MLR model
- We are particularly interested in residual plot of log(charges) and log(BMI)*smoker (right, bottom left)

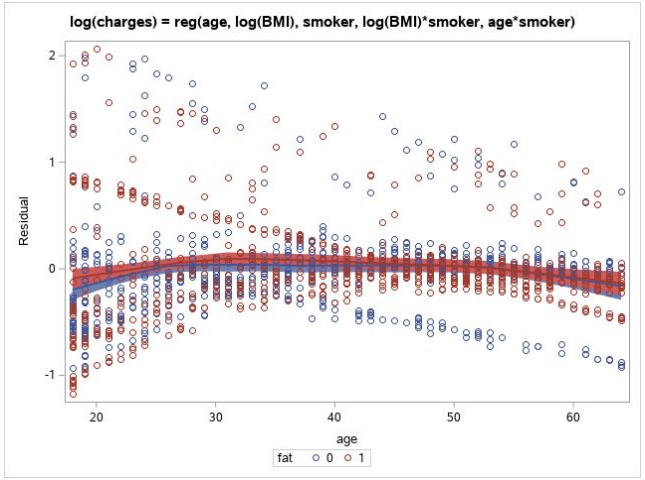


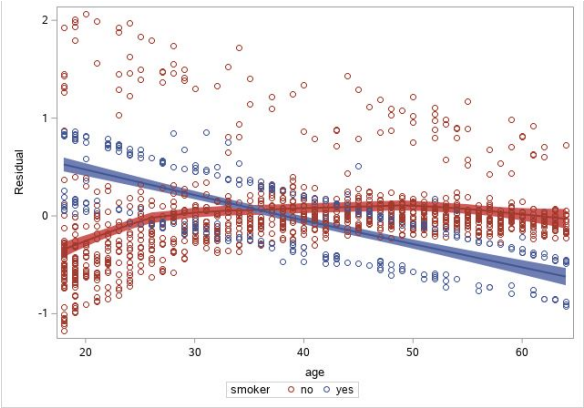




MLR: Model 5.5

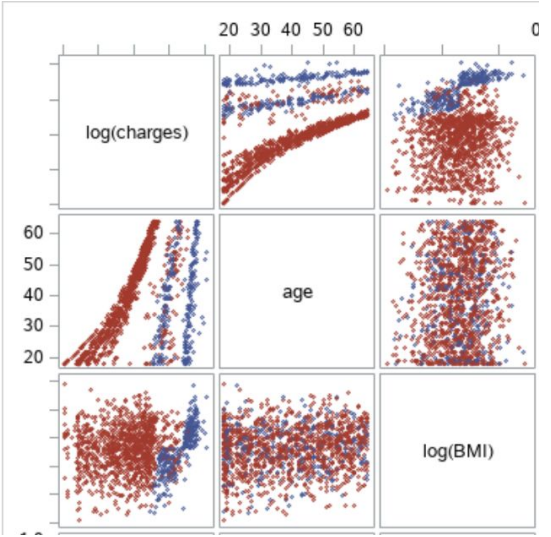
- Creation of “fat” variable
  - fat variable is a binary where X=0 when log(BMI)<mean(log(bmi)) and X=1 when log(BMI)≥mean(log(BMI)), and we found that mean(log(BMI))=3.40295
- Residual plot of log(charges) and age grouped by fat (top) displays improvements from residual plot of log(charges) and age grouped by smoker (bottom)

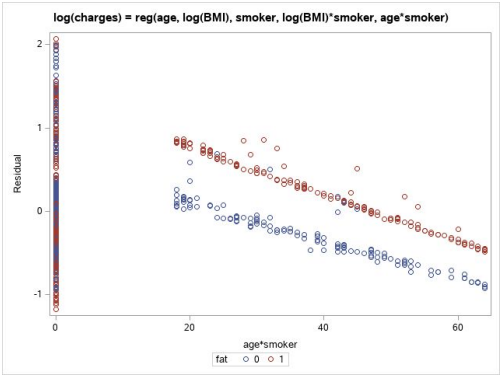




MLR: Model 5.5

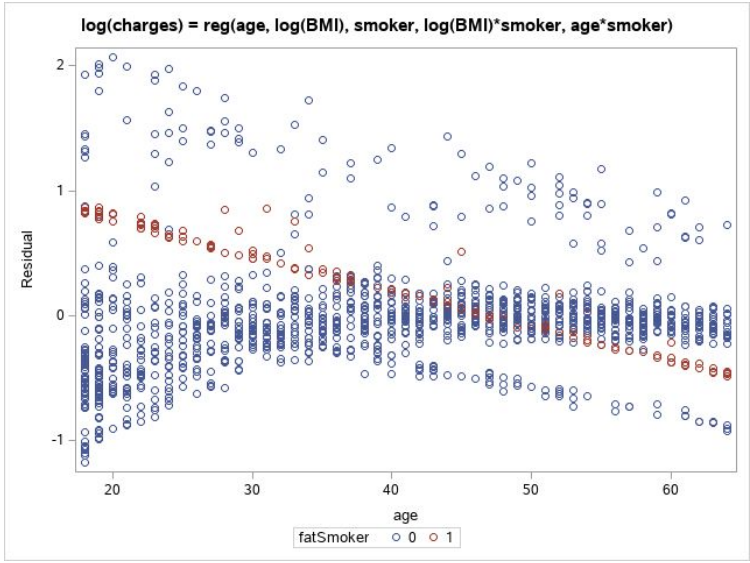
- By grouping the residual plot of log(charges) and age*smoker by fat (right), we are able to see the relationship that we wanted to find in the residual plot of log(charges) and log(BMI)*smoker
- The fat variable was responsible for creating this pattern
- Those with a log(BMI) below the mean amount are consistently grouped together in terms of charges and age when they are a smoker, and, similarly, those with a log(BMI) above the mean amount are consistently grouped together in terms of charges and age when they are a smoker





MLR: Model 5.5

- From here we created a new “fatSmoker” variable, which is fat*smoker
- Residual plot of log(charges) and age grouped by fatSmoker explains the model much more clearly
- When fatSmoker=1, there is a strong negative linear pattern with little deviation
- When fatSmoker=0, there is a slight upward pattern hovering around 0 with lots of scattering and variation, appearing more random
- This model is much more random and suits the MLR assumptions as best as possible



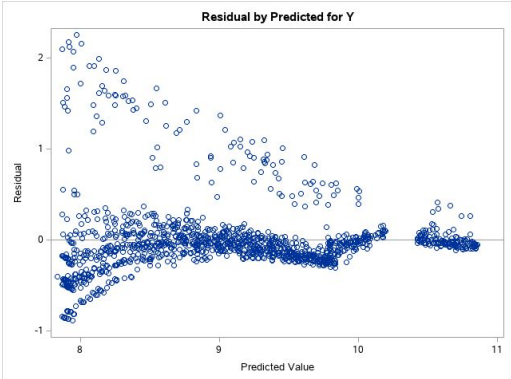
Continuation of Model 5.5

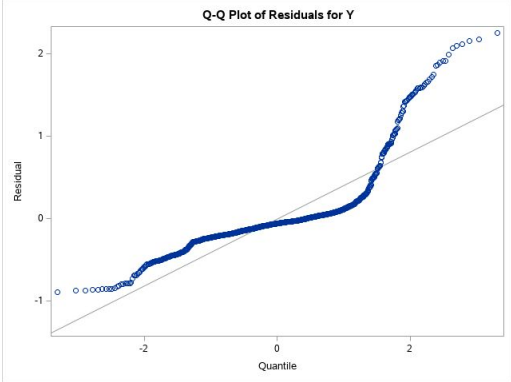
- With our variables of fat and fatSmoker, we now need to find a way to arrange these into an MLR model
- The models we attempted are to the right
- We deemed Models 5.5.2 and 5.5.4 adequate, as they have extremely similar results, but we will choose Model 5.5.2 since it has both variables fat and fatSmoker

Model Numbers	Model
5.5.1	Model 5 + fat
5.5.2	Age + log(BMI) + smoker + fat + fatSmoker + age*smoker
5.5.3	Model 5 + fat + fatSmoker
5.5.4	Age + log(BMI) + smoker + fatSmoker + age*smoker

Continuation of Model 5.5: Model 5.5.2

- The residual and Q-Q plot of Model 5.5.2 are not perfect, and do not meet an MLR completely
- The VIF values for each of the parameters within this model are below 10, and thus display no multicollinearity, which makes us believe that Model 5.5.2 is on the right track to finding the best suited model for our data





Goal Resolutions

- Determine how much weight each field has on individual charges
  - log(BMI), age, and smoker were found to be significant; sex, children and region are not significant
    - However, children treated as continuous (instead of categorical) was significant
- Find correlations between variables
  - log(charges), smoker and log(BMI)
  - log(charges), smoker and age
  - log(charges), fat and smoker
- Find the direct weight and influence that the number of children as dependents cause on an individual's insurance plan
  - The number of children was found to be quadratically related to charges

Goal Resolutions Continued

- Is there a linear relationship between number of dependents and overall charge?
  - The LOESS curve of children does not display a linear pattern, but rather a quadratic. The p-values show a significance of the quadratic term under 10% confidence level.
- Do charges increase at different age stages in life due to certain societal standards?
  - We did not find any supporting evidence of this, although we did find the three linear patterns within age for each group of health (healthy, moderate health, unhealthy)

Conclusions

- We conclude Model 5.5.2 for now, but note that in the future this Model needs to be adjusted to better suit our data
    - Making the children variable continuous and using a quadratic term as done in model 1.5 may be a good next step
- Limitation: This is real world data which makes it difficult to fit a perfect model since we can’t shape the data to fit our needs.
- Thus, we conclude our research on Medical Insurance Costs