Final Report

Group 11: Isiah Montalvo, Tristan Dull, Winston Zheng, Trisha Agrawal

2023-12-14

Contents

List of Figures

List of Tables

Introduction

Research Question

Question: Does age, BMI, number of children, and smoker status impact health insurance costs?

Various factors could impact the amount a person pays for health insurance, and we were particularly interested in age, BMI, children, and smoker status. We chose these variables because the data is diverse in terms of these four variables, so the data set was representative of people with a variety of lifestyles.

Hypothesis: There is no relationship between the predictor variables age, BMI, number of children, smoker status, and the response variable, charges.

Alternative Hypothesis: There is a relationship between the predictor variables age, BMI, number of children, smoker status, and the response variable, charges.

Data Description

This data set consists of six predictors and one response variable. Each row in the data set represents one person and their medical information. As shown in Table 1, this data set consists of numerical and categorical values. Our analysis investigates our hypothesis using categorical and numerical variables. The predictor values that are **bold** in Table 1 are the values we chose to use in our analysis. The decision-making process in terms of what variables we use will be explained in the Regression Analysis section.

Variable	Type	Description	Predictor or Response
Age	Numerical, Continuous	Age of primary beneficiary	Predictor
Sex	Categorical	Insurance contractor gender	Predictor
BMI	Numerical, Continuous	Body mass index of beneficiary	Predictor
Children	Numerical, Discrete	Number of dependents	Predictor
Smoker	Categorical	Smoker status of primary beneficiary	Predictor
Region	Categorical	The beneficiary's residential area	Predictor
Charges	Numerical, Continuous	Individual medical costs billed by health insurance	Response

Table 1: Variables Description

Exploratory Data Analysis

When conducting exploratory data analysis, we created scatter plots for each numerical value and bar graphs for each categorical value.

At first glance, the age vs. charges scatter plot in Figure 1 (below) shows three distinct groups with a positive correlation. The BMI vs. charges plot show two distinct groups with different slopes, indicating that another factor may be impacting insurance charges when a person's BMI is above 30. Finally, the children vs. charges plot displays a negative correlation as number of children increases. It is also clear that less people in the data set have 4 or 5 kids, which could imply that the data set is not representative of the people with 4 or 5 children.

The bar graphs in Figure 2 (below) show the distribution of each value for every categorical variable in the data set. It's clear that there is an even distribution of sex and region in the data set. There is a large amount of people who do not smoke compared to those who do smoke. With the large discrepancy between the number of people who do and do not smoke, the data set may not be representative of the population of people who do not smoke.

Table 2 (below) displays the descriptive statistics for the numerical variables in the data set. The counts for each value in a categorical value is printed on the bars in Figure 2 (below).

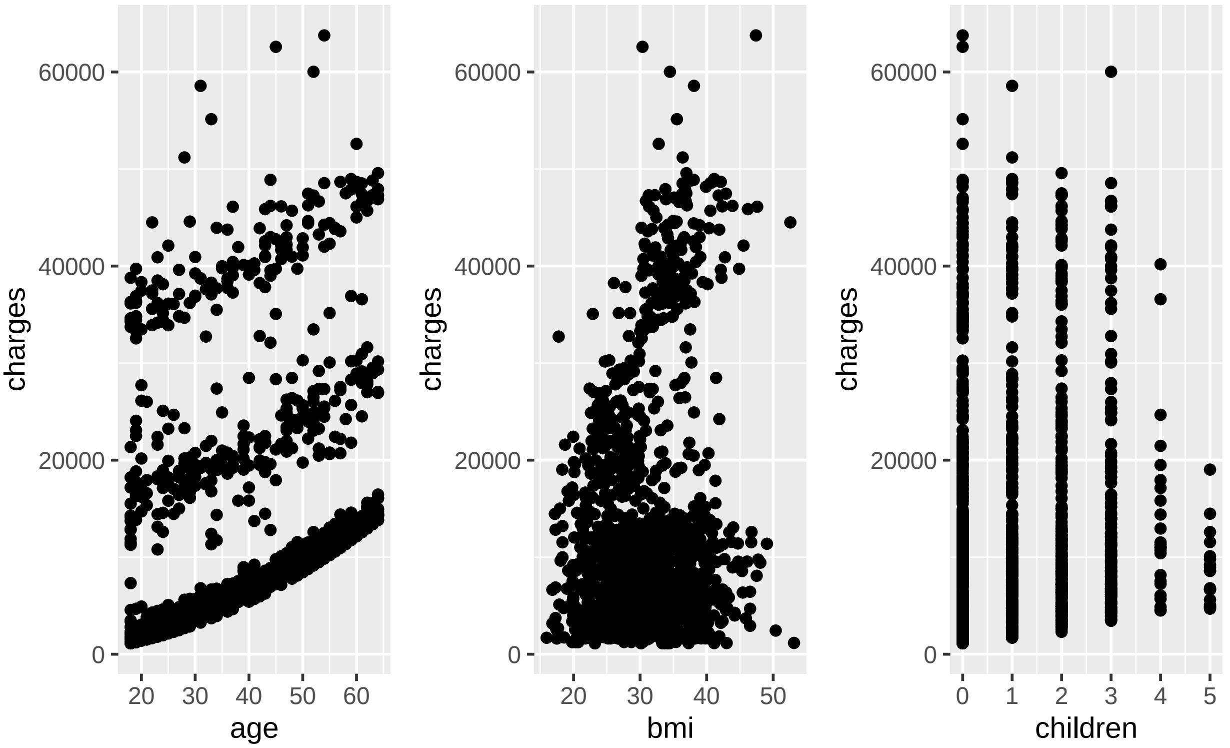

Figure 1: Exploratory Data Analysis: Numerical Variables

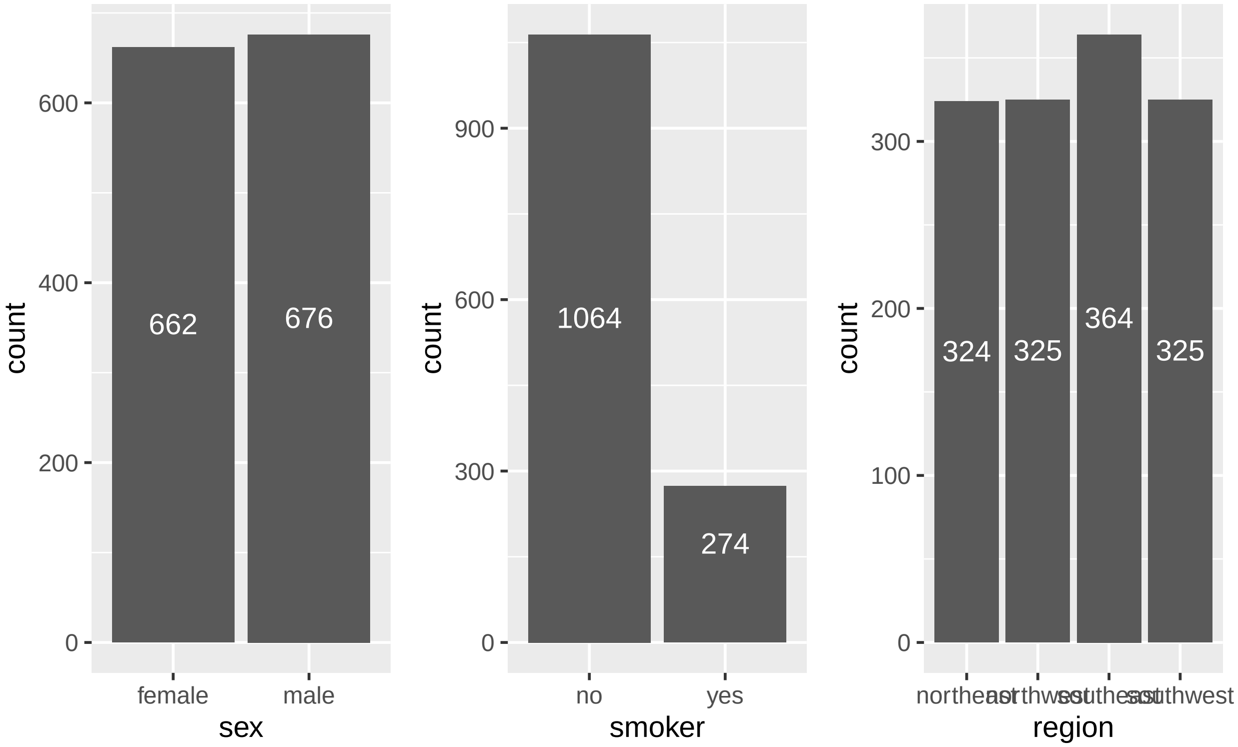

Figure 2: Exploratory Data Analysis: Categorical Variables

Variable	Min	Max	Mean
Age	18	64	39.21
BMI	15.96	53.13	30.66
Children	0	5	1.095
Charges	1122	63770	13270

Table 2: Descriptive Statistics: Numerical Variables

Regression Analysis

Transformations

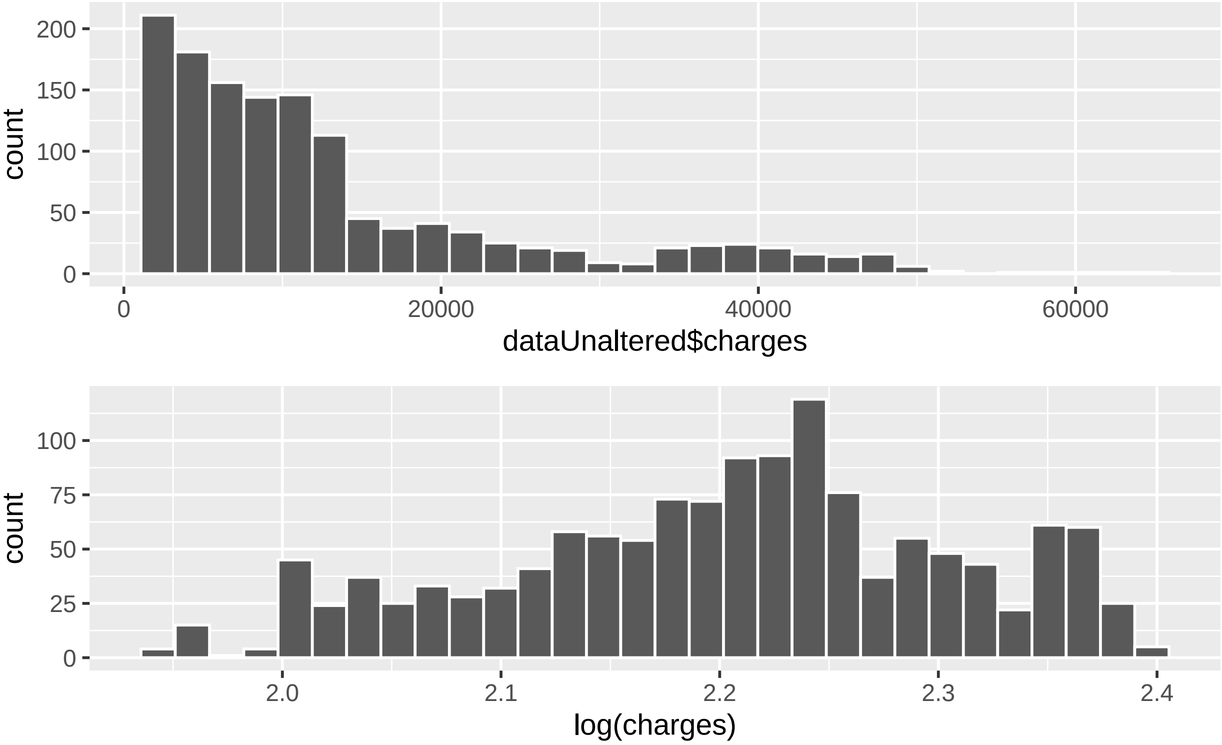

Figure 3: Histogram of Charges and log(charges)

charges → $\log(charges)$

Age → e^{Age}

We transformed Charges and Age to remove the skewness in the data and to satisfy assumptions.

x_1	e^{age}
x_2	bmi
x_3	children
x_4	smokeryes
$\widehat{y}$	estimated log($charges$)

Table 3: Variables in Model

Full Model

$\widehat{y} = \beta 0 + \beta 1x_1 + \beta 2x2 + \beta_3x_3 + \beta_4x4 + \beta 11x1^2 + \beta 12x_1x2 + \beta 13x_1x3 + \beta 14x_1x4 + \beta 22x2^2 + \beta 23x_2x3 + \beta 24x_2x4 + \beta 33x3^2 + \beta 34x_3x_4 + \epsilon$

Our full model (above) includes all first order and and single interaction terms. We found that the variables sex and region did not significantly impact charges, so we did not include them in our full model.

Reduced Model

$\widehat{y} = \beta 0 + \beta 1x_1 + \beta 2x2 + \beta_3x_3 + \beta_4x4 + \beta 11x1^2 + \beta 14x_1x4 + \beta 22x2^2 + \beta 24x_2x4 + \beta 33x3^2 + \beta 34x_3x_4 + \epsilon$

To get our reduced model we did stepwise model selection ranging from our full model to our null model (no independent variables). The F test and the AIC concluded that the reduced model was not significantly different from our full model (See Appendix “Final Model”).

Correlation Coefficient

$R_a^2 = .5447$

The adjusted R^2 value suggests that about 54% of the variability in log(charges) can be explained by our reduced model.

Assumptions

Linearity

The Residual vs fitted value plot (See Appendix: Figure 5) and the Q-Q plot (See Appendix: Figure 8) suggest that the linearity assumption is satisfied.

Constant Variance

The Residual vs. Fitted value plot (See Appendix: Figure 5) shows that, while there is some underlying pattern, our variance is close to constant.

Normality

To test normality we used a Q-Q plot (See Appendix: Figure 8) and a Shapiro-Wilk test. The Shapiro-Wilk test resulted in a p-value of $6.363*10^{-5}$ which is less than $\alpha = 0.05$, suggesting that the normality assumption is not satisfied. On the contrary, our Q-Q plot looks very close to normal, so we considered this assumption satisfied.

Independence

We ran a Durbin Watson test to check independence, which resulted in a p-value of .65, which is greater than $\alpha = .05$. Because the null hypothesis is that the variables are independent, we can conclude that this assumption is satisfied.

Multicollinearity Analysis

After checking variance inflation factors (See Appendix: Final Model Building and Multicolinearity Analysis) for our variables, our outputs were all very close to 1, suggesting that there is no multicollinearity in the data.

Conclusion

Our goal of this regression analysis is to see which variable have the most impact on health insurance costs. From the final model, the factors age, BMI, number of children, and smoking status are identified as significant factors affecting insurance costs, but this is to be expected since insurance costs are presumed to increase base on these factors. According to our final model, the smoker status is the most correlated to insurance costs.

There is also correlation between smoker status and other variables. People who smoke have higher charges regardless of the predictor variable it is paired with. We also noticed that the interaction terms between the variable age and smoker results in the insurance costs decreasing slightly. This means the interaction terms does not significantly affect the insurance cost compared to the individual terms. The quadratic terms in our equation indicate that after a certain threshold, the values start flattening out and decreasing. For example, the age range 70-80 will have a smaller decrease in charges compared to the age range 20-50.

For future models, we could add second-order terms and increase the sample size for more accurate results and add in another factor such as race/ethnicity. A larger sample size would ensure that the each population in the dataset is represented accurately.

Going back to our original research question and hypothesis, our p-value for our final reduced model is $2.2e - 16$, since this is less than our α (0.05) value, we can reject the null hypothesis. This means that there is a statistically significant relationship between the predictor variables we selected for our final reduced model and charges, the response variable.

Limitations

Assumptions

Based on the p-value obtained from the Shapiro-Wilk test, the normality assumption is not satisfied. However, once we transformed our predictor variable, charges, the Q-Q plot suggested normality, so we decided to proceed with our analysis.

Adjusted R^2 Value

The low Adjusted R^2 value we obtain from our model indicates that there is some randomness in the model. This could also suggest that there are some variables that impact insurance charges that are not included in this data set, such as ethnicity.

Statistically Insignificant Variables

Based on our research question and hypothesis, we determined that region and sex were statistically insignificant.

Appendix

Libraries Used to Complete Analysis

```
library(tidyverse)
library(car)
library("GGally")
library("ggplot2")
```

Data Reading

```
data <- read_csv("insurance.csv")
```

Response and Predictor Transformations

```
data$charges <- log(data$charges) # model transformation
data$age <- exp(data$age)
```

Final Model Building and Multicolinearity Analysis

```
ggpairs(data) + theme()
```

```
# reduced model found from stepwise regression below
reduced <- lm(charges ~ age + bmi + children + smoker + age*bmi +
    age*smoker + bmi*smoker + children*smoker, data = data)
vif(reduced)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif

##             age             bmi        children          smoker         age:bmi
##       41.237707        1.321357        1.236788       26.201782       38.373904
##      age:smoker      bmi:smoker children:smoker
##        1.654207       25.477501        2.158604
```

```
# model version with no interaction terms
individual_terms <- lm(charges ~ age + bmi + children + smoker, data = data)
vif(individual_terms)
```

```
##      age      bmi children   smoker
## 1.009381 1.004174 1.004483 1.001317
```

Second Order and Interaction Terms Stepwise Model Transformation

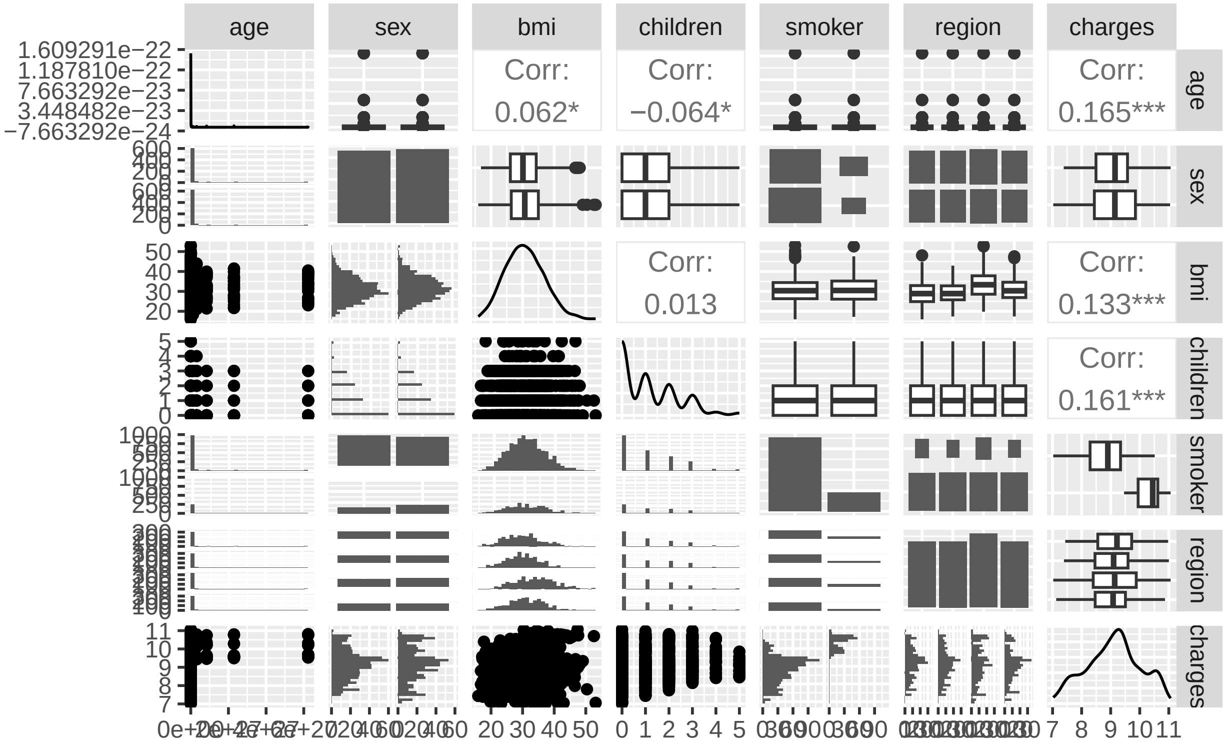

Figure 4: Correlation PairPlot

```
# full model with first order, second order, and interaction terms
# NOTE: model transformation is done above, where the data is read in
full <- lm(charges ~ age + bmi + children + smoker + age^2 + bmi^2 +
             children^2 + age*bmi + age*children + age*smoker
           + bmi*children + bmi*smoker + children*smoker
            , data = data)
null <- lm(charges ~ 1, data = data)
step_model1 <- step(full, scope = list(lower = null, upper = full),
                    direction = "both",test="F")
```

```
## Start:  AIC=-1203.02
## charges ~ age + bmi + children + smoker + age^2 + bmi^2 + children^2 +
##     age * bmi + age * children + age * smoker + bmi * children +
##     bmi * smoker + children * smoker
## 
##                   Df Sum of Sq    RSS     AIC F value    Pr(>F)    
## - bmi:children     1    0.0008 535.59 -1205.0  0.0020 0.9647022    
## - age:children     1    0.5673 536.16 -1203.6  1.4056 0.2360076    
## <none>                         535.59 -1203.0                      
## - age:bmi          1    0.8814 536.47 -1202.8  2.1837 0.1397122    
## - children:smoker  1    4.7602 540.35 -1193.2 11.7941 0.0006125 ***
## - age:smoker       1    5.3602 540.95 -1191.7 13.2807 0.0002785 ***
## - bmi:smoker       1   14.6456 550.24 -1168.9 36.2866 2.202e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Step:  AIC=-1205.02
## charges ~ age + bmi + children + smoker + age:bmi + age:children +
##     age:smoker + bmi:smoker + children:smoker
##
##                   Df Sum of Sq    RSS     AIC F value    Pr(>F)
## - age:children     1    0.5750 536.17 -1205.6  1.4258 0.2326695
## <none>                          535.59 -1205.0
## - age:bmi          1    0.8820 536.47 -1204.8  2.1870 0.1394155
## + bmi:children     1    0.0008 535.59 -1203.0  0.0020 0.9647022
## - children:smoker  1    4.7629 540.35 -1195.2 11.8097 0.0006074 ***
## - age:smoker       1    5.3644 540.95 -1193.7 13.3012 0.0002755 ***
## - bmi:smoker       1   14.6892 550.28 -1170.8 36.4221 2.058e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Step:  AIC=-1205.59
## charges ~ age + bmi + children + smoker + age:bmi + age:smoker +
##     bmi:smoker + children:smoker
##
##                   Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                          536.17 -1205.6
## - age:bmi          1    0.8938 537.06 -1205.4  2.2154 0.1368751
## + age:children     1    0.5750 535.59 -1205.0  1.4258 0.2326695
## + bmi:children     1    0.0085 536.16 -1203.6  0.0211 0.8845493
## - children:smoker  1    4.7956 540.96 -1195.7 11.8869 0.0005830 ***
## - age:smoker       1    5.2628 541.43 -1194.5 13.0449 0.0003154 ***
## - bmi:smoker       1   14.4833 550.65 -1171.9 35.8999 2.671e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Final Model

```
# final reduced model
reduced <- lm(charges ~ age + bmi + children + smoker + age*bmi +
    age*smoker + bmi*smoker + children*smoker, data = data)
anova(full,reduced)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + bmi + children + smoker + age^2 + bmi^2 + children^2 +
##     age * bmi + age * children + age * smoker + bmi * children +
##     bmi * smoker + children * smoker
## Model 2: charges ~ age + bmi + children + smoker + age * bmi + age * smoker +
##     bmi * smoker + children * smoker
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1327 535.59
## 2   1329 536.17 -2  -0.57581 0.7133 0.4902
```

```
summary(reduced)
```

```
##
```

```
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + age *
##     bmi + age * smoker + bmi * smoker + children * smoker, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73656 -0.38910  0.02948  0.42461  1.75896
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         8.316e+00  1.031e-01  80.629  < 2e-16 ***
## age                 3.985e-28  1.317e-28   3.025 0.002533 **
## bmi                 9.036e-03  3.274e-03   2.759 0.005869 **
## children            1.507e-01  1.603e-02   9.404  < 2e-16 ***
## smokeryes           4.048e-01  2.203e-01   1.838 0.066321 .
## age:bmi            -5.672e-30  3.810e-30  -1.488 0.136875
## age:smokeryes      -1.706e-28  4.723e-29  -3.612 0.000315 ***
## bmi:smokeryes       4.130e-02  6.892e-03   5.992 2.67e-09 ***
## children:smokeryes -1.276e-01  3.702e-02  -3.448 0.000583 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6352 on 1329 degrees of freedom
## Multiple R-squared:  0.5257, Adjusted R-squared:  0.5229
## F-statistic: 184.1 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Residual diagnostics of final model

Linearity and Constant Variance Assumption

```
# First/Second: Linearity check and constant variance:
ggplot(data = reduced, aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Predicted values", y = "Residuals")
```

Normality Assumption

```
# Third:Normality Check
ggplot(data = reduced, aes(x = .resid)) + geom_histogram()
```

```
ggplot(data = reduced, aes(x = .resid)) + geom_boxplot()
```

```
qqnorm(resid(reduced))
qqline(resid(reduced))
```

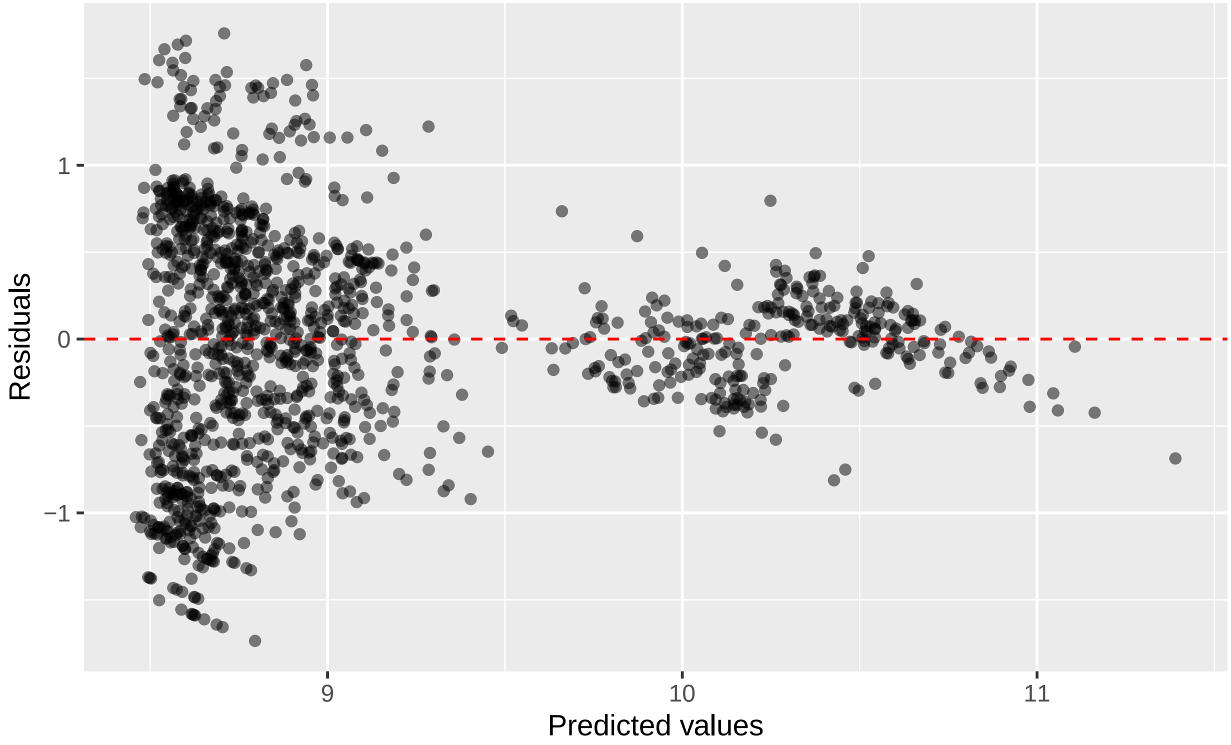

Figure 5: Residuals vs Fitted

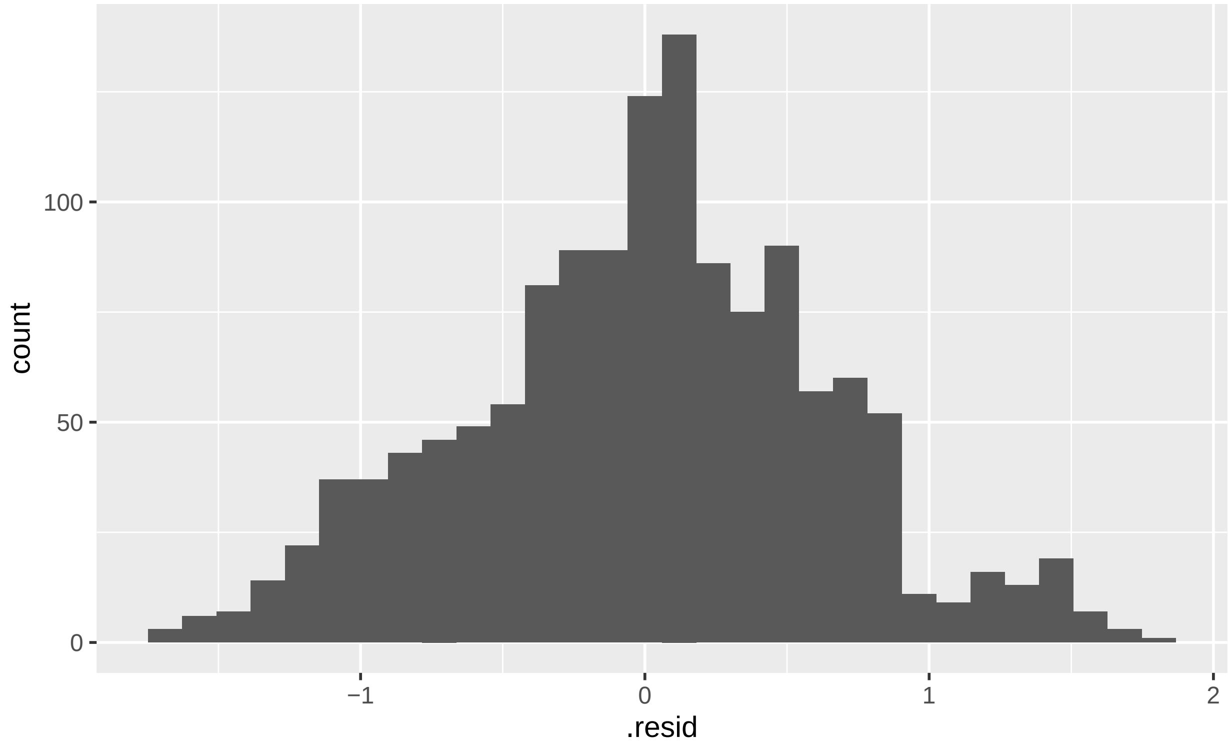

Figure 6: Normality Histogram

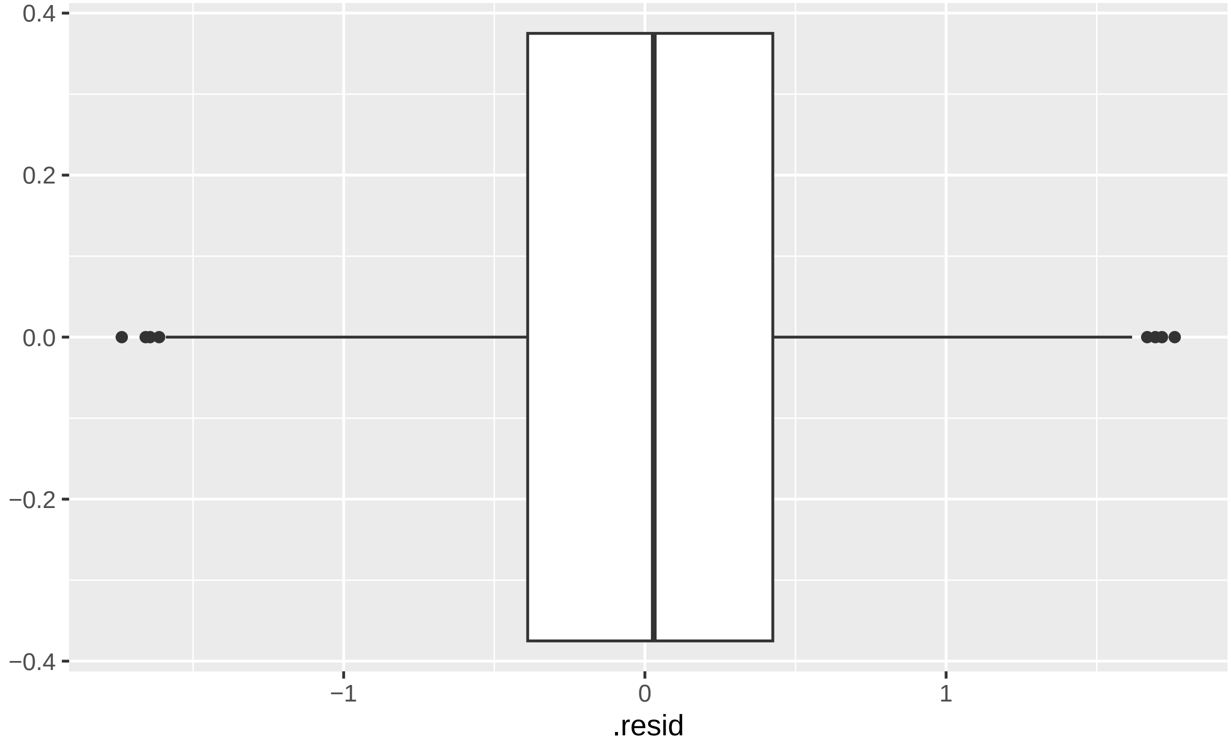

Figure 7: Normality BoxPlot

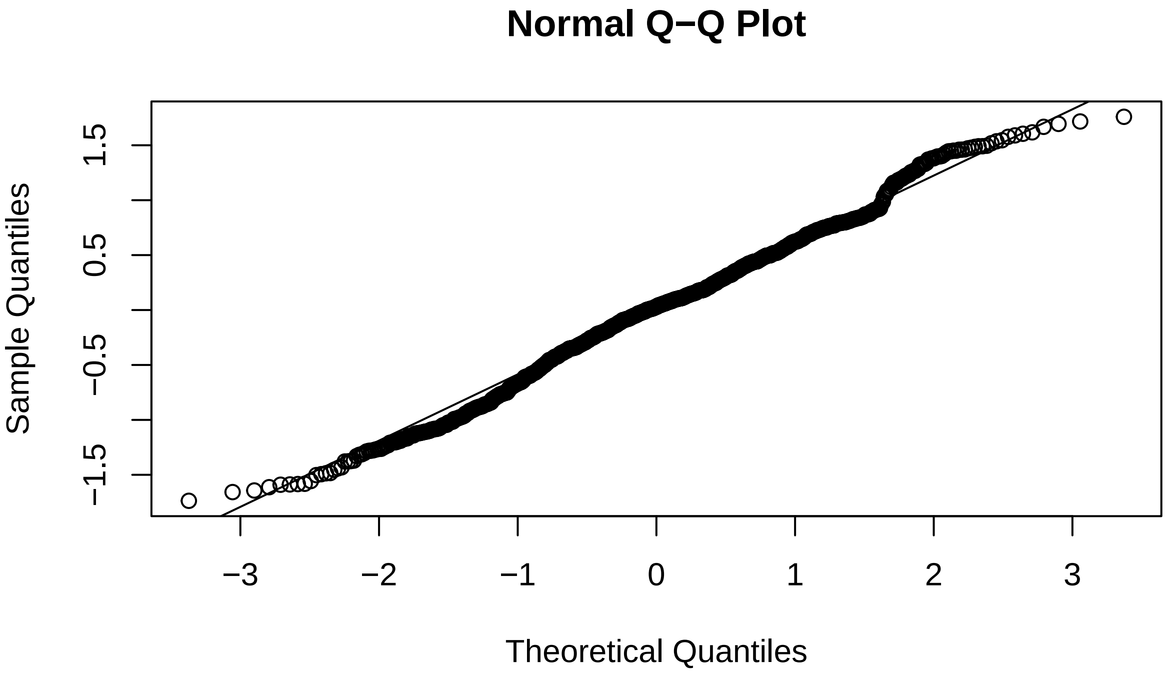

Figure 8: Normality QQPlot

```
shapiro.test(resid(reduced))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(reduced)
## W = 0.99421, p-value = 4.631e-05
```

Independence Assumption

```
# Fourth: Independence check
check_ind <- lm(charges ~ age + bmi + children + smoker , data = data)
set.seed(5)
dwt(reduced)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.01183687      2.023378   0.656
##  Alternative hypothesis: rho != 0
```

```
dwt(check_ind)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.01230911      2.024375   0.644
##  Alternative hypothesis: rho != 0
```

Considered Models

```
full <- lm(charges ~ age + bmi + children + smoker + age^2 + bmi^2
           + children^2 + age*bmi + age*children + age*smoker
           + bmi*children + bmi*smoker + children*smoker, data = data)
```