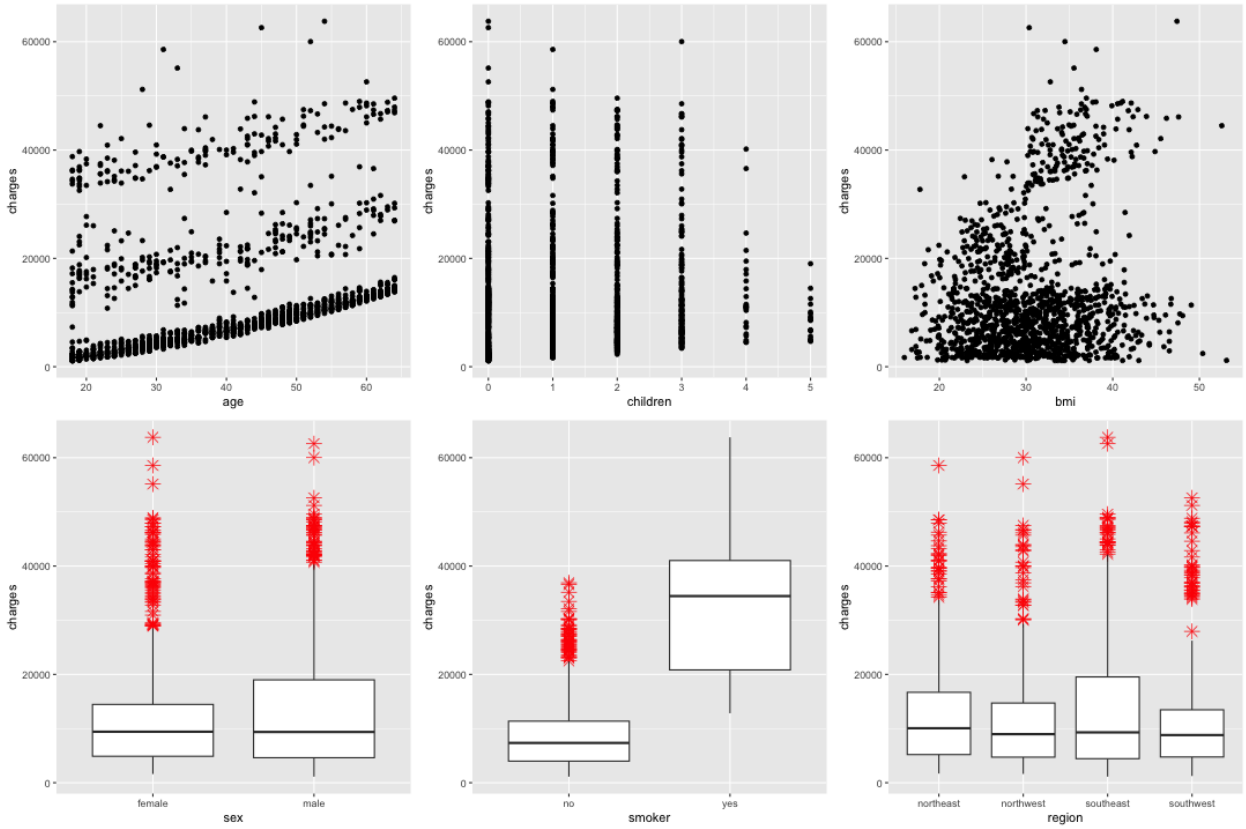
STAT170 Final Project Report

Nicole Carter, Jonathan Emmons, Dylan Nunes, Joonho Han

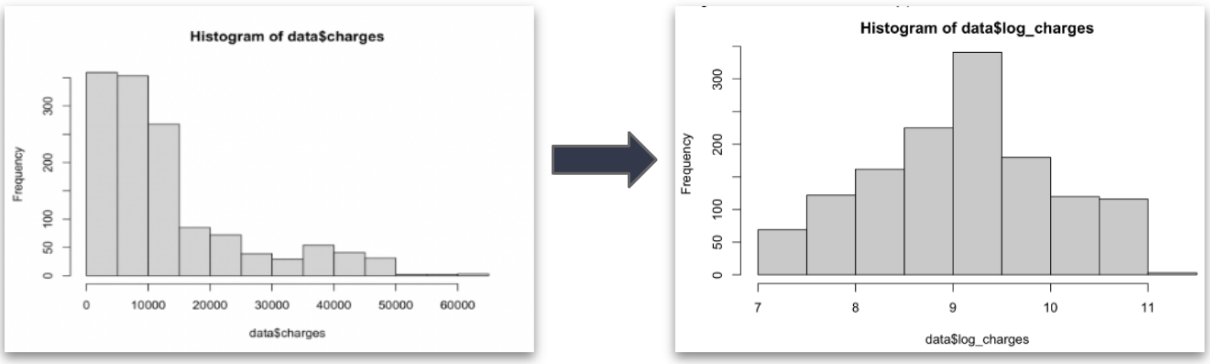
Introduction

For the final project, we are using the Medical Cost Personal Datasets (Insurance dataset). According to the book Machine Learning with R by Brett Lantz: “For this analysis, we will use a simulated dataset containing medical expenses for patients in the United States. These data were created for this book using demographic statistics from the U.S. Census Bureau, and thus approximately reflect real-world conditions.”

The given data includes information on an individual’s age, sex, BMI, number of dependents covered by their insurance, smoking status, beneficiary’s place of residence based on the region in the US (region), and the individual yearly medical expenses in dollars (charges). Our quantitative variables are age, BMI, number of dependents, and charges, while our qualitative variables are sex, smoking status, and region. Based on our given data and what we know about insurance, we want to explore which of these factors influence individual medical costs billed by health insurance? In terms of hypothesis testing: Our Null Hypothesis would be that none of the factors have an effect on the individual medical costs billed by health insurance. The Alternative Hypothesis would be that at least one of the variables have an effect on health insurance charges.



Based on the scatterplots of our EDA when comparing our predictor variables with the response, we can conclude that an individual's age and their smoking status seem to have the strongest correlation with insurance charges.



And when visualizing the response variable, the graph was not normally distributed, which we decided would further require a logarithmic transformation in order to normalize the data.

Regression Analysis

This is our final model:

$Y=\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_{3}x_{3} + \beta_4x_4 - \beta_5x_5 - \beta_6x_6 - \beta_7x_7 - \beta_8x_8 - \beta_9x_1^2 - \beta_{10}x_2^2 - \beta_{11}x_3^2 - \beta_{12}x_1x_4+ \beta_{13}x_3x_4 - \beta{14}x_1x_2 - \beta{15}x_2x_4 + \beta{16}x_1x_5 + \beta{16}x_1x_6 + \beta{17}x_1x_7 + \beta{18}x_1x_8 + \beta{19}x_4x_5 + \beta{20}x_4x_6 + \beta{21}x_4x_7 + \beta{22}x_4x_8 + \epsilon$

Assumptions of ϵ

* Normality $\epsilon \sim N(0, \sigma^2)$ (Appexdix #)

The histogram is bell-shaped strongly resembling a normal distribution. QQplot shows most values following the normal line closely except for a small portion deviating upward towards the end. We decide that this small deviation is being caused by the same unknown variable that was making the constant variance assumption fail. The Shapiro Wilks test failed as well so we conclude that the assumption is violated. Normality assumption violated

* Constant Variance $\sigma\neq \sigma(X)$ (Appendix #)

This plot shows that there is no pattern for most of the data with respect to the residuals. There is a small portion that does show a pattern in the residuals and we hypothesize this is being caused by the unknown variable. Since we are unable to take the unknown variable into account ,we conclude that the assumption is violated.. Constant Variance assumption violated

* Linearity (Appendix #)

This plot shows most of the data is clustered at the bottom of the plot with no clear pattern. There is a small portion that deviates and seems to have a linear pattern. There is clearly something causing this data to deviate in such a distinct way and we hypothesize that the unknown variable may be causing this. Since we are unable to take the unknown variable into account, we conclude that the assumption is violated. Linearity Assumption violated

* Independence (Appendix #)

P-value > .05 so we fail to reject null and assume that the data is independent. Independence assumption not violated.

How was our final model chosen?

Through our exploratory data analysis, we concluded that our response variable had to be log transformed to satisfy having a normal distribution. We then proceeded to create both a stepwise model and the all-purpose-regression best subset model including second order terms and interactions. This also results in the terms and interactions automatically being considered and compared throughout the process. Through comparing both models using ANOVA and the F-test, we concluded that the stepwise model was a statistically better predictor of insurance charges, resulting in that output being the final model. The process and code can be shown in the appendix (pg #).

Model discussion and diagnostics, model fit statistics

Our model ran into a trend where a good chunk of our data was affected by a hidden variable, resulting in a majority of our assumptions in our diagnostics being violated. For the fit of the model itself however, the model proved to be fairly strong in predicting the log insurance

charges, as the Adjusted R-Squared value was 0.8391, and it also proved to be statistically significant as the p-value of 2.2e-16 was less than 0.05.

Conclusion

Variables: x_1 - age, x_2 - children, x_3 - bmi, x_4 - smoker, x_5 region north, x_6 region southeast, x_7 regionsouthwest, x_8 sexmale

Model:

$$\hat{y} = 6.341 + 0.05238x_1 + 0.3477x_2 + 0.04569x_3 + 1.306x_4 - 0.1856x_5 - 0.4361x_6 - 0.4779x_7 - 0.2975x_8 - 0.0001832x_1^2 - 0.0176x_2^2 - 0.0007x_3^2 - 0.03286(x_1*x_4) + 0.04959(x_3*x_4) - 0.004225(x_1*x_2) - 0.127(x_2*x_4) + 0.002682(x_1*x_5) + 0.007159(x_1*x_6) + 0.0073(x_1*x_7) + 0.004858(x_1*x_8) + 0.07285(x_4*x_5) + 0.09043(x_4*x_6) + 0.1928(x_4*x_7) + 0.09103(x_4*x_8)$$

Our model produced some unexpected results. Since the dataset is about insurance, one can assume that more likely than not, having traits such as being a smoker, old age, and such would cause an increase in insurance charges. We can see this as a majority of our predictors have positive coefficients, leading towards increasing insurance charges. Surprisingly, we can also note how the region one is in as well as being of the male sex have negative coefficients, resulting in a decrease in insurance charges. We can also see how all the quadratic predictors have negative coefficients, indicating a negative concave relationship, showing a reduction in the rate of increase for those predictors. Most notably, interactions between being a smoker and age, age and number of children, and also smoking and having children actually cause a decrease in charges for every 1 unit increase in their respective interactions.

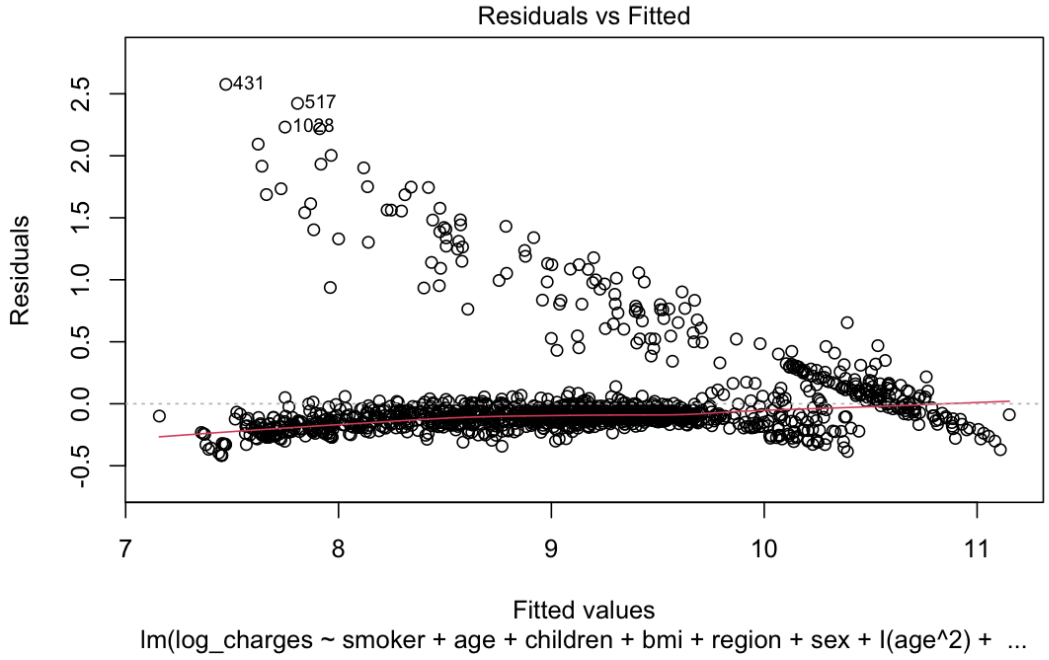
In terms of our hypotheses, as our final model has a p-value of 2.2e-16 which is less than 0.05, we can reject the null. We can then conclude that at least one of the variables would have an effect on the cost of health insurance. We can also see how in terms of our quantitative variables, that age, bmi, and the number of children respectfully, have the biggest impact on the

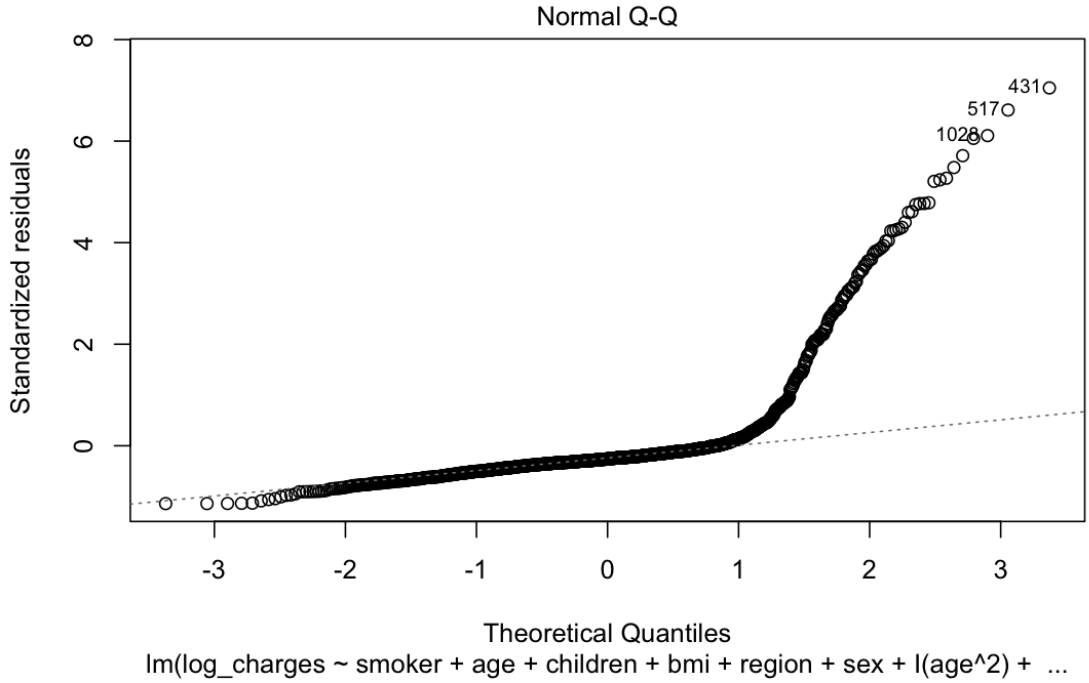
cost of insurance. We can also see how in terms of our categorical variables, smoking had the largest impact by far, having even more impact than our quantitative variables, which was also a bit unexpected.

Limitations

The biggest limitation we found in our report was the lack of satisfied assumptions. Since there are clearly two groups in the Residuals vs Fitted plot of our final model, we do not satisfy the homoskedasticity, and since the Normal Q-Q plot for our final model is very non-linear, we do not satisfy normality. Since we do not satisfy these assumptions, we cannot generalize our model to the population, which is a limitation in its usefulness. However, since our model has a very high Multiple R-squared value (0.8419), and a p-value (2.2e-16) less than 0.05, we can conclude that our model at least has a high prediction power for the data it was fitted to. Another potential limitation we may have had is the presence of a hidden variable in the data. We hypothesize that this is the case due to the presence of two distinct groups in the Residuals vs Fitted plot. Since our final model was selected through Stepwise Regression with the full model including all first order, second order, and interaction terms, we know that the groupings in the Residuals vs Fitted plot is not the result of a missing variable or interaction from the data. Thus, we hypothesize that it must be the result of a hidden variable that was not in our data. This was

a large limitation in our model, as it caused our assumptions to not be satisfied and made our model non-generalizable to the population.





Appendix

Libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(leaps)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(ggplot2)
```

Data

```
setwd('/Users/Jonathan/Documents/School/Fall 2024/STAT170/Final Project')
data <- read.csv('./insurance.csv')
names <- colnames(data)
predictors <- select(data, -charges)
log_charges <- log(data$charges)
```

Stepwise Model Selection

```
full_model <- lm(log_charges ~.^2 + I(age^2) + I(bmi^2) + I(children^2), data
= predictors)
null_model <- lm(log_charges ~ 1, data = data)
step_model <- step(null_model, scope = list(lower = null_model, upper =
full_model), direction = "both",test="F")
```

```
## Start:  AIC=-223.51
## log_charges ~ 1
```

```
##
##                 Df Sum of Sq     RSS      AIC   F value    Pr(>F)
## + smoker         1    500.68  629.79 -1004.24 1062.1239 < 2.2e-16 ***
## + age            1    314.96  815.51  -658.46  515.9771 < 2.2e-16 ***
## + I(age^2)       1    293.08  837.39  -623.04  467.5871 < 2.2e-16 ***
## + children       1     29.43 1101.05  -256.79   35.7047 2.941e-09 ***
## + bmi            1     19.90 1110.58  -245.27   23.9365 1.117e-06 ***
## + I(bmi^2)       1     17.64 1112.83  -242.55   21.1799 4.579e-06 ***
## + I(children^2)  1     15.92 1114.55  -240.48   19.0828 1.348e-05 ***
## <none>                       1130.47  -223.51
## + region         3      3.55 1126.92  -221.72    1.4020    0.2406
## + sex            1      0.04 1130.44  -221.55    0.0424    0.8369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-1004.24
## log_charges ~ smoker
##
##                 Df Sum of Sq     RSS      AIC   F value    Pr(>F)
## + age            1    335.35  294.44 -2019.56 1520.5251 < 2.2e-16 ***
## + I(age^2)       1    313.65  316.14 -1924.38 1324.4587 < 2.2e-16 ***
## + children       1     27.59  602.20 -1062.19   61.1727 1.055e-14 ***
## + bmi            1     19.16  610.63 -1043.57   41.8806 1.359e-10 ***
## + I(children^2)  1     17.59  612.20 -1040.15   38.3656 7.796e-10 ***
## + I(bmi^2)       1     16.34  613.45 -1037.42   35.5661 3.152e-09 ***
## + sex            1      2.31  627.48 -1007.16    4.9140   0.02681 *
## <none>                       629.79 -1004.24
## + region         3      2.48  627.31 -1003.52    1.7546   0.15403
## - smoker         1    500.68 1130.47  -223.51 1062.1239 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2019.56
## log_charges ~ smoker + age
##
##                 Df Sum of Sq    RSS      AIC   F value    Pr(>F)
## + age:smoker     1     42.38 252.06 -2225.49  224.2885 < 2.2e-16 ***
## + children       1     20.03 274.41 -2111.83   97.3779 < 2.2e-16 ***
## + I(children^2)  1     14.16 280.28 -2083.49   67.3849 5.214e-16 ***
## + I(age^2)       1      6.57 287.87 -2047.75   30.4392 4.133e-08 ***
## + bmi            1      5.70 288.73 -2043.72   26.3401 3.286e-07 ***
## + I(bmi^2)       1      4.86 289.57 -2039.85   22.4117 2.435e-06 ***
## + region         3      2.31 292.12 -2024.11    3.5158   0.01469 *
## + sex            1      1.37 293.06 -2023.81    6.2507   0.01253 *
## <none>                       294.44 -2019.56
## - age            1    335.35 629.79 -1004.24 1520.5251 < 2.2e-16 ***
## - smoker         1    521.08 815.51  -658.46 2362.6176 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Step:  AIC=-2225.49
## log_charges ~ smoker + age + smoker:age
##
##                 Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + children       1    21.101 230.96 -2340.5 121.7901 < 2.2e-16 ***
## + I(children^2)  1    14.643 237.41 -2303.6  82.2154 < 2.2e-16 ***
## + I(age^2)       1     6.551 245.51 -2258.7  35.5666 3.153e-09 ***
## + bmi            1     4.991 247.07 -2250.2  26.9274 2.440e-07 ***
## + I(bmi^2)       1     4.190 247.87 -2245.9  22.5323 2.289e-06 ***
## + region         3     2.438 249.62 -2232.5   4.3341  0.004759 **
## + sex            1     1.274 250.78 -2230.3   6.7708  0.009369 **
## <none>                       252.06 -2225.5
## - smoker:age     1    42.379 294.44 -2019.6 224.2885 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2340.47
## log_charges ~ smoker + age + children + smoker:age
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + bmi              1     4.817 226.14 -2366.7  28.3750 1.172e-07 ***
## + age:children     1     4.552 226.40 -2365.1  26.7798 2.630e-07 ***
## + children:smoker  1     4.551 226.40 -2365.1  26.7766 2.634e-07 ***
## + I(bmi^2)         1     4.030 226.93 -2362.0  23.6544 1.290e-06 ***
## + region           3     2.695 228.26 -2350.2   5.2336  0.001360 **
## + I(age^2)         1     1.557 229.40 -2347.5   9.0400  0.002691 **
## + sex              1     1.460 229.50 -2347.0   8.4763  0.003658 **
## + I(children^2)    1     1.164 229.79 -2345.2   6.7453  0.009503 **
## <none>                       230.96 -2340.5
## - children         1    21.101 252.06 -2225.5 121.7901 < 2.2e-16 ***
## - smoker:age       1    43.450 274.40 -2111.8 250.7765 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2366.68
## log_charges ~ smoker + age + children + bmi + smoker:age
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + bmi:smoker       1    20.864 205.27 -2494.2 135.2824 < 2.2e-16 ***
## + age:children     1     5.002 221.14 -2394.6  30.1067 4.890e-08 ***
## + children:smoker  1     4.429 221.71 -2391.1  26.5872 2.900e-07 ***
## + region           3     4.435 221.70 -2387.2   8.8618 8.123e-06 ***
## + I(age^2)         1     1.763 224.38 -2375.2  10.4573  0.001252 **
## + sex              1     1.734 224.40 -2375.0  10.2842  0.001374 **
## + I(bmi^2)         1     1.644 224.49 -2374.4   9.7472  0.001835 **
## + I(children^2)    1     1.086 225.05 -2371.1   6.4235  0.011376 *
## <none>                       226.14 -2366.7
## + age:bmi          1     0.256 225.88 -2366.2   1.5072  0.219783
## + bmi:children     1     0.001 226.14 -2364.7   0.0042  0.948291
## - bmi              1     4.817 230.96 -2340.5  28.3750 1.172e-07 ***
```

```
## - children         1    20.928 247.07 -2250.2 123.2689 < 2.2e-16 ***
## - smoker:age       1    42.738 268.88 -2137.1 251.7328 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2494.19
## log_charges ~ smoker + age + children + bmi + smoker:age + smoker:bmi
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + age:children     1     4.935 200.34 -2524.8  32.7649 1.284e-08 ***
## + region           3     4.854 200.42 -2520.2  10.7213 5.788e-07 ***
## + children:smoker  1     4.223 201.05 -2520.0  27.9350 1.465e-07 ***
## + sex              1     2.393 202.88 -2507.9  15.6843 7.881e-05 ***
## + I(age^2)         1     1.863 203.41 -2504.4  12.1808 0.0004986 ***
## + I(bmi^2)         1     1.862 203.41 -2504.4  12.1737 0.0005005 ***
## + I(children^2)    1     0.838 204.44 -2497.7   5.4534 0.0196779 *
## <none>                         205.27 -2494.2
## + bmi:children     1     0.109 205.16 -2492.9   0.7094 0.3998014
## + age:bmi          1     0.104 205.17 -2492.9   0.6710 0.4128665
## - smoker:bmi       1    20.864 226.14 -2366.7 135.2824 < 2.2e-16 ***
## - children         1    21.478 226.75 -2363.1 139.2628 < 2.2e-16 ***
## - smoker:age       1    46.981 252.25 -2220.4 304.6266 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2524.76
## log_charges ~ smoker + age + children + bmi + smoker:age + smoker:bmi +
##     age:children
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + children:smoker  1     4.470 195.87 -2552.9  30.3263 4.378e-08 ***
## + region           3     4.743 195.60 -2550.8  10.7251 5.757e-07 ***
## + sex              1     2.725 197.61 -2541.1  18.3261 1.995e-05 ***
## + I(bmi^2)         1     1.980 198.36 -2536.1  13.2685 0.0002803 ***
## + I(age^2)         1     1.758 198.58 -2534.6  11.7640 0.0006224 ***
## + I(children^2)    1     1.137 199.20 -2530.4   7.5840 0.0059692 **
## <none>                         200.34 -2524.8
## + age:bmi          1     0.074 200.26 -2523.2   0.4899 0.4840998
## + bmi:children     1     0.000 200.34 -2522.8   0.0021 0.9637911
## - age:children     1     4.935 205.27 -2494.2  32.7649 1.284e-08 ***
## - smoker:bmi       1    20.797 221.14 -2394.6 138.0686 < 2.2e-16 ***
## - smoker:age       1    46.249 246.59 -2248.8 307.0384 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2552.94
## log_charges ~ smoker + age + children + bmi + smoker:age + smoker:bmi +
##     age:children + smoker:children
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
```

```
## + region           3     4.768 191.10 -2579.9  11.0290 3.738e-07 ***
## + sex              1     2.519 193.35 -2568.3  17.3047 3.388e-05 ***
## + I(children^2)    1     1.781 194.09 -2563.2  12.1830 0.0004981 ***
## + I(age^2)         1     1.741 194.13 -2562.9  11.9067 0.0005769 ***
## + I(bmi^2)         1     1.702 194.17 -2562.6  11.6438 0.0006635 ***
## <none>                         195.87 -2552.9
## + age:bmi          1     0.069 195.80 -2551.4   0.4659 0.4950100
## + bmi:children     1     0.011 195.86 -2551.0   0.0758 0.7831570
## - smoker:children  1     4.470 200.34 -2524.8  30.3263 4.378e-08 ***
## - age:children     1     5.182 201.05 -2520.0  35.1611 3.864e-09 ***
## - smoker:bmi       1    20.584 216.45 -2421.2 139.6667 < 2.2e-16 ***
## - smoker:age       1    43.940 239.81 -2284.1 298.1360 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2579.92
## log_charges ~ smoker + age + children + bmi + region + smoker:age +
##     smoker:bmi + age:children + smoker:children
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + sex              1     2.546 188.56 -2595.9  17.8906 2.501e-05 ***
## + I(age^2)         1     1.770 189.33 -2590.4  12.3900 0.0004463 ***
## + I(bmi^2)         1     1.695 189.41 -2589.8  11.8576 0.0005922 ***
## + I(children^2)    1     1.652 189.45 -2589.5  11.5549 0.0006958 ***
## + age:region       3     2.125 188.98 -2588.9   4.9587 0.0019971 **
## + bmi:region       3     1.527 189.57 -2584.7   3.5523 0.0139807 *
## <none>                         191.10 -2579.9
## + smoker:region    3     0.739 190.36 -2579.1   1.7111 0.1628475
## + age:bmi          1     0.071 191.03 -2578.4   0.4931 0.4826698
## + children:region  3     0.607 190.49 -2578.2   1.4053 0.2395874
## + bmi:children     1     0.013 191.09 -2578.0   0.0873 0.7677086
## - region           3     4.768 195.87 -2552.9  11.0290 3.738e-07 ***
## - smoker:children  1     4.495 195.60 -2550.8  31.1922 2.832e-08 ***
## - age:children     1     5.065 196.17 -2546.9  35.1430 3.901e-09 ***
## - smoker:bmi       1    20.988 212.09 -2442.5 145.6286 < 2.2e-16 ***
## - smoker:age       1    43.820 234.92 -2305.7 304.0583 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2595.87
## log_charges ~ smoker + age + children + bmi + region + sex +
##     smoker:age + smoker:bmi + age:children + smoker:children
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + I(age^2)         1     1.746 186.81 -2606.3  12.3740 0.0004501 ***
## + I(bmi^2)         1     1.718 186.84 -2606.1  12.1717 0.0005011 ***
## + age:sex          1     1.634 186.92 -2605.5  11.5727 0.0006892 ***
## + age:region       3     2.175 186.38 -2605.4   5.1415 0.0015472 **
## + I(children^2)    1     1.608 186.95 -2605.3  11.3869 0.0007610 ***
## + bmi:region       3     1.373 187.18 -2599.7   3.2335 0.0216029 *
```

```
## + sex:smoker       1     0.601 187.95 -2598.1   4.2305 0.0398995 *
## <none>                         188.56 -2595.9
## + smoker:region    3     0.791 187.76 -2595.5   1.8566 0.1350746
## + children:region  3     0.623 187.93 -2594.3   1.4618 0.2232641
## + age:bmi          1     0.054 188.50 -2594.2   0.3785 0.5385350
## + sex:bmi          1     0.052 188.50 -2594.2   0.3671 0.5446856
## + sex:children     1     0.039 188.52 -2594.2   0.2756 0.5997070
## + bmi:children     1     0.014 188.54 -2594.0   0.0985 0.7537449
## + sex:region       3     0.116 188.44 -2590.7   0.2721 0.8455389
## - sex              1     2.546 191.10 -2579.9  17.8906 2.501e-05 ***
## - region           3     4.795 193.35 -2568.3  11.2314 2.804e-07 ***
## - smoker:children  1     4.287 192.84 -2567.8  30.1288 4.839e-08 ***
## - age:children     1     5.383 193.94 -2560.2  37.8304 1.020e-09 ***
## - smoker:bmi       1    21.676 210.23 -2452.3 152.3197 < 2.2e-16 ***
## - smoker:age       1    43.766 232.32 -2318.6 307.5498 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2606.31
## log_charges ~ smoker + age + children + bmi + region + sex +
##     I(age^2) + smoker:age + smoker:bmi + age:children + smoker:children
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + I(bmi^2)         1     1.685 185.12 -2616.4  12.0425 0.0005367 ***
## + age:region       3     2.224 184.59 -2616.3   5.3043 0.0012324 **
## + age:sex          1     1.537 185.27 -2615.4  10.9735 0.0009492 ***
## + I(children^2)    1     1.055 185.75 -2611.9   7.5147 0.0062020 **
## + bmi:region       3     1.366 185.44 -2610.1   3.2438 0.0213014 *
## + sex:smoker       1     0.573 186.24 -2608.4   4.0697 0.0438611 *
## + smoker:region    3     0.840 185.97 -2606.3   1.9884 0.1138639
## <none>                         186.81 -2606.3
## + sex:bmi          1     0.044 186.76 -2604.6   0.3088 0.5785238
## + sex:children     1     0.026 186.78 -2604.5   0.1854 0.6668137
## + children:region  3     0.578 186.23 -2604.5   1.3657 0.2516686
## + bmi:children     1     0.016 186.79 -2604.4   0.1118 0.7381240
## + age:bmi          1     0.010 186.80 -2604.4   0.0677 0.7947113
## + sex:region       3     0.110 186.70 -2601.1   0.2591 0.8548735
## - I(age^2)         1     1.746 188.56 -2595.9  12.3740 0.0004501 ***
## - sex              1     2.521 189.33 -2590.4  17.8704 2.527e-05 ***
## - region           3     4.824 191.63 -2578.2  11.3974 2.215e-07 ***
## - smoker:children  1     4.273 191.08 -2578.1  30.2841 4.475e-08 ***
## - age:children     1     5.273 192.08 -2571.1  37.3699 1.283e-09 ***
## - smoker:bmi       1    21.771 208.58 -2460.8 154.3034 < 2.2e-16 ***
## - smoker:age       1    43.664 230.47 -2327.3 309.4666 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2616.44
## log_charges ~ smoker + age + children + bmi + region + sex +
##     I(age^2) + I(bmi^2) + smoker:age + smoker:bmi + age:children +
```

```
##     smoker:children
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + age:region       3     2.411 182.71 -2628.0   5.8060 0.0006101 ***
## + age:sex          1     1.497 183.63 -2625.3  10.7798 0.0010529 **
## + I(children^2)    1     1.088 184.04 -2622.3   7.8129 0.0052624 **
## + sex:smoker       1     0.448 184.68 -2617.7   3.2060 0.0735968 .
## <none>                         185.12 -2616.4
## + smoker:region    3     0.762 184.36 -2615.9   1.8178 0.1420182
## + sex:bmi          1     0.107 185.02 -2615.2   0.7664 0.3815061
## + children:region  3     0.581 184.54 -2614.6   1.3855 0.2455479
## + sex:children     1     0.012 185.11 -2614.5   0.0886 0.7660527
## + bmi:children     1     0.005 185.12 -2614.5   0.0336 0.8544964
## + age:bmi          1     0.002 185.12 -2614.4   0.0131 0.9089810
## + bmi:region       3     0.486 184.64 -2613.9   1.1577 0.3246795
## + sex:region       3     0.118 185.01 -2611.3   0.2814 0.8388227
## - I(bmi^2)         1     1.685 186.81 -2606.3  12.0425 0.0005367 ***
## - I(age^2)         1     1.713 186.84 -2606.1  12.2447 0.0004821 ***
## - sex              1     2.544 187.67 -2600.2  18.1803 2.152e-05 ***
## - smoker:children  1     3.981 189.10 -2590.0  28.4527 1.128e-07 ***
## - region           3     4.816 189.94 -2588.1  11.4720 1.992e-07 ***
## - age:children     1     5.390 190.51 -2580.0  38.5197 7.239e-10 ***
## - smoker:bmi       1    22.015 207.14 -2468.1 157.3318 < 2.2e-16 ***
## - smoker:age       1    43.937 229.06 -2333.5 313.9998 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2627.98
## log_charges ~ smoker + age + children + bmi + region + sex +
##     I(age^2) + I(bmi^2) + smoker:age + smoker:bmi + age:children +
##     smoker:children + age:region
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + age:sex          1     1.472 181.24 -2636.8  10.7094 0.0010935 **
## + I(children^2)    1     1.076 181.64 -2633.9   7.8127 0.0052633 **
## + sex:smoker       1     0.434 182.28 -2629.2   3.1410 0.0765760 .
## + smoker:region    3     0.888 181.82 -2628.5   2.1435 0.0929803 .
## <none>                         182.71 -2628.0
## + sex:bmi          1     0.147 182.57 -2627.1   1.0622 0.3028937
## + bmi:region       3     0.681 182.03 -2627.0   1.6424 0.1777701
## + age:bmi          1     0.108 182.60 -2626.8   0.7790 0.3775989
## + children:region  3     0.592 182.12 -2626.3   1.4271 0.2331738
## + sex:children     1     0.008 182.71 -2626.0   0.0582 0.8093963
## + bmi:children     1     0.003 182.71 -2626.0   0.0245 0.8756841
## + sex:region       3     0.124 182.59 -2622.9   0.2985 0.8264984
## - I(age^2)         1     1.762 184.47 -2617.1  12.7301 0.0003727 ***
## - age:region       3     2.411 185.12 -2616.4   5.8060 0.0006101 ***
## - I(bmi^2)         1     1.873 184.59 -2616.3  13.5280 0.0002445 ***
## - sex              1     2.597 185.31 -2611.1  18.7633 1.592e-05 ***
## - smoker:children  1     3.696 186.41 -2603.2  26.7006 2.741e-07 ***
```

```
## - age:children     1     5.586 188.30 -2589.7  40.3546 2.909e-10 ***
## - smoker:bmi       1    22.486 205.20 -2474.7 162.4466 < 2.2e-16 ***
## - smoker:age       1    44.519 227.23 -2338.2 321.6238 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2636.8
## log_charges ~ smoker + age + children + bmi + region + sex +
##     I(age^2) + I(bmi^2) + smoker:age + smoker:bmi + age:children +
##     smoker:children + age:region + age:sex
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + I(children^2)    1     1.093 180.15 -2642.9   7.9941 0.0047640 **
## + sex:smoker       1     0.469 180.77 -2638.3   3.4206 0.0646099 .
## + smoker:region    3     0.908 180.33 -2637.5   2.2084 0.0853788 .
## <none>                         181.24 -2636.8
## + age:bmi          1     0.142 181.10 -2635.8   1.0333 0.3095644
## + bmi:region       3     0.662 180.58 -2635.7   1.6087 0.1855256
## + sex:bmi          1     0.061 181.18 -2635.2   0.4453 0.5046979
## + sex:children     1     0.002 181.24 -2634.8   0.0113 0.9152391
## + bmi:children     1     0.001 181.24 -2634.8   0.0082 0.9277158
## + children:region  3     0.522 180.72 -2634.7   1.2679 0.2839365
## + sex:region       3     0.131 181.11 -2631.8   0.3168 0.8132776
## - age:sex          1     1.472 182.71 -2628.0  10.7094 0.0010935 **
## - I(age^2)         1     1.668 182.91 -2626.5  12.1385 0.0005101 ***
## - I(bmi^2)         1     1.830 183.07 -2625.4  13.3145 0.0002736 ***
## - age:region       3     2.385 183.63 -2625.3   5.7862 0.0006273 ***
## - smoker:children  1     3.798 185.04 -2611.1  27.6370 1.706e-07 ***
## - age:children     1     5.499 186.74 -2598.8  40.0216 3.433e-10 ***
## - smoker:bmi       1    22.926 204.17 -2479.4 166.8429 < 2.2e-16 ***
## - smoker:age       1    45.346 226.59 -2340.0 330.0084 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2642.89
## log_charges ~ smoker + age + children + bmi + region + sex +
##     I(age^2) + I(bmi^2) + I(children^2) + smoker:age + smoker:bmi +
##     age:children + smoker:children + age:region + age:sex
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + sex:smoker       1     0.454 179.69 -2644.3   3.3287 0.0683083 .
## + smoker:region    3     0.943 179.21 -2643.9   2.3058 0.0750884 .
## <none>                         180.15 -2642.9
## + bmi:region       3     0.706 179.44 -2642.1   1.7239 0.1602156
## + age:bmi          1     0.162 179.99 -2642.1   1.1838 0.2767840
## + sex:bmi          1     0.047 180.10 -2641.2   0.3457 0.5566739
## + sex:children     1     0.009 180.14 -2640.9   0.0628 0.8020986
## + bmi:children     1     0.004 180.15 -2640.9   0.0310 0.8601772
## + children:region  3     0.347 179.80 -2639.5   0.8456 0.4689159
## + sex:region       3     0.102 180.05 -2637.6   0.2476 0.8630726
```

```
## - I(children^2)    1     1.093 181.24 -2636.8   7.9941 0.0047640 **
## - I(age^2)         1     1.122 181.27 -2636.6   8.2078 0.0042373 **
## - age:sex          1     1.488 181.64 -2633.9  10.8891 0.0009932 ***
## - age:region       3     2.372 182.52 -2631.4   5.7855 0.0006279 ***
## - I(bmi^2)         1     1.866 182.01 -2631.1  13.6522 0.0002290 ***
## - smoker:children  1     4.250 184.40 -2613.7  31.0914 2.983e-08 ***
## - age:children     1     5.835 185.98 -2602.2  42.6927 9.127e-11 ***
## - smoker:bmi       1    22.581 202.73 -2486.9 165.2089 < 2.2e-16 ***
## - smoker:age       1    45.462 225.61 -2343.8 332.6106 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2644.27
## log_charges ~ smoker + age + children + bmi + region + sex +
##     I(age^2) + I(bmi^2) + I(children^2) + smoker:age + smoker:bmi +
##     age:children + smoker:children + age:region + age:sex + smoker:sex
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## + smoker:region    3     0.915 178.78 -2645.1   2.2419 0.0817012 .
## <none>                         179.69 -2644.3
## + bmi:region       3     0.681 179.01 -2643.3   1.6667 0.1723485
## + age:bmi          1     0.143 179.55 -2643.3   1.0467 0.3064573
## - smoker:sex       1     0.454 180.15 -2642.9   3.3287 0.0683083 .
## + sex:bmi          1     0.049 179.65 -2642.6   0.3586 0.5493625
## + sex:children     1     0.010 179.69 -2642.3   0.0713 0.7895074
## + bmi:children     1     0.008 179.69 -2642.3   0.0612 0.8046225
## + children:region  3     0.324 179.37 -2640.7   0.7910 0.4989409
## + sex:region       3     0.109 179.59 -2639.1   0.2648 0.8507763
## - I(children^2)    1     1.078 180.77 -2638.3   7.8983 0.0050213 **
## - I(age^2)         1     1.105 180.80 -2638.1   8.1006 0.0044935 **
## - age:sex          1     1.523 181.22 -2635.0  11.1618 0.0008584 ***
## - I(bmi^2)         1     1.730 181.43 -2633.4  12.6824 0.0003823 ***
## - age:region       3     2.356 182.05 -2632.8   5.7556 0.0006548 ***
## - smoker:children  1     4.427 184.12 -2613.7  32.4494 1.507e-08 ***
## - age:children     1     5.904 185.60 -2603.0  43.2683 6.866e-11 ***
## - smoker:bmi       1    21.443 201.14 -2495.4 157.1612 < 2.2e-16 ***
## - smoker:age       1    45.228 224.92 -2345.9 331.4801 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=-2645.1
## log_charges ~ smoker + age + children + bmi + region + sex +
##     I(age^2) + I(bmi^2) + I(children^2) + smoker:age + smoker:bmi +
##     age:children + smoker:children + age:region + age:sex + smoker:sex +
##     smoker:region
##
##                   Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## <none>                         178.78 -2645.1
## - smoker:region    3     0.915 179.69 -2644.3   2.2419 0.0817012 .
## + age:bmi          1     0.129 178.65 -2644.1   0.9470 0.3306560
```

```
## + bmi:region       3     0.661 178.12 -2644.1   1.6229 0.1822296
## - smoker:sex       1     0.427 179.21 -2643.9   3.1350 0.0768601 .
## + sex:bmi          1     0.061 178.72 -2643.6   0.4482 0.5033211
## + sex:children     1     0.005 178.78 -2643.1   0.0361 0.8493773
## + bmi:children     1     0.002 178.78 -2643.1   0.0146 0.9038061
## + children:region  3     0.344 178.44 -2641.7   0.8415 0.4711416
## + sex:region       3     0.109 178.67 -2639.9   0.2655 0.8502822
## - I(children^2)    1     1.116 179.90 -2638.8   8.1988 0.0042584 **
## - I(age^2)         1     1.144 179.92 -2638.6   8.4111 0.0037914 **
## - age:sex          1     1.541 180.32 -2635.6  11.3227 0.0007877 ***
## - I(bmi^2)         1     1.637 180.42 -2634.9  12.0297 0.0005405 ***
## - age:region       3     2.485 181.26 -2632.6   6.0884 0.0004103 ***
## - smoker:children  1     4.570 183.35 -2613.3  33.5870 8.522e-09 ***
## - age:children     1     5.844 184.62 -2604.1  42.9557 8.022e-11 ***
## - smoker:bmi       1    18.603 197.38 -2514.7 136.7304 < 2.2e-16 ***
## - smoker:age       1    44.286 223.06 -2351.0 325.4926 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```