Final Report: Multiple Regression Analysis on Variables that may Predict Health Insurance Charges

Liam Daly, Luis Colin Cornejo, Nicholas Khemvisay

2023-12-08

Contents

Introduction **5**
- Subject Matter . . . 5
- Research Question . . . 5
- Motivation . . . 5
- Hypothesis . . . 5
- Data Description and Defintion of Key Variables . . . 5
- Exploratory Data Analysis . . . 6
  - Checking the Normality Assumption . . . 6

Regression Analysis **6**
- Assumptions . . . 6
- Model Building . . . 6
- Final Model . . . 7
  - Final Model Assumptions . . . 7
  - Interpretation of Coefficients . . . 8
  - Checking for Multicollineairty . . . 8
  - Interpretation of the Coefficient of Determination, R^2 . . . 8
  - Interpretation of F Test . . . 8

Conclusion **9**

Limitations **9**

Appendix **12**
- Introduction . . . 12
  - Data Collection . . . 12
  - Supportive Material for Hypothesis . . . 12
  - Exploratory Data Analysis . . . 12
  - A1.1: Summary of Insurance Data . . . 12
  - A1.2: Predictor Plots . . . 12
  - A1.3: Response Plot . . . 13
  - A1.4: Predictors vs. Response Plots . . . 14
  - A1.5: Inital Full Model Output and Residual Plot . . . 15
  - A1.6: Multicollinearity Analysis I: ggpairs() . . . 15
  - A1.7: Multicollinearity Analysis I: Variance Inflation Factor . . . 17
- Regression Analysis . . . 17
  - A2.1: Finding Functional Form . . . 17
  - A2.2: Model Transformations . . . 24
  - A2.3: Considering Various Models . . . 30
  - A2.4: Stepwise Regression . . . 37
  - A2.5: Final Model Diagnostics . . . 41
  - A2.6: Final Model Output . . . 42
  - A2.7: Multicollinearity Analysis on Final Model: VIF . . . 44
- Future Work . . . 44

Works Cited **44**

List of Figures

1 Full Model Residual Plot 6
2 Final Model Assumption Checking Plots 7
3 Residuals v. Fitted Linear Patterns 9
4 Residuals v. Fitted with Color Coding Based on BMI 10
5 Residuals v. Fitted with Color Coding Based on Smoker 10

List of Tables

1 Data Summary Statistics 5

Introduction

Subject Matter

For this project, we are working with a health insurance dataset. Our main objective is to assess the impact of several predictor variables on the health insurance charges, the response variable. To accomplish this, we'll build a multiple regression model that can predict charges as an outcome.

Research Question

Which predictor variables in the health insurance data set are most important for influencing changes in insurance charges?

Motivation

Understanding the impact of several independent variables on insurance charges can have implications for healthcare resource planning. Different demographic groups may have distinct healthcare needs and utilization patterns. Hospitals, healthcare providers, and insurers can leverage this information to allocate resources efficiently, anticipate demand for specific services, and tailor preventive care programs for different groups (National Academies of Sciences)

Primarily, we are interested in determining which individual variables have the strongest linear relationship with with health insurance charges. Parties of interest or stakeholders can use the results of our analysis to make data-driven decisions based on the key variables that have the most influence on insurance charges.

Hypothesis

We believe that age, BMI, an individual's smoking habits, and number of children are the best predictors for insurance charges. Specifically, we hypothesize that all four variables and interactions of interest will demonstrate positive linear relationships with insurance charges. Additional explanation on our reasoning for this hypothesis can be found in the appendix.

Data Description and Defintion of Key Variables

```
ncol(insurance)
```

```
## [1] 7
```

```
nrow(insurance)
```

```
## [1] 1338
```

As you can see from the R output above, each row of the insurance data set contains seven variables. There are a total of 1338 rows, meaning data was collected from 1338 separate individuals. Details of the seven variables are listed below.

Variable	Type	Min	Max	Descr.	Mean	SD
Age x_1	Num. Con.	18	64	Identifies patient's age in discrete numbers.	39.21	14.05
Sex x_2	Categ.	0	1	Identifies patient's sex (male or female)	N/A	N/A
BMI x_3	Num. Con.	15.96	54.13	Identifies patient's Body Mass Index	30.66	6.10
Children x_4	Num. Disc.	0	5	Identifies patient's number of children	1.095	1.20
Smoker x_5	Categ.	0	1	Identifies patient's smoker status (yes or no)	N/A	N/A
Region x_6	Categ.	0	1	Identifies region where patient lives (4 possibilities)	N/A	N/A
Charges y	Num. Cont.	1122	63770	Patient's total health insurance charge	13270	12110.01

Table 1: Data Summary Statistics

Exploratory Data Analysis

By using various exploratory data analysis methods, we' hope we'll thoroughly check all variables in the full model below.

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

To explore the distributions and relationships within our data we first created predictor plots, response plots, and predictors vs. response plots. After this, we analyzed the output of our full model and tested for multicollinearity. Our key findings are below

- Predictor plots indicate that there are more younger individuals and non-smokers represented in the data set (Appendix A1.2).
- Right-skewed response plot shows that the majority of individuals have lower charges, with few cases having substantially higher charges (Appendix A1.3).
- Response v. predictor plots most notably indicated age, number of children, and smoker would likely be the strongest predictors (Appendix A1.4).
- Summary output from our full model showed that all predictor coefficients except sex were statistically significant. R^2 indicates that the model explains approximately 75.09% of the variability in insurance charges, therefore there's room for potential improvement in R^2 (Appendix A1.5).
- ggpairs(insurance) and vif(insurance) do not indicate the presence of multicollinearity (Appendix A1.6 & A1.7).

Checking the Normality Assumption

Lastly, below we'll check if the residuals of y are normally distributed

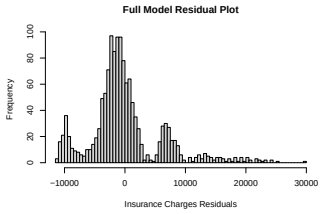


Figure 1: Full Model Residual Plot

Based on the figure above, the residuals violate normality.

Regression Analysis

Assumptions

Primarily, we aimed to satisfy the four assumptions below.

1. Normality: The distribution of the residuals are approximately normally distributed.
2. Constant Variance: The variability of the least squares line is constant.
3. Linearity: There is a linear relationship between response and predictor variables.
4. Independence: Residuals are independent from each other.

Model Building

To satisfy the normality assumption, we went through the process of finding the functional form of our x variables. We constructed multiple linear models to check if our numerical independent variables and interactions with categorical variables effectively contributed to normality and linearity.

After examining the outputs, we decided to keep age x_1, BMI x_3, and number of children x_4 in our model. We found that smoker also contributes to linear patterns, so we decided to keep smoker's x_5 interactions with all three numerical predictors as well (Appendix A2.1).

In spite of our model reductions normality still did not hold, so we found it necessary to use the transformation $log(y)$ in order to satisfy normality. We also decided to remove intercepts because an individual cannot generate insurance charges if all three numeric x variables equal zero (Appendix A2.2).

Next, we individually removed each independent variable and interactions one by one to determine if any removals would improve assumptions or the R^2 value (Appendix A2.3). After this, we decided to conduct stepwise regression in both directions to narrow down our candidates even further (Appendix A2.4).

Final Model

We decided on this as our final model:

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

The full summary output is found in Appendix A2.6.

Final Model Assumptions

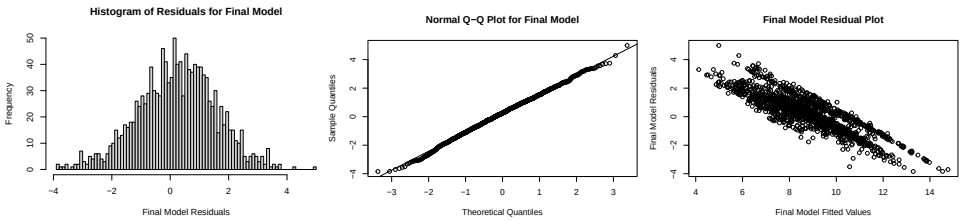


Figure 2: Final Model Assumption Checking Plots

```
shapiro.test(lm14$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm14$residuals
## W = 0.99843, p-value = 0.259
```

```
dwt(lm14)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.0148239      1.969075    0.62
##  Alternative hypothesis: rho != 0
```

1. Normality: There's a clear bell curve pattern on the residuals histogram, and the points on the normal Q-Q plot follow the line extremely closely. The high p-value from the Shapiro-Wilk test further proves that the normality assumption holds.
2. Constant Variance: The points seem evenly distributed on the y axis of the residuals v. fitted plot, so we conclude that constant variance holds.
3. Linearity: Strangely, there seems to be three parallel downward linear patterns on our residuals vs. fitted plot. This may indicate the presence of a hidden variable in the data set. For now, we'll assume that linearity is approximately satisfied but we'll discuss this more in our limitations.
4. The Durbin-Watson p value .356 > .05, so we'll conclude that the residuals are not correlated, therefore independent from each other.

Interpretation of Coefficients

Full output of the summary function on our final model is found in Appendix A2.6

- $x1$ (age): $\beta_1 = .075854$
  - A one-unit increase in age is associated with an average change of 100 * 0.075854 % in insurance charges (while holding other predictors constant). T test result p < 2e-16 indicates the existence of the positive linear association between x_1 and log(y) when all other independent variables are held fixed.
- $x3$ (BMI): $\beta_2 = .173248$
  - A one-unit increase in BMI is associated with a average change of 100 * 0.173248 % insurance charges (while holding other predictors constant). T test result p < 2e-16 indicates the existence of the positive linear association between x_3 and log(y) when all other independent variables are held fixed.
- $x4$ (children): $\beta_3 = .248180$
  - A one-unit increase in the number of children is associated with an average change of 100 * 0.248180% increase in insurance charges (while holding other predictors constant). T test result p = 1.23e-15 indicates the existence of the positive linear association between x_4 and log(y) when all other independent variables are held fixed.
- $(x_1)(x_5)$: β_4 (age * smoker): $\beta_4 =$ -.020711
  - The interaction term between smoker and age is associated with an average change of 100 * -0.020711% decrease in insurance charges (while holding other predictors constant). T test result p = .00543 indicates the existence of the slight negative linear association between $x_1 * x_5$ and log(y) when all other independent variables are held fixed.
- $(x_3)(x_5)$: β_5 (BMI * smoker): $\beta_5 = .075454$
  - The interaction term between smoker and BMI is associated with an average change of 100 * 0.075454% increase in insurance charges (while holding other predictors constant). T test result p < 2e-16 indicates the existence of the positive linear association between $x_3 * x_5$ and log(y) when all other independent variables are held fixed.

Age, BMI, children, and BMI * smoker contribute positively to the natural log of charges while age * smoker contributes negatively.

Checking for Multicollineairty

The statistical significance of all coefficient p-values indicates that multicollineaity is not present (Appendix A2.6). Furthermore, no variables display values greater than 10 in our VIF test, so there are no indications for multicollinearity in our final model (Appendix A2.7).

Interpretation of the Coefficient of Determination, R^2

Our R^2 value .9779 is extremely close to 1. This indicates that our model is an excellent fit for data. Furthermore, our model explains approximately 97.8% of the variance in the log of insurance charges (Appendix A2.6).

Interpretation of F Test

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$ (all model terms are unimportant for predicting log(y)

H_A: At least one model term is useful for predicting log(y)

The p-value for the F test statistic is < 2.2e-16. Given that all of our coefficient p-values were statistically significant, we can conclude that all of our model terms are useful for predicting log(y) (Appendix A2.6).

Conclusion

Our final model suggests that age, BMI, number of children, and their interactions with smoking status are the strongest predictors of health insurance charges. The high R-squared value indicates an excellent fit, and the coefficients provide insights into the directions and strengths of the relationships between predictors and charges.

Our initial hypothesis was somewhat correct. Age, BMI, number of children, and smoker turned out to be the best predictors for insurance charges. However, the crucial interactions were smoker with age and BMI respectively, since we removed the interaction between smoker and number of children due to insignificance.

All numerical predictors plus the interaction between BMI and smoker turned out to have positive slope coefficients. This indicates that higher values of the natural log of charges are associated with the categories below:

- older patients
- overweight patients
- patients with more children
- overweight smokers

We believe interaction between age and smoker's negative slope coefficient is attributed to the higher amount of younger individuals are represented in the insurance data set. Younger smokers are expected to generate lower charges than older smokers, due to the positive linear relationship between age and charges.

Based on our results, we recommend healthcare providers to anticipate higher insurance charges associated with services for older patients, overweight patients, those who have more children, and overweight smokers.

Limitations

While we believe our final model model is extremely accurate for predicting log(charges), there are a few limitations to be aware of.

Out of the 6 predictor variables only 4 hold some significance to predicting insurance charges. From experimenting with the model, we found that sex and region hold little to almost no significance to predicting the outcome insurance charges, so we quickly scrapped the use of these variables after experimenting more with the models to get our final (Appendix A2.1).

Upon analyzing the residuals v. fitted plot of our final model, we noticed a pattern that looked like three downward lines with different intercepts yet similar slopes.

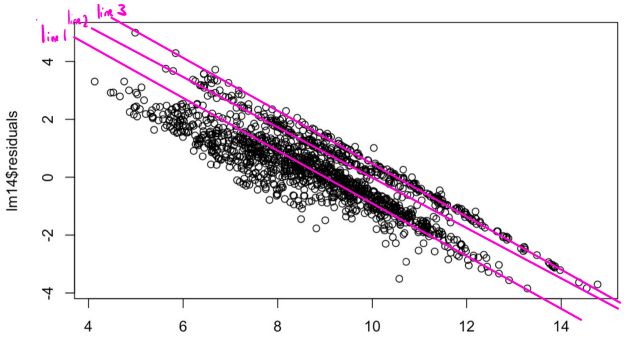


Figure 3: Residuals v. Fitted Linear Patterns

We thought bmi was causing this pattern somehow, so we decided to color code distinct bmi zones.

The CDC states a healthy BMI is between 18.5 and 24.9, and that an overweight BMI is above 29.9. We'll color code according to these observations (CDC).

BMI < 24.9 = green

BMI greater than 24.9 and less than 29.9 = black

BMI > 29.9 = red

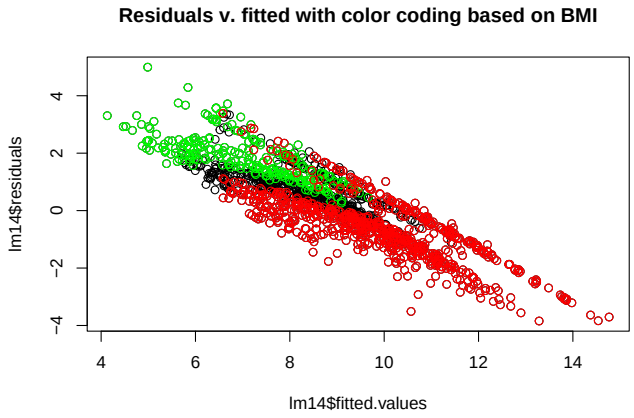


Figure 4: Residuals v. Fitted with Color Coding Based on BMI

As you can see, the red BMI may be able to partially explain lower line, but the middle and higher lines seem to be a mix off all zones. We also found it strange how the red zone is highly prevalent in the higher and lower lines, but not in the middle.

Next, we decided to color code the smoker variable, x_5.

red = an individual smokes

blue = an individual does not smoke

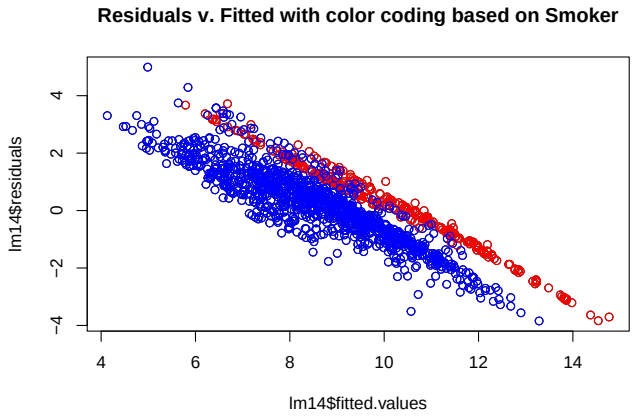


Figure 5: Residuals v. Fitted with Color Coding Based on Smoker

As you can see, "smoker = no" clearly explains the lower line, but the higher line seems to be a mix of "yes" and "no" points. We only know that the lower line is largely attributed to smoker = "no", but we still have no conclusive information about the middle and higher lines.

Throughout our entire analysis, we thoroughly analyzed all relationships between numerical and categorical independent variables. The pattern displayed above may suggest the existence of another categorical variable which may have been omitted from the insurance data set. Therefore, a hidden interaction between an independent categorical variable and our independent numerical variables may be present. Linearity may be satisfied on a residuals v. fitted plot that takes the hidden interaction into account. Given the data we worked with, we can assume that linearity is only approximately satisfied.

Our analysis is limited to the data that's present in insurance.csv, so we are unable to make any concrete conclusions about the unknown factor. We can only hypothesize that it seems to affect these two assumptions.

Appendix

Introduction

Data Collection

The original data was collected from a series of data sets found in the book "Machine Learning with R" by Brett Lantz. It is a simulated data set based on demographic statistics from the US Census Bureau (Lantz).

Supportive Material for Hypothesis

Data from the US Department of Health and Human Services shows that "out of pocket healthcare expenses for adults 65 and older rose 41% from 2009 to 2019" and according to a Gallup Article by Nicole Willcoxon, older individuals tend to experience an "increased demand for health services" (Wilcoxon).

Furthermore, data from the Medical Expenditure Panel Survey suggests that higher BMI and obesity is "associated with $1,861" excess annual medical costs per person (Ward et al.).

Additionally, a 2013 analysis of 2006-2010 Medical Expenditure Panel Survey indicated that "8.7% (95% CI=6.8%, 11.2%) of annual healthcare spending in the U.S. could be attributed to cigarette smoking." This percentage amounted to "as much as $170 billion per year" (Xu et al.).

Lastly, data extracted from the Institute of Health Metrics and Evaluation Disease Expenditure's 2013 Database shows an "increase in healthcare spending on children from $149.6 billion [in 1996] to $233.5 billion in 2013" (Bui et al).

Exploratory Data Analysis

A1.1: Summary of Insurance Data

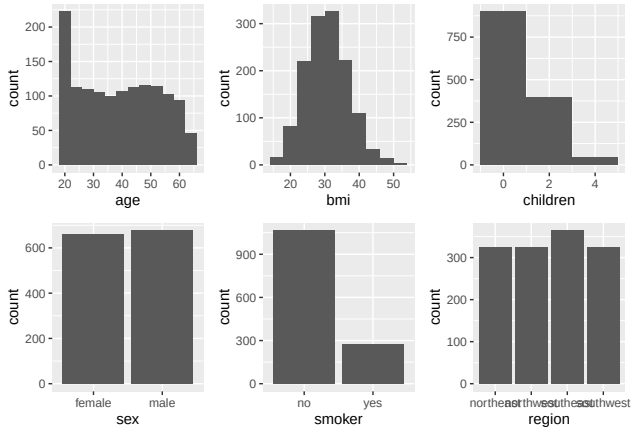
```
summary(insurance)
```

```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     smoker             region            charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

A1.2: Predictor Plots

```
# Histograms and barplots for predictors
# Create six plots and store in variable
h1<-ggplot(insurance, aes(x=age)) + geom_histogram(binwidth=4)
h2<-ggplot(insurance, aes(x=bmi)) + geom_histogram(binwidth=4)
h3<-ggplot(insurance, aes(x=children)) + geom_histogram(binwidth=2)
b1<-ggplot(data=insurance, aes(x=sex))+geom_bar()
b2<-ggplot(data=insurance, aes(x=smoker)) + geom_bar()
b3<-ggplot(data=insurance, aes(x=region)) + geom_bar()
```

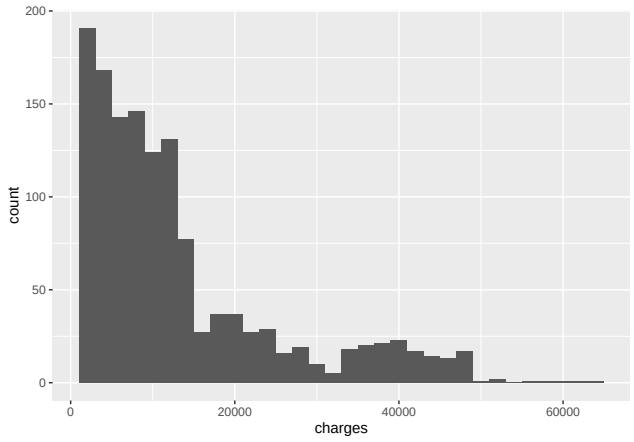
```
# Divide frame with grid.arrange function
# and put above created plots int it

grid.arrange(h1, h2,h3, b1,b2,b3, ncol = 3,nrow=2)
```



- The histogram for age indicates that the distribution is roughly skewed to the right. This means that there are more individuals with ages towards the lower end of the scale.
- The histogram for BMI appears to be approximately normally distributed. The shape suggests a balanced spread of BMI values without a significant skew. The histogram for the number of children shows a right-skewed distribution. Most individuals seem to have fewer children.
- The bar plot for sex shows that the counts of male and female individuals are relatively close, suggesting a balanced distribution between the sexes.
- The bar plot for smoker highlights a significant imbalance between smokers and non-smokers. There are notably more non-smokers than smokers in the dataset.
- Bar plot for region indicates a higher amount of individuals are from the southeast.

A1.3: Response Plot

```
ggplot(insurance, aes(x=charges)) + geom_histogram(binwidth=2000)
```

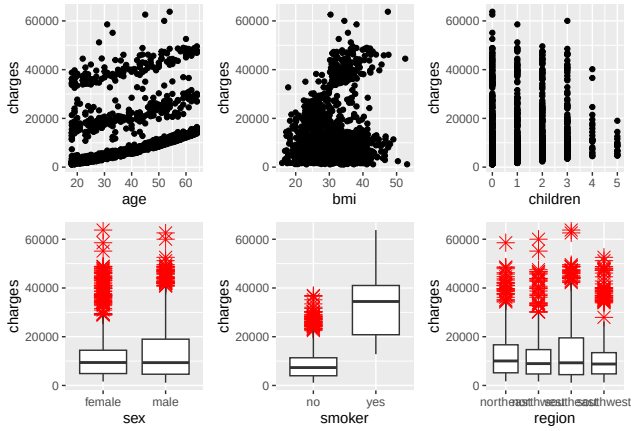


The right-skewed histogram of insurance charges above suggests that the majority of individuals have lower charges, with a few cases having substantially higher charges. This observation aligns with the typical

distribution of healthcare costs, where a relatively small proportion of individuals may incur significantly higher expenses.

A1.4: Predictors vs. Response Plots

```
# Create six plots and store in variable

# Scatter plots for numerical variables
s1<-ggplot(insurance, aes(x=age, y=charges)) + geom_point()
s2<-ggplot(insurance, aes(x=bmi, y=charges)) + geom_point()
s3<-ggplot(insurance, aes(x=children, y=charges)) + geom_point()

# Box plots for categorical variables
box1<-ggplot(insurance, aes(x=sex, y=charges)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=4)
box2<-ggplot(insurance, aes(x=smoker, y=charges)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=4)

box3<-ggplot(insurance, aes(x=region, y=charges)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=4)

# Divide frame in grid using grid.arrange function
# and put above created plot int it
grid.arrange(s1, s2,s3,box1, box2, box3, ncol = 3,nrow=2)
```



- Age vs. Charges (Scatter Plot): The scatter plot of age against insurance charges shows the distribution of charges across different age groups. From the graph, we can see that insurance charges tend to increase with age, suggesting a possible positive correlation.
- BMI vs. Charges (Scatter Plot): The scatter plot of BMI against insurance charges shows whether there is a relationship between an individual's BMI and their insurance charges. On this plot we don't see a clear pattern yet.
- Number of Children vs. Charges (Scatter Plot): The scatter plot of the number of children against insurance charges demonstrates whether there is an association between number of children covered by insurance and insurance charges. From this plot we see that as the number of children increases the charges of insurance seem to decrease, suggesting a possible negative correlation.
- Sex vs. Charges (Box Plot): This box plot compares insurance charges between genders. From this

plot, we see that there is some overlap between both genders and many outliers present, which might indicate gender to not be a great predictor for insurance charges.

- Smoker vs. Charges (Box Plot): This box plot examines the impact of smoking status on insurance charges. Here we see that there is no overlapping between the yes and no boxes, which might indicate that smoker is a good predictor for insurance charges.
- Region vs. Charges (Box Plot): This box plot examines the impact of region on charges. There's overlapping between all four categories, indicating region may not be a good predictor for charges.

A1.5: Inital Full Model Output and Residual Plot

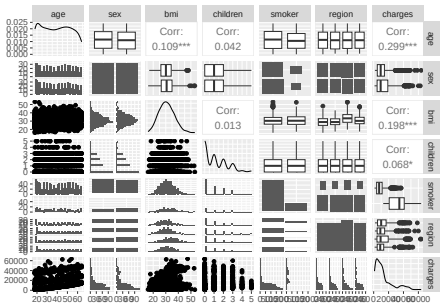
```
insurance.mlr<-lm(charges~age+sex+bmi+children+smoker+region,data=insurance)
summary(insurance.mlr)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

From our initial linear model we can identify the significant predictors. In the output of our model we see that the F test's p-value is statistically significant suggesting that at least some of the predictor variables are associated with the response variable. The significant predictors include age, BMI, children, and smoking status. Sex is not a great significant predictors. We also see that the model explains approximately 75.09% of the variability in insurance charges. The adjusted R-squared value of 0.7494 tells us that, considering all the predictors in the model, this value is very close to the multiple R-squared.

A1.6: Multicollinearity Analysis I: ggpairs()

```
ggpairs(insurance)
```



We used the ggpairs function, part of the GGally package, to create a matrix of plots to explore the pairwise relationships between variables. In our case, the variables include age, sex, BMI, children, smoker, and charges. Let's interpret the key information from this matrix:

- Smooth density curves: The BMI density curve tells us that the data is roughly normally distributed as we have seen before. The density curve for age tells us that the distributions of ages is fairly consistent with a skewness to the right. The children density curve has many drops possibly indicating that there are distinct groups or clusters within the data related to the number of children. Our last density curve for charges indicates again a big skewness to the right in the data as we have observed before.
- Scatter plots: Most scatter plots in our output tell us that the variables have noting to do with each other such as with age and children, and bmi and children since the points on the plot do not follow any pattern. On the other side, we do see some patterns (possibly positive linear relationships) with the predictors and charges (response) as expected.
- Histograms: Along the diagonal of the ggpairs plot, we see some individual histograms for each variable. The shape of each histogram reinforce the insights we discovered previously from the data such as their distribution and skewness. The histograms of individual predictors across the other different predictors are very consistent with their different categories having a similar spread.
- Bar plots for frequency: In our ggpairs plot we also see bar plots for our categorical data which give us insights on how these variables are distributed. Same as with the individual and predictor vs predictor histograms, our bar plots highlight the consitency of the data across the different categories.
- Box plots: The box plots for sex asnd smoker against age appear to show roughly similar distributions. This suggests that, based on age, there might not be significant differences between the groups represented by these categorical variables. We see the same thing in box plots against sex, BMI and children. In general, there may not be significant variations or differences in the distributions of our variables due to similar spread, medians and skewness and outliers.

Correlations: The correlation coefficients provide insights into the strength and direction of relationships between pairs of variables.

- BMI and Age (0.109***): The correlation coefficient of 0.109 indicates a positive but weak correlation between BMI and age. This suggests a slight tendency for individuals with higher ages to have higher BMI values.
- Children and Age (0.042): The correlation coefficient of 0.042 suggests a very weak positive correlation between the number of children and age. This correlation is not strong.
- Charges and Age (0.299***): The correlation coefficient of 0.299 indicates a moderate positive correlation between charges and age. This suggests that, on average, older individuals tend to have higher insurance charges.
- BMI and Charges (0.198***): The correlation coefficient of 0.198 indicates a moderate positive correlation between BMI and charges. This suggests that individuals with higher BMI values may, on average, have higher insurance charges.
- Children and Chargezs (0.068*): The correlation coefficient of 0.068 suggests a weak positive correlation between the number of children and charges. This correlation is not strong. I'ts important to rembember that correlation does not imply causation. Further analysis is needed to understand these relationships.

Given how none of the plots or correlation coefficients indicate notable high correlation, the ggpairs() function does not output any clear instances of multicollinearity.

A1.7: Multicollinearity Analysis I: Variance Inflation Factor

```
library(car)
vif(insurance.mlr)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## age      1.016822  1        1.008376
## sex      1.008900  1        1.004440
## bmi      1.106630  1        1.051965
## children 1.004011  1        1.002003
## smoker   1.012074  1        1.006019
## region   1.098893  3        1.015841
```

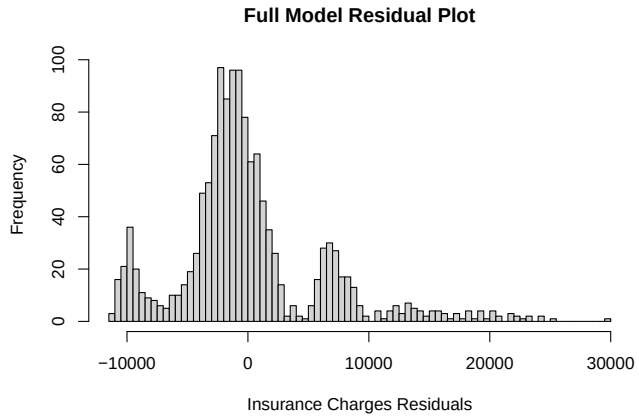
```
# No VIF parameters below are greater than 10
```

After running our variance inflation factor analysis we see that the VIF values for all predictor variables are well below the common threshold of 10, indicating a very low level of multicollinearity. This tells us that multicollinearity is not a serious concern in our full model.

Regression Analysis

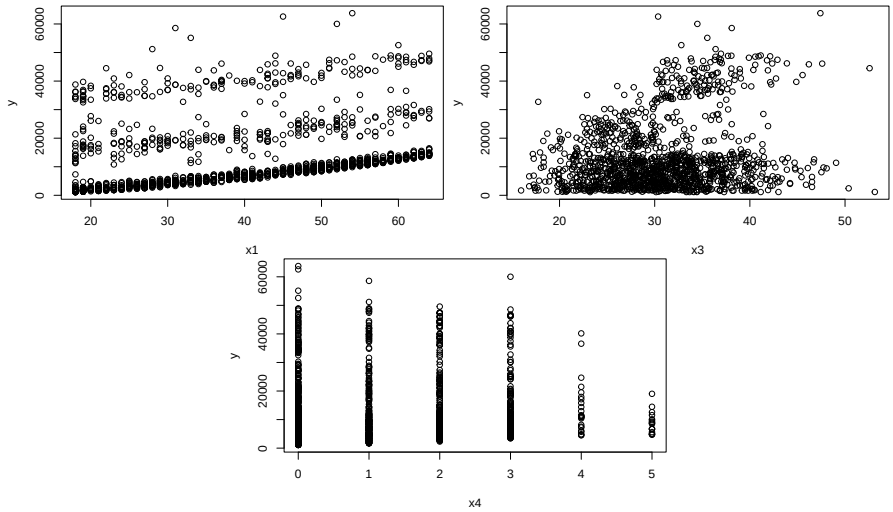
A2.1: Finding Functional Form

```
###Figure 1 Code
insurance.mlr<-lm(charges~age+sex+bmi+children+smoker+region,data=insurance)
hist(insurance.mlr$residuals, breaks = 90, xlab = "Insurance Charges Residuals", main =
       "Full Model Residual Plot")
```



First, we'll plot the numerical independent variables to examine their relationship with y:

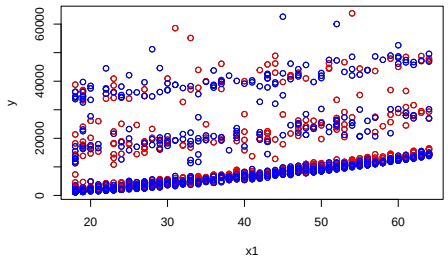
```
plot(x1, y)
plot(x3, y)
plot(x4, y)
```



Seeing as though x1 displays a clear positive linear relationship with the appearance of multiple y intercepts, we'll plot x1 by color according to each categorical variable.

Plot of Age and Sex against Charges

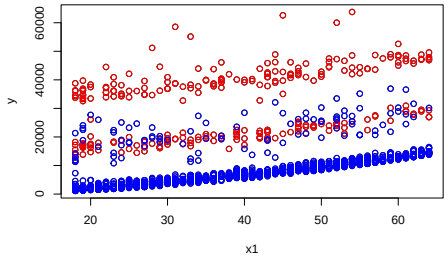
```
plot(x1, y)
points(x1[x2=="female"], y[x2=="female"], col = "red")
points(x1[x2=="male"], y[x2=="male"], col = "blue")
```



No discernable interaction pattern

Plot of Age and Smoker against Charges

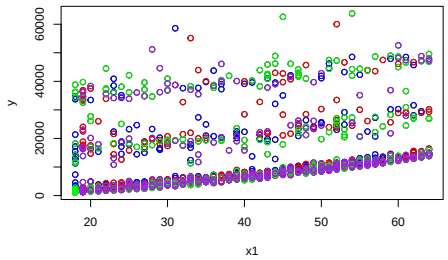
```
# categorical variable smoker and age against charges
plot(x1, y)
points(x1[x5=="yes"], y[x5=="yes"], col = "red")
points(x1[x5=="no"], y[x5=="no"], col = "blue")
# smokers seem to have a higher y intercept
```



We observe a notable interaction between smoker (x5) and age. Specifically, the intercept for smoker = "yes" on age and charges seem to be noticeably higher than for smoker = "no."

Plot of Age and Region against Charges

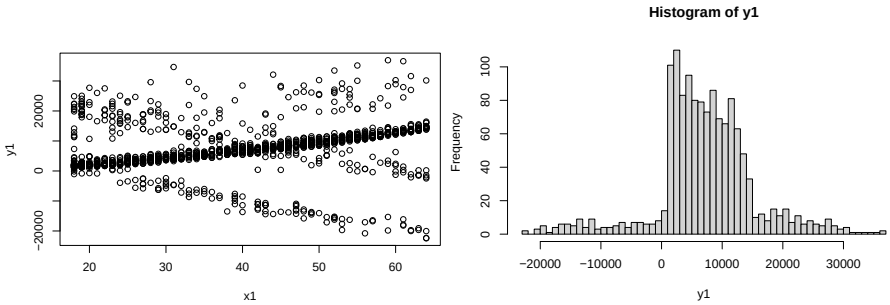
```
#region with x1
plot(x1, y)
points(x1[x6=="northeast"], y[x6=="northeast"], col = "blue")
points(x1[x6=="northwest"], y[x6=="northwest"], col = "red")
points(x1[x6=="southeast"], y[x6=="southeast"], col = "green")
points(x1[x6=="southwest"], y[x6=="southwest"], col = "purple")
#evenly distributed, no pattern
```



No discernable interaction pattern

Given how we noticed an interaction pattern between age and smoker against y, we will remove x1 and x5's interaction contribution and examine the relationship with the new y y1.

```
lm1<-lm(y~0+z1)
y1<-lm1$residuals
plot(x1,y1)
hist(y1, breaks = 80)
```

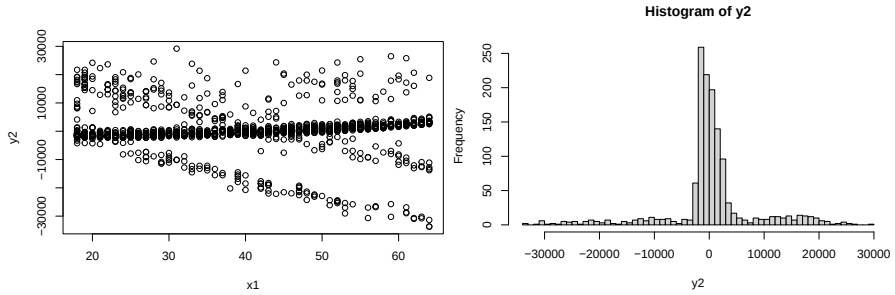


Now, the linear relationship between x1 and y1 seems slightly more evenly distributed without the interaction between age and smoker. The interaction between x1 and x5 contributes to a more normal y1, so we'll include z1 in our normal y model.

The histogram for y1 is more normal as well.

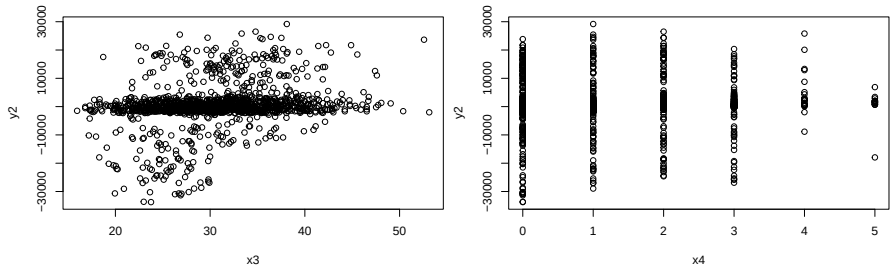
We'll now remove age x1 to examine the relationship between new y2 with the other numerical variables.

```
lm2<-lm(y1~0+x1)
y2<-y1-lm2$fitted.values
plot(x1, y2)
hist(y2, breaks = 80)
```



The histogram for y2 is noticeably more normal. There is a somewhat linear pattern present in x1 and y2, so we'll include x1 in the normal y model. Now we will examine the two remaining numerical independent variables.

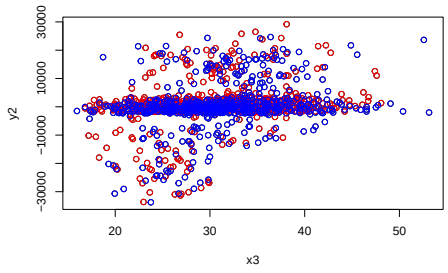
```
plot(x3, y2)
plot(x4, y2)
```



We observe a somewhat positive linear relationship in bmi x3 and y that seems to be influenced by a categorical variable.

Plot of bmi and sex against y2

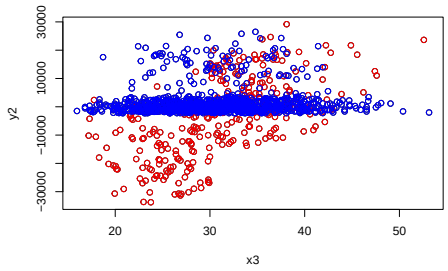
```
plot(x3, y2)
points(x3[x2=="female"], y2[x2=="female"], col = "red")
points(x3[x2=="male"], y2[x2=="male"], col = "blue")
```



Evenly distributed, no discernable pattern.

Plot of bmi and smoker against y2

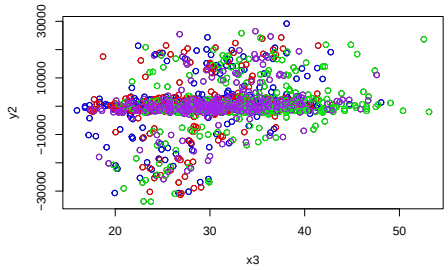
```
#interaction between bmi and smoker against y2
plot(x3, y2)
points(x3[x5=="yes"], y2[x5=="yes"], col = "red")
points(x3[x5=="no"], y2[x5=="no"], col = "blue")
```



Seems to be different intercepts and slopes for bmi depending on if an individual smokes or not

Plot of bmi and region against y2

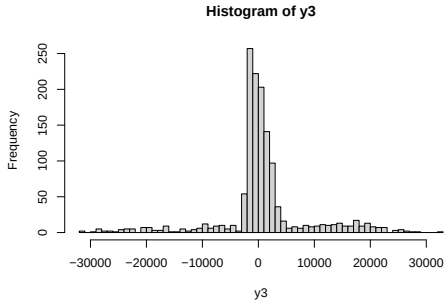
```
#interaction between bmi and region against y2
plot(x3, y2)
points(x3[x6=="northeast"], y2[x6=="northeast"], col = "blue")
points(x3[x6=="northwest"], y2[x6=="northwest"], col = "red")
points(x3[x6=="southeast"], y2[x6=="southeast"], col = "green")
points(x3[x6=="southwest"], y2[x6=="southwest"], col = "purple")
```



Evenly distributed, no discernable pattern.

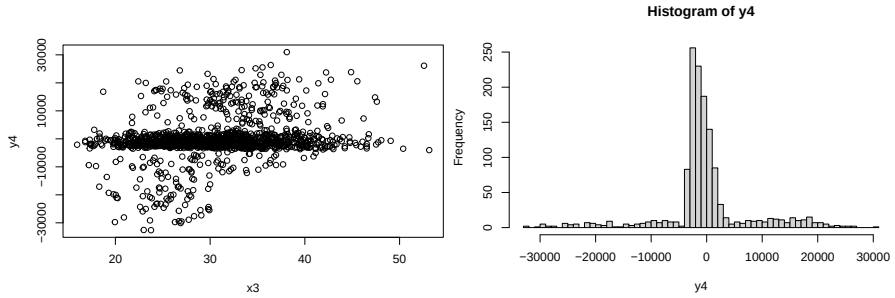
We observed a potential interaction between bmi x3 and smoker x5. Now, we will remove the interaction contribution to examine the new y3.

```
z2<-smoker * x3

lm3<-lm(y2~0+z2)
y3<-lm3$residuals
hist(y3, breaks = 80)
```



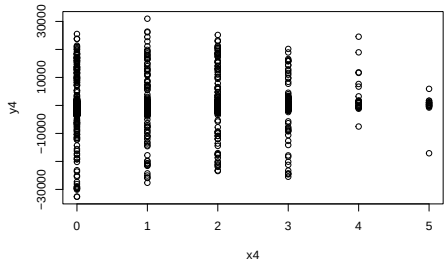
The interaction between x3 and x5 contributes to a more normal y5, so we'll include z2 in our normal y model.The y3 plot is still somewhat normal. The code below removes bmi. We will include z2 in out normal y model

```
lm4<-lm(y3~0+x3)
y4<-y3-lm4$fitted.values
plot(x3, y4)
hist(y4, breaks = 80)
```



The plot of x3 on y4 displays a somewhat positive linear pattern. Therefore, x3 will be included. Now, we'll examine the relationship between x4 and y4.

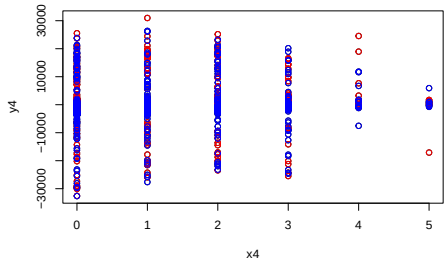
```
plot(x4, y4)
```



This looks evenly distributed.

Plot of Number of Children and Sex against y4

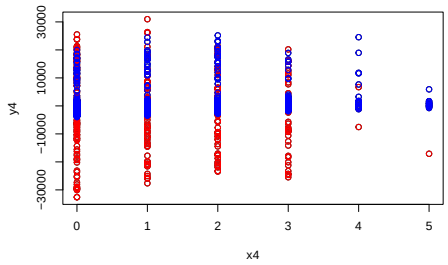
```
plot(x4, y4)
points(x4[x2=="female"], y4[x2=="female"], col = "red")
points(x4[x2=="male"], y4[x2=="male"], col = "blue")
```



No discernable pattern

Plot of Number of Children and Smoker against y4

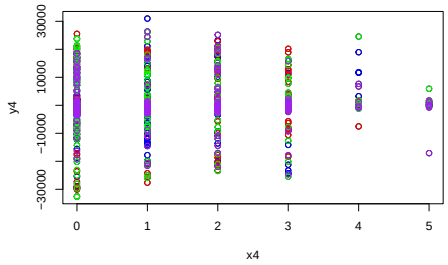
```
plot(x4, y4)
points(x4[x5=="yes"], y4[x5=="yes"], col = "red")
points(x4[x5=="no"], y4[x5=="no"], col = "blue")
# non-smokers seem to have a higher y intercept
```



Smokers seem to have a lower intercept. Below, we'll remove the interaction between x4 and x5.

Plot of Number of Children and Region against y4

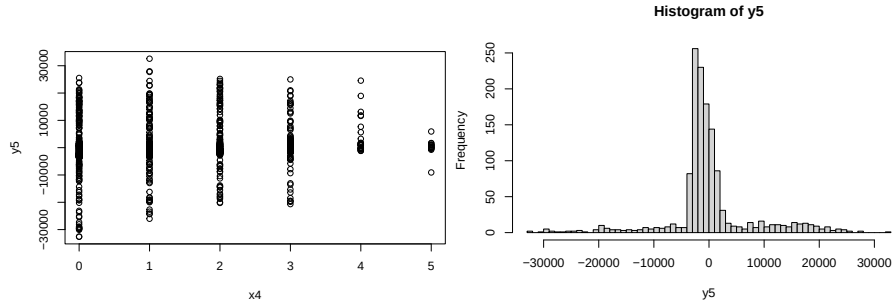
```
#interaction between Number of Children and Region against y4
plot(x4, y4)
points(x4[x6=="northeast"], y4[x6=="northeast"], col = "blue")
points(x4[x6=="northwest"], y4[x6=="northwest"], col = "red")
points(x4[x6=="southeast"], y4[x6=="southeast"], col = "green")
points(x4[x6=="southwest"], y4[x6=="southwest"], col = "purple")
```



No discernable pattern.

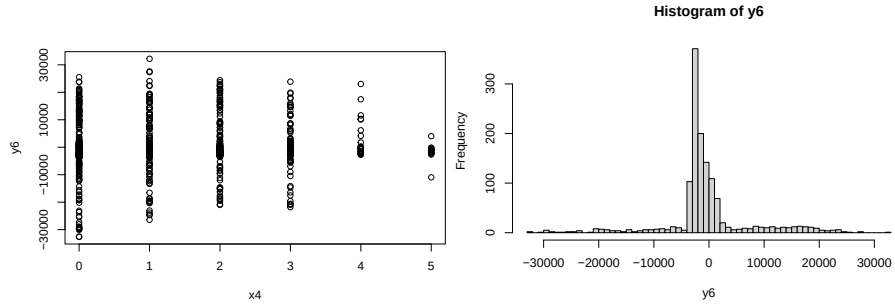
Removing the interaction between x4 and x5.

```
z3<-smoker * x4
lm5<-lm(y4~0+z3)
y5<-lm5$residuals
plot(x4, y5)
hist(y5, breaks = 80)
```



The interaction between x4 and x5 seems to contribute to a more normal y5, so we'll include z3 in our normal y model. Below, x4 will be removed so we can examine the final y6.

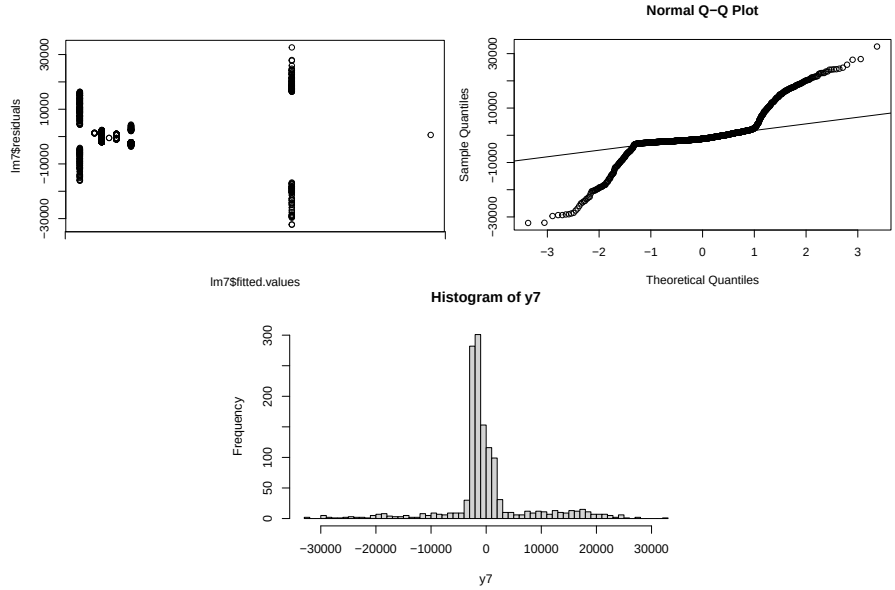
```
lm6<-lm(y5~0+x4)
y6<-y5-lm6$fitted.values
plot(x4, y6)
hist(y6, breaks = 80)
```



There seems to be a slight positive linear relationship between x4 and y5, so x4 will be included. We'll also include z3 in our normal linear model. Most importantly, y6 is still fairly normal. Next, we'll examine residuals when the intercept is removed.

```
lm7 <- lm(y6~1)
y7 <- lm7$residuals
plot(lm7$fitted.values, lm7$residuals)
qqnorm(y7)
qqline(y7)
hist(y7, breaks = 90)
shapiro.test(y7)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  y7
## W = 0.80163, p-value < 2.2e-16
```



7 fails the normality assumption. There is strange behavior on the residuals vs. fitted plot, which indicates potential mistakes. It may also necessitate transformations on y for normality.

Given these results, we'll keep x_1, x_3, x_4, z_1, z_2, and z_3 in consideration for our final model.

A2.2: Model Transformations

As displayed above, various steps of inclusion/exclusion of x variables contributes to a far more normal model compared to our initial full model. However, normality still doesn't hold.

Because of this, we'll have to transform our y variable in order to get an approximately normal model. We're choosing a log() transformation on y because our original response plot was right-skewed.

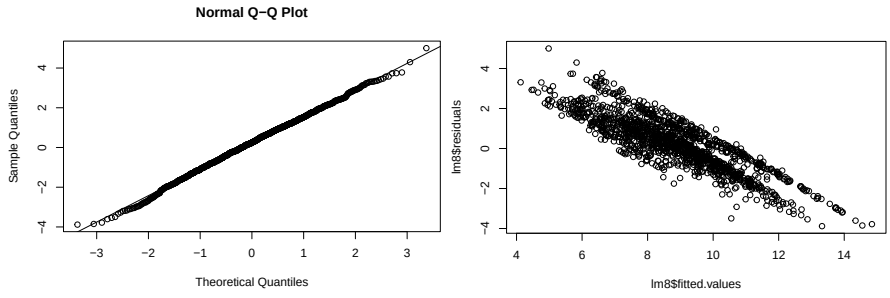
Additionally, we're choosing to remove the intercept because insurance charges (y) simply should be zero if all other numerical independent x variables are zero as well.

These two changes are the only way to achieve normality for charges against our x variables.

Following these transformations, we hope to observe normality in our output below.

```
lm8 <-lm(log(y)~0+ x1 + x3 + x4 + z1 + z2 + z3)
y8 <- lm8$residuals
qqnorm(lm8$residuals)
qqline(lm8$residuals)
plot(lm8$fitted.values, lm8$residuals)
summary(lm8)
```

```
##
## Call:
## lm(formula = log(y) ~ 0 + x1 + x3 + x4 + z1 + z2 + z3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8821 -0.6583  0.2450  1.1378  5.0004
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## x1  0.075775   0.002761  27.448  < 2e-16 ***
## x3  0.172964   0.003782  45.731  < 2e-16 ***
## x4  0.259314   0.034008   7.625 4.61e-14 ***
## z1 -0.020038   0.006040  -3.317 0.000933 ***
## z2  0.076689   0.008013   9.571  < 2e-16 ***
## z3 -0.059215   0.078426  -0.755 0.450356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.361 on 1332 degrees of freedom
## Multiple R-squared:  0.9779, Adjusted R-squared:  0.9778
## F-statistic:  9845 on 6 and 1332 DF,  p-value: < 2.2e-16
```




```
shapiro.test(lm8$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm8$residuals
## W = 0.99847, p-value = 0.2841
```

```
dwt(lm8)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01528498      1.968188   0.532
##  Alternative hypothesis: rho != 0
```

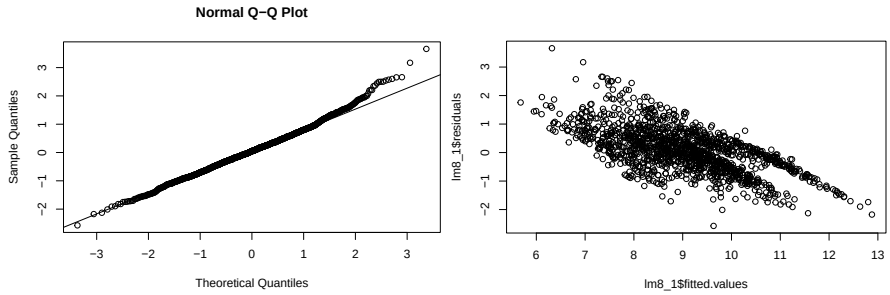
Thanks to log(y) and the removal of intercepts, linear model 8 is normal. The points look evenly distributed across the residual axis, therefore indicating constant variance. The residuals v. fitted plot implies independence and constant variance, but we'll preform model diagnostics to see if we can improve these assumptions. Lastly, the Durbin Watson p-value .542 indicates that autocorrelation is not present, therefore independence is satisfied.

Additionally, given the strange patterns we observed in bmi against charges in our functional form section, we're going to need to observe if transformations on x3 improve linearity and constant variance assumptions

First, we'll consider the square root of x3 as well as its interactions with smoker

```
z2_1 = sqrt(x3) * sqrt(smoker)
lm8_1 <-lm(log(y)~0+ x1 + sqrt(x3) + x4 + z1 + z2_1 + z3)
y8_1 <- lm8_1$residuals
qqnorm(lm8_1$residuals)
qqline(lm8_1$residuals)
plot(lm8_1$fitted.values, lm8_1$residuals)
summary(lm8_1)
```

```
##
## Call:
## lm(formula = log(y) ~ 0 + x1 + sqrt(x3) + x4 + z1 + z2_1 + z3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5739 -0.4472  0.0475  0.5516  3.6596
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## x1        0.048986   0.001722  28.443  < 2e-16 ***
## sqrt(x3)  1.200609   0.013442  89.317  < 2e-16 ***
## x4        0.162619   0.019940   8.155 7.96e-16 ***
## z1       -0.029704   0.003803  -7.810 1.15e-14 ***
## z2_1      0.507354   0.028894  17.559  < 2e-16 ***
## z3       -0.093262   0.045974  -2.029   0.0427 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7929 on 1332 degrees of freedom
## Multiple R-squared:  0.9925, Adjusted R-squared:  0.9925
## F-statistic: 2.944e+04 on 6 and 1332 DF,  p-value: < 2.2e-16
```




```
shapiro.test(lm8_1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm8_1$residuals
## W = 0.99254, p-value = 2.758e-06
```

```
dwt(lm8_1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.01898383      2.037811   0.518
##  Alternative hypothesis: rho != 0
```

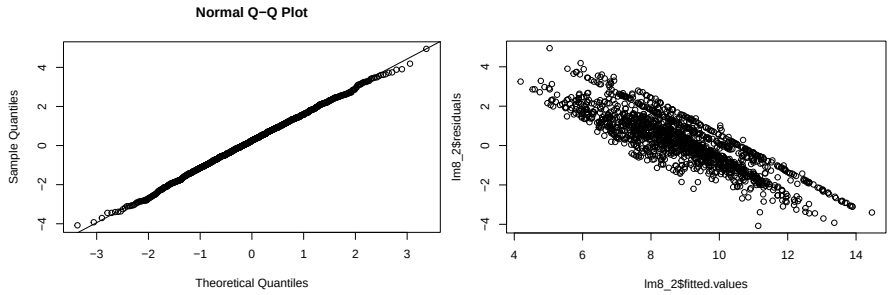
There is a slight improvement to the spread of our residuals, but normality is violated so we cannot keep this transformation.

Next, we'll consider squaring x3 and its interaction with smoker.

```
z2_2 = x3^2 * (smoker)^2
lm8_2 <-lm(log(y)~0+ x1 + (x3)^2 + x4 + z1 + z2_2 + z3)
y8_2 <- lm8_2$residuals
qqnorm(lm8_2$residuals)
qqline(lm8_2$residuals)
plot(lm8_2$fitted.values, lm8_2$residuals)
summary(lm8_2)
```

```
##
## Call:
## lm(formula = log(y) ~ 0 + x1 + (x3)^2 + x4 + z1 + z2_2 + z3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0793 -0.6975  0.2504  1.1848  4.9457
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## x1   0.0666910  0.0028119  23.717  < 2e-16 ***
## x3   0.1870509  0.0038373  48.745  < 2e-16 ***
## x4   0.2303019  0.0350848   6.564 7.48e-11 ***
## z1   0.0231519  0.0050052   4.626 4.10e-06 ***
## z2_2 0.0003314  0.0001872   1.771   0.0768 .
## z3   0.0727992  0.0801736   0.908   0.3640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.405 on 1332 degrees of freedom
```

```
## Multiple R-squared:  0.9765, Adjusted R-squared:  0.9764
## F-statistic:  9219 on 6 and 1332 DF,  p-value: < 2.2e-16
```




```
shapiro.test(lm8_2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm8_2$residuals
## W = 0.99904, p-value = 0.7261
```

```
dwt(lm8_2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01712449      1.962958   0.526
##  Alternative hypothesis: rho != 0
```

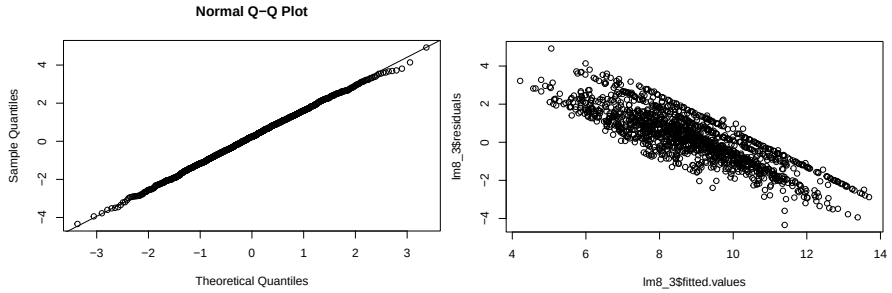
Again, there's slight improvement to the residuals v. fitted plot but normality is violated so we can't keep this transformation.

Next, we'll consider cubing x3 and its interactions

```
z2_3 = x3^3 * (smoker)^3
lm8_3 <-lm(log(y)~0+ x1 + (x3)^3 + x4 + z1 + z2_3 + z3)
y8_3 <- lm8_3$residuals
qqnorm(lm8_3$residuals)
qqline(lm8_3$residuals)
plot(lm8_3$fitted.values, lm8_3$residuals)
summary(lm8_3)
```

```
##
## Call:
## lm(formula = log(y) ~ 0 + x1 + (x3)^3 + x4 + z1 + z2_3 + z3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3401 -0.7036  0.2348  1.1742  4.9213
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## x1    6.264e-02  2.760e-03  22.693  < 2e-16 ***
## x3    1.933e-01  3.747e-03  51.596  < 2e-16 ***
## x4    2.174e-01  3.502e-02   6.207 7.21e-10 ***
## z1    3.635e-02  4.197e-03   8.662  < 2e-16 ***
## z2_3 -8.982e-06  4.063e-06  -2.210   0.0272 *
## z3    1.140e-01  7.975e-02   1.430   0.1530
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.405 on 1332 degrees of freedom
## Multiple R-squared:  0.9765, Adjusted R-squared:  0.9764
## F-statistic:  9232 on 6 and 1332 DF,  p-value: < 2.2e-16
```




```
shapiro.test(lm8_3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm8_3$residuals
## W = 0.99934, p-value = 0.9359
```

```
dwt(lm8_3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01681362      1.963181    0.47
##  Alternative hypothesis: rho != 0
```

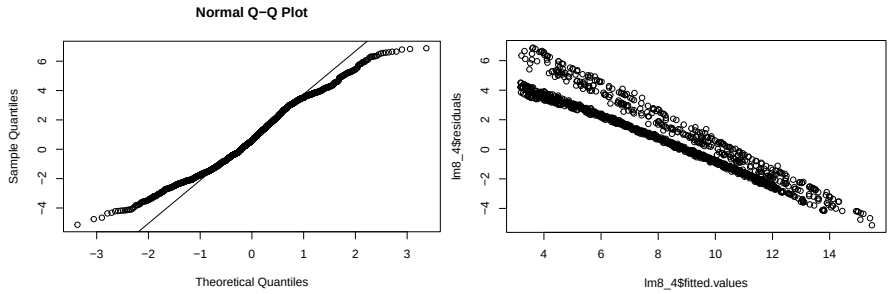
There aren't any noticeable improvements to our residuals v. fitted plot and normality is violated again.

Next, we'll try a cosine transformation

```
z2_4 = cos(x3) * cos(smoker)
lm8_4 <-lm(log(y)~0+ x1 + cos(x3) + x4 + z1 + z2_4 + z3)
y8_4 <- lm8_4$residuals
qqnorm(lm8_4$residuals)
qqline(lm8_4$residuals)
plot(lm8_4$fitted.values, lm8_4$residuals)
summary(lm8_4)
```

```
##
## Call:
## lm(formula = log(y) ~ 0 + x1 + cos(x3) + x4 + z1 + z2_4 + z3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1401 -1.1594  0.6066  2.7990  6.8809
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## x1       0.187365   0.002418  77.495  < 2e-16 ***
## cos(x3) -0.571470   0.485590  -1.177    0.239
## x4       0.615277   0.061869   9.945  < 2e-16 ***
## z1       0.031734   0.005627   5.640 2.08e-08 ***
## z2_4     0.382927   0.527087   0.726    0.468
```

```
## z3       0.098819   0.143961   0.686    0.493
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.544 on 1332 degrees of freedom
## Multiple R-squared:  0.923,  Adjusted R-squared:  0.9226
## F-statistic:  2660 on 6 and 1332 DF,  p-value: < 2.2e-16
```




```
shapiro.test(lm8_4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm8_4$residuals
## W = 0.98248, p-value = 1.166e-11
```

```
dwt(lm8_4)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1156335      1.764263       0
##  Alternative hypothesis: rho != 0
```

Cosine worsens the residuals v. fitted plot and violates normality, therefore we cannot keep this transformation.

For the sake of brevity, we've omitted various other transformations that did not improve linearity and constant variance. The patterns that are visible in the residuals v. fitted plot of model 8 will be discussed in our limitations section.

Now we'll move on with model 8 into our model diagnostics.

A2.3: Considering Various Models

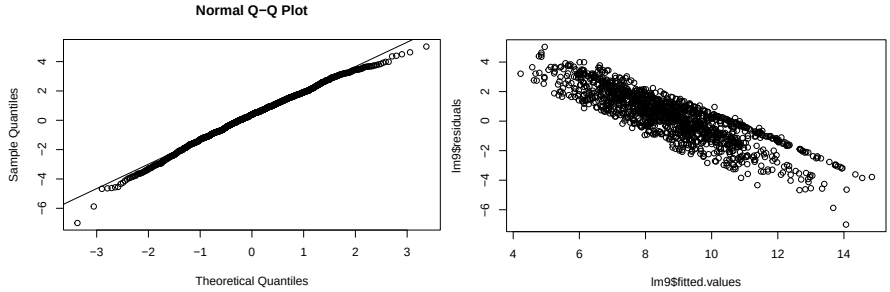
```
#removing x1
lm9 <-lm(log(y)~ 0 + x3 + x4 + z1 + z2 + z3)
qqnorm(lm9$residuals)
qqline(lm9$residuals)
plot(lm9$fitted.values, lm9$residuals)

summary(lm9)
```

```
##
## Call:
## lm(formula = log(y) ~ 0 + x3 + x4 + z1 + z2 + z3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0031 -0.7884  0.3802  1.4554  5.0268
```

```
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## x3  0.264675   0.002217 119.407  < 2e-16 ***
## x4  0.341433   0.042371   8.058 1.71e-15 ***
## z1  0.055737   0.006720   8.294 2.64e-16 ***
## z2 -0.015023   0.009109  -1.649   0.0993 .
## z3 -0.141334   0.098021  -1.442   0.1496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.702 on 1333 degrees of freedom
## Multiple R-squared:  0.9655, Adjusted R-squared:  0.9653
## F-statistic:  7455 on 5 and 1333 DF,  p-value: < 2.2e-16
```

```
dwt(lm9)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    -0.009617465      2.018441   0.736
##  Alternative hypothesis: rho != 0
```




```
shapiro.test(lm9$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm9$residuals
## W = 0.99432, p-value = 5.677e-05
```

P-value is much lower, normality no longer holds when x1 is removed. This shows us we definitely need to keep x1. We'll exclude model 9.

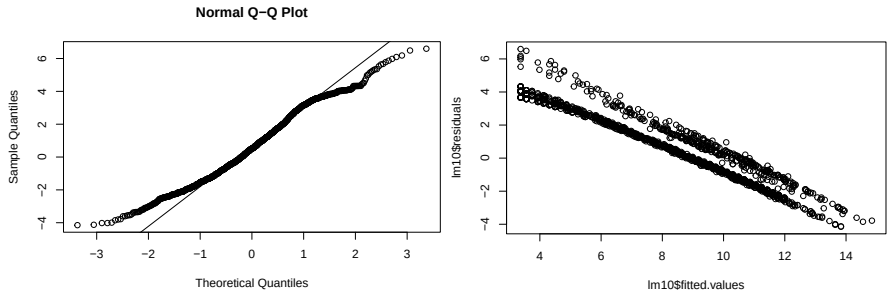
```
#removing x3
lm10 <-lm(log(y)~ 0 + x1 + x4 + z1 + z2 + z3)
qqnorm(lm10$residuals)
qqline(lm10$residuals)
plot(lm10$fitted.values, lm10$residuals)
shapiro.test(lm10$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm10$residuals
## W = 0.98268, p-value = 1.412e-11
```

```
summary(lm10)
```

```
##
```

```
## Call:
## lm(formula = log(y) ~ 0 + x1 + x4 + z1 + z2 + z3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1487 -1.0173  0.5157  2.2395  6.5970
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## x1  0.187308   0.002073  90.357  < 2e-16 ***
## x4  0.615539   0.053051  11.603  < 2e-16 ***
## z1 -0.131571   0.008856 -14.857  < 2e-16 ***
## z2  0.249653   0.011320  22.054  < 2e-16 ***
## z3 -0.415440   0.125059  -3.322 0.000918 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.181 on 1333 degrees of freedom
## Multiple R-squared:  0.9433, Adjusted R-squared:  0.9431
## F-statistic:  4437 on 5 and 1333 DF,  p-value: < 2.2e-16
```



Normality no longer holds when x3 is removed. Just like for x1, we must keep x3 for normality to hold. We'll exclude model 10.

```
#removing x4
lm11 <-lm(log(y)~ 0 + x1 + x3 + z1 + z2 + z3)
qqnorm(lm11$residuals)
qqline(lm11$residuals)
plot(lm11$fitted.values, lm11$residuals)
shapiro.test(lm11$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm11$residuals
## W = 0.99833, p-value = 0.214
```

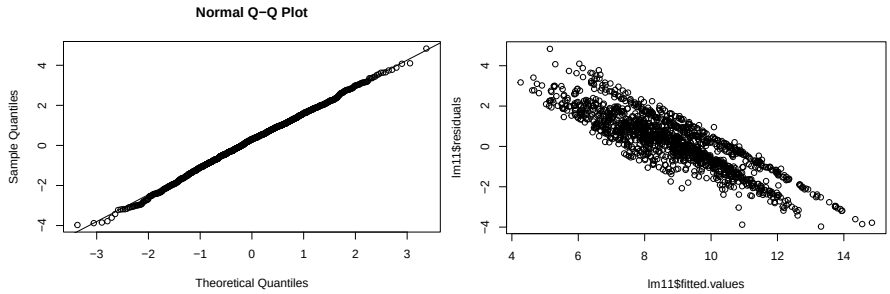
```
summary(lm11)
```

```
##
## Call:
## lm(formula = log(y) ~ 0 + x1 + x3 + z1 + z2 + z3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9723 -0.6778  0.3063  1.1335  4.8342
```

```
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## x1  0.077627   0.002808  27.642  < 2e-16 ***
## x3  0.179570   0.003760  47.761  < 2e-16 ***
## z1 -0.021890   0.006163  -3.552 0.000396 ***
## z2  0.070083   0.008135   8.615  < 2e-16 ***
## z3  0.200099   0.072167   2.773 0.005637 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.39 on 1333 degrees of freedom
## Multiple R-squared:  0.977,  Adjusted R-squared:  0.9769
## F-statistic: 1.132e+04 on 5 and 1333 DF,  p-value: < 2.2e-16
```

```
dwt(lm11)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.02354817      1.951713   0.362
##  Alternative hypothesis: rho != 0
```



Normality still holds. Excluding x4 ever so slightly increases the spread of points on residuals v. fitted. However, given how the R^2 value is slightly worse than the R^2 for model 8, we won't exclude x4. Linearity and constant variance still hold according to the residuals v. fitted plot. The Durbin-Watson p value .356 $>$ .05, so we'll conclude that the residuals are not correlated, therefore independent from each other. The assumptions still hold, but we prefer a model with a higher R^2. For now, we'll keep x4 in the model and consider other models to see if R^2 improves.

```
#removing z1
lm12 <-lm(log(y)~ 0 + x1 + x3 + x4 + z2 + z3)
qqnorm(lm12$residuals)
qqline(lm12$residuals)
plot(lm12$fitted.values, lm12$residuals)


shapiro.test(lm12$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm12$residuals
## W = 0.9984, p-value = 0.2461
```

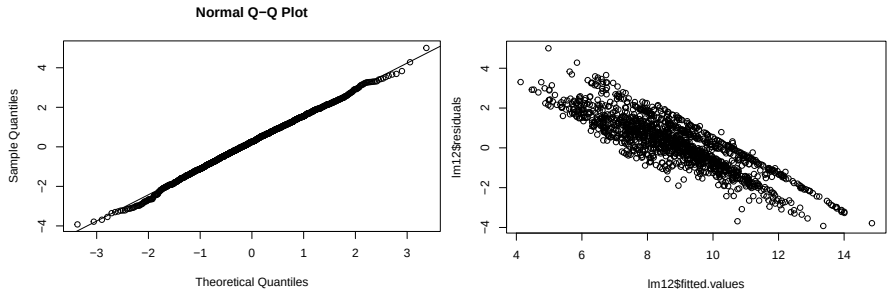
```
summary(lm12)
```

```
##
## Call:
```

```
## lm(formula = log(y) ~ 0 + x1 + x3 + x4 + z2 + z3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9239 -0.6580  0.2513  1.1387  5.0019
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## x1  0.071589   0.002465  29.046  < 2e-16 ***
## x3  0.178030   0.003473  51.259  < 2e-16 ***
## x4  0.263850   0.034108   7.736 2.02e-14 ***
## z2  0.053635   0.004003  13.398  < 2e-16 ***
## z3 -0.097584   0.077859  -1.253     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.366 on 1333 degrees of freedom
## Multiple R-squared:  0.9778, Adjusted R-squared:  0.9777
## F-statistic: 1.172e+04 on 5 and 1333 DF,  p-value: < 2.2e-16
```

```
dwt(lm12)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01672732      1.964807    0.55
##  Alternative hypothesis: rho != 0
```



Excluding z1 has no noticeable effect on residuals vs. fitted, but the R^2 is slightly reduced compared to model 8. Linearity and constant variance still hold according to the residuals v. fitted plot. The Durbin-Watson p value .498 > .05, so we'll conclude that residuals are not correlated, therefore independent from each other. Given R^2 is slightly lessened when z1 is removed, we'll proceed with keeping z1.

```
#removing z2
lm13 <-lm(log(y)~ 0 + x1 + x3 + x4 + z1+ z3)
y13 = lm13$residuals
qqnorm(lm13$residuals)
qqline(lm13$residuals)
plot(lm13$fitted.values, lm13$residuals)

shapiro.test(lm13$residuals)
```

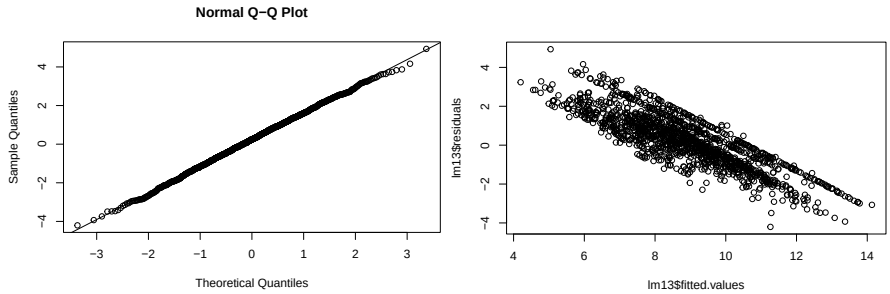
```
##
##  Shapiro-Wilk normality test
##
## data:  lm13$residuals
## W = 0.99927, p-value = 0.9015
```

```
summary(lm13)
```

```
##
## Call:
## lm(formula = log(y) ~ 0 + x1 + x3 + x4 + z1 + z3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2039 -0.6939  0.2500  1.1681  4.9340
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## x1 0.064756   0.002593  24.973  < 2e-16 ***
## x3 0.190051   0.003446  55.155  < 2e-16 ***
## x4 0.224123   0.034939   6.415 1.96e-10 ***
## z1 0.030104   0.003107   9.689  < 2e-16 ***
## z3 0.093955   0.079342   1.184    0.237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.407 on 1333 degrees of freedom
## Multiple R-squared:  0.9764, Adjusted R-squared:  0.9763
## F-statistic: 1.104e+04 on 5 and 1333 DF,  p-value: < 2.2e-16
```

```
dwt(lm13)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01687762       1.96322    0.52
##  Alternative hypothesis: rho != 0
```



Excluding z2 has no effect on residuals v. fitted or normality, but R^2 is further reduced compared to model 8. Linearity and constant variance still hold according to the residuals v. fitted plot. The Durbin-Watson p value .488 > .05, so we'll conclude that the residuals are not correlated, therefore independent from each other. Given R^2 is slightly lessened when z2 is removed, we'll proceed with keeping z2.

```
#removing z3
lm14 <-lm(log(y)~ 0 + x1 + x3 + x4 + z1 + z2)

y14 <- lm14$residuals
qqnorm(lm14$residuals)
qqline(lm14$residuals)
plot(lm14$fitted.values, lm14$residuals)
shapiro.test(lm14$residuals)
```

```
##
##  Shapiro-Wilk normality test
```

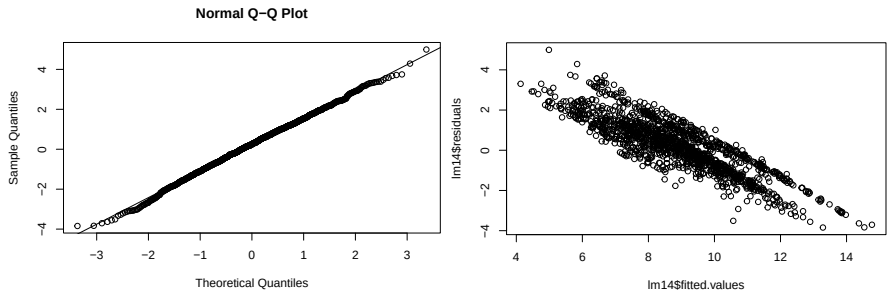
```
##
## data:  lm14$residuals
## W = 0.99843, p-value = 0.259
```

```
summary(lm14)
```

```
##
## 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
```

```
dwt(lm14)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.0148239      1.969075   0.594
##  Alternative hypothesis: rho != 0
```




```
shapiro.test(lm14$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm14$residuals
## W = 0.99843, p-value = 0.259
```

Removing z3 does not make any changes to the residuals vs. fitted plot, but the R^2 value is the exact same as model 8. Additionally, the adjusted R^2 value is higher than for model 8. Since the exclusion of z3 has a completely neutral effect on the coefficient of determination while normality holds, we can safely exclude it. Without the interaction between number of children (x4) and smoker (x5), the model still fits the data equally as well. Therefore, since z3 does not contribute positively to R^2, we can safely exclude it.

Linearity and constant variance still hold according to the residuals v. fitted plot. The Durbin-Watson p value .572 > .05, so we'll conclude that the residuals are not correlated, therefore independent from each other.

Considered But Excluded Models

Given the results of our testing above, we're heavily considering models 8 and 14, and we're excluding models 9 through 13.

Also, none of these models demonstrate significant improvements over model 8. To more thoroughly determine our final model, we'll build a stepwise regression model.

A2.4: Stepwise Regression

```
#null model with no predictors
stepwise_null_model = lm(log(y) ~ 1)

#full model with all relevant predictors
stepwise_full_model <- lm(log(y)~0+ x1 + x3 + x4 + z1 + z2 + z3)
```

```
stepwise_model1 <- step(stepwise_null_model, scope = list(lower = stepwise_null_model, upper =
stepwise_full_model), direction = "both", test="F")
```

```
## Start:  AIC=-223.51
## log(y) ~ 1
##
##        Df Sum of Sq     RSS      AIC  F value    Pr(>F)
## + z2    1    527.32  603.15 -1062.07 1168.045 < 2.2e-16 ***
## + z1    1    468.11  662.36  -936.77  944.192 < 2.2e-16 ***
## + x1    1    314.96  815.51  -658.46  515.977 < 2.2e-16 ***
## + z3    1    218.46  912.01  -508.82  320.020 < 2.2e-16 ***
## + x4    1     29.43 1101.05  -256.79   35.705 2.941e-09 ***
## + x3    1     19.90 1110.58  -245.27   23.936 1.117e-06 ***
## <none>              1130.47  -223.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-1062.07
## log(y) ~ z2
##
##        Df Sum of Sq     RSS      AIC   F value    Pr(>F)
## + x1    1    330.23  272.92 -2121.11 1615.3736 < 2.2e-16 ***
## + x4    1     27.89  575.26 -1123.42   64.7252 1.885e-15 ***
## + x3    1      3.90  599.25 -1068.76    8.6926  0.003251 **
## + z1    1      3.45  599.70 -1067.75    7.6802  0.005660 **
## <none>              603.15 -1062.07
## + z3    1      0.10  603.05 -1060.29    0.2196  0.639428
## - z2    1    527.32 1130.47  -223.51 1168.0454 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2121.11
## log(y) ~ z2 + x1
##
##        Df Sum of Sq    RSS      AIC   F value Pr(>F)
## + z1    1     31.30 241.62 -2282.09  172.8066 <2e-16 ***
```

```
## + x4    1     20.35 252.56 -2222.80  107.4917 <2e-16 ***
## <none>              272.92 -2121.11
## + z3    1      0.05 272.87 -2119.35    0.2427 0.6224
## + x3    1      0.00 272.91 -2119.13    0.0173 0.8953
## - x1    1    330.23 603.15 -1062.07 1615.3736 <2e-16 ***
## - z2    1    542.60 815.51  -658.46 2654.1796 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2282.09
## log(y) ~ z2 + x1 + z1
##
##        Df Sum of Sq    RSS     AIC   F value    Pr(>F)
## + x4    1     21.39 220.23 -2404.1  129.4690 < 2.2e-16 ***
## + x3    1      2.65 238.96 -2294.9   14.8042 0.0001249 ***
## + z3    1      0.49 241.13 -2282.8    2.6877 0.1013613
## <none>              241.62 -2282.1
## - z1    1     31.30 272.92 -2121.1  172.8066 < 2.2e-16 ***
## - z2    1    202.58 444.20 -1469.4 1118.4825 < 2.2e-16 ***
## - x1    1    358.08 599.70 -1067.8 1977.0210 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2404.12
## log(y) ~ z2 + x1 + z1 + x4
##
##        Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + x3    1      2.86 217.37 -2419.6   17.524 3.023e-05 ***
## + z3    1      2.01 218.21 -2414.4   12.295 0.0004694 ***
## <none>              220.23 -2404.1
## - x4    1     21.39 241.62 -2282.1  129.469 < 2.2e-16 ***
## - z1    1     32.34 252.56 -2222.8  195.734 < 2.2e-16 ***
## - z2    1    204.62 424.85 -1527.0 1238.522 < 2.2e-16 ***
## - x1    1    352.23 572.46 -1128.0 2131.998 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2419.6
## log(y) ~ z2 + x1 + z1 + x4 + x3
##
##        Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + z3    1      2.36 215.01 -2432.2   14.607 0.0001386 ***
## <none>              217.37 -2419.6
## - x3    1      2.86 220.23 -2404.1   17.524 3.023e-05 ***
## - x4    1     21.60 238.96 -2294.9  132.335 < 2.2e-16 ***
## - z1    1     35.19 252.56 -2220.8  215.632 < 2.2e-16 ***
## - z2    1    201.04 418.41 -1545.4 1231.954 < 2.2e-16 ***
## - x1    1    349.79 567.16 -1138.4 2143.467 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2432.21
## log(y) ~ z2 + x1 + z1 + x4 + x3 + z3
##
```

```
##        Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## <none>              215.01 -2432.2
## - z3    1      2.36 217.37 -2419.6   14.607 0.0001386 ***
## - x3    1      3.21 218.21 -2414.4   19.841 9.121e-06 ***
## - x4    1     23.58 238.59 -2294.9  145.999 < 2.2e-16 ***
## - z1    1     31.82 246.83 -2249.5  197.010 < 2.2e-16 ***
## - z2    1    201.46 416.47 -1549.6 1247.144 < 2.2e-16 ***
## - x1    1    346.98 561.99 -1148.7 2147.970 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#forward selection using null model
summary(stepwise_model1)
```

```
##
## Call:
## lm(formula = log(y) ~ z2 + x1 + z1 + x4 + x3 + z3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95283 -0.14934 -0.06793  0.02344  2.26369
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.3375693  0.0621366 118.088  < 2e-16 ***
## z2           0.0835824  0.0023668  35.315  < 2e-16 ***
## x1           0.0402696  0.0008689  46.346  < 2e-16 ***
## z1          -0.0250427  0.0017842 -14.036  < 2e-16 ***
## x4           0.1221502  0.0101093  12.083  < 2e-16 ***
## x3          -0.0084606  0.0018994  -4.454 9.12e-06 ***
## z3          -0.0885135  0.0231598  -3.822 0.000139 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4019 on 1331 degrees of freedom
## Multiple R-squared:  0.8098, Adjusted R-squared:  0.8089
## F-statistic: 944.5 on 6 and 1331 DF,  p-value: < 2.2e-16
```

When intercepts are included and variables are added one-by-one, forward selection decides to keep all variables. There may be complications to consider since forward selection includes intercepts, so we'll have to closely examine the results of backward selection.

The last variable to be included was z3, so it may be a likely candidate for removal once we conduct backwards selection using the full model.

```
#backwards selection using full model
stepwise_model2 <- step(stepwise_full_model, scope = list(lower = stepwise_null_model, upper =
stepwise_full_model), direction = "both", test = "F")
```

```
## Start:  AIC=830.96
## log(y) ~ 0 + x1 + x3 + x4 + z1 + z2 + z3
##
##        Df Sum of Sq    RSS     AIC   F value    Pr(>F)
## - z3    1       1.1 2468.7  829.53    0.5701 0.4503557
## <none>              2467.6  830.96
## - z1    1      20.4 2488.0  839.97   11.0049 0.0009332 ***
## - x4    1     107.7 2575.3  886.12   58.1419 4.615e-14 ***
```

```
## - z2    1     169.7 2637.3  917.95   91.6042 < 2.2e-16 ***
## - x1    1    1395.7 3863.3 1428.74  753.3795 < 2.2e-16 ***
## - x3    1    3874.3 6342.0 2091.94 2091.3407 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=829.53
## log(y) ~ x1 + x3 + x4 + z1 + z2 - 1
##
##        Df Sum of Sq    RSS     AIC   F value    Pr(>F)
## <none>              2468.7  829.53
## + z3    1       1.1 2467.6  830.96    0.5701 0.4503557
## - z1    1      22.3 2490.9  839.54   12.0213 0.0005427 ***
## - x4    1     121.5 2590.2  891.82   65.6104  1.23e-15 ***
## - z2    1     171.4 2640.1  917.36   92.5616 < 2.2e-16 ***
## - x1    1    1400.7 3869.3 1428.83  756.3057 < 2.2e-16 ***
## - x3    1    3925.8 6394.5 2100.98 2119.7945 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#backwards selection using full model
summary(stepwise_model2)
```

```
##
## Call:
## lm(formula = log(y) ~ x1 + x3 + x4 + z1 + z2 - 1)
##
## 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
```

In backwards selection using the full model, z3 was removed. Based on forward selection, we've determined that z3 doesn't contribute significantly to the improvement of R^2. Because of this, we've decided to remove z3 from our final model.

```
summary(stepwise_model2)
```

```
##
## Call:
## lm(formula = log(y) ~ x1 + x3 + x4 + z1 + z2 - 1)
##
## 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
```

```
summary(lm14)
```

```
##
## 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
```

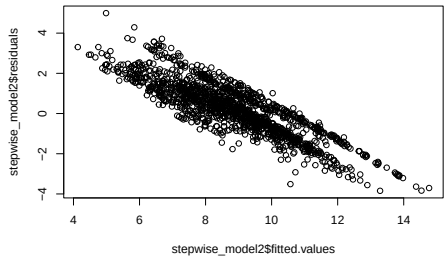
Notice how stepwise_model_2 is identical to model 14. This further supports our decision to move forward using model 14.

A2.5: Final Model Diagnostics

```
shapiro.test(lm14$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lm14$residuals
## W = 0.99843, p-value = 0.259
```

```
plot(stepwise_model2$fitted.values, stepwise_model2$residuals)
```



The Shapiro-Wilk normality test's p-value of .259 indicates that the normality assumption is satisfied. Upon examining the residuals v. fitted plot, this model approximately satisfies linearity and constant variance with complications worth considering. We'll use model 14 as our final model but we'll discuss these complications in our limitations section.

Additionally, this model satisfies the normality and independence assumptions according to the tests we ran before stepwise regression. All necessary assumptions are satisfied, and the R^2 is very high, so we've decided to select model 14 as our final model.

A2.6: Final Model Output

```
summary(lm14)
```

```
##
## 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
```

Our final multiple linear regression model is of the form:

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

This multiple linear regression model aims to analyze the impact of several predictor variables on health insurance charges. The model uses the logarithm of the response variable (y), since the data for insurance charges was not satisfying the normality assumption. The model includes the predictor variables x1, x3, x4, z1, and z2. The coefficients represent the estimated change in insurance charges for a one-unit change in each predictor.

After fitting our model:

$$log(\hat{y}) = 0 + 0.075854x_1 + 0.0173248x_3 + 0.248180x_4 - 0.020711x_1x_5 + 0.075454x_3x_5 + \epsilon$$

The interaction between age and smoker has a slightly negative coefficient, but this isn't entirely unexpected. Younger smokers may be attributed with lower insurance charges, and we initially observed there were many more younger patients included in the data compared to older patients. Because of this, none of our Beta parameters display any opposite signs from what's expected, so this indicates multicollinearity isn't present in our final model. We'll also run a VIF test in A2.6.

```
summary(lm14)$coefficients
```

```
##       Estimate  Std. Error   t value      Pr(>|t|)
## x1  0.07585424 0.002758234 27.501013 3.002555e-132
## x3  0.17324765 0.003762881 46.041226 8.988284e-278
## x4  0.24817962 0.030639366  8.100025  1.229812e-15
## z1 -0.02071067 0.005973352 -3.467178  5.426875e-04
## z2  0.07545408 0.007842731  9.620894  3.135365e-21
```

All coefficients are statistically significant with p-values < 0.05, indicating strong evidence that these variables have an impact on insurance charges.

R-squared:

```
summary(lm14)$r.squared
```

```
## [1] 0.9779381
```

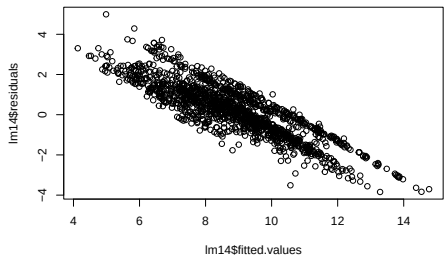
```
summary(lm14)$adj.r.squared
```

```
## [1] 0.9778553
```

The multiple R-squared value is 0.9779 and the adjusted R-squared is 0.9778553, indicating that the model explains approximately 97.8% of the variance in insurance charges.

Residuals:

```
plot(lm14$fitted.values, lm14$residuals)
```



Residuals (differences between observed and predicted values) exhibit a symmetric distribution with no clear patterns in the residuals vs. fitted values plot.

Overall Fit:

```
summary(lm14)$fstatistic
```

```
##    value    numdf    dendf
## 11817.58     5.00  1333.00
```

```
pf(summary(lm14)$fstatistic[1], summary(lm14)$fstatistic[2], summary(lm14)$fstatistic[3], lower.tail = 
```

```
## value
##     0
```

The F-statistic is significant (p-value < 0.001), suggesting that the model as a whole is a good fit for the data.

Our final model suggests that age, BMI, number of children, and the interactions between smoker and age as well as BMI are strong predictors of health insurance charges. The high R-squared value indicates an excellent fit, and the coefficients provide insights into the direction of the relationships between predictors and charges.

A2.7: Multicollinearity Analysis on Final Model: VIF

```
vif(lm14)
```

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

None of our variables display values greater than 10, so there are no indicates for multicollinearity present in our final model.

Future Work

We believe future work is required for a more thorough analysis on the factors that impact insurance charges

Considering additional variables variables can help in creating a more accurate model. Additional important inferences could have been made if other important factors such as history of chronic illness or the number of hospital visits were present in the data.

Works Cited

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/

CDC, Centers for Disease Control and Prevention. Assessing Your Weight, 2022, Jun 3. https://www.cdc.gov/healthyweight/assessing/index.html#:~:text=If%20your%20BMI%20is%20less,falls%20within%20the%20obese%20range

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/

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/

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

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/