Car Insurance Analysis

Nadia Khoo

2024-09-04

Introduction

The data I am working with is sourced from Allstate Indemnity Company's Private Passenger Automobile Maryland insurance dataset 2020. Obtained from Kaggle dataset: https://www.kaggle.com/datasets/thedevastator/insurance-companies-secret-sauce-finally-exposed?select=cgr-premiums-table.csv

It contains car insurance data with columns:
`territory` - territory the individual lives in
`gender` - gender of individual
`birthdate` - individual's birthdate
`ypc` - individual's years of prior coverage
`current_premium` - individual's current premium, what is being paid
`indicated _premium` - individual's indicated suggested by model premium
`selected _premium` - individual's selected by insurer premium
`underlying _premium` - individual's underlying base amount before adjustment premium
`fixed_expenses` - individual's fixed expenses
`underlying_total_premium` - individual's underlying total premium including adjustments
`cgr` - individual's CGR `cgr_factor` - individual's CGR factor, risk of claims

I approach this dataset with the question: *How does different factors such as age, gender, living areas, etc., affect the premium charged to policyholders?*

I think that current premium which is the variable `current_premium` would be the best use choice as it reflects the actual amounts individuals are paying for their insurance premiums. Hence,I will be dropping all other premium variables

Loading in libraries and dataset

```r
library(googledrive)
library(tidyverse)
library(ggplot2)
library(lubridate)
library(patchwork)
library(olsrr)
library(car)
```

```r
# downloading zip csv data from Google Drive
temp = tempfile(fileext = ".zip")

dl = drive_download(
  as_id("1fzpzgte8p3z_LJ7dLD4Xzmj5Rv5D1cQe"), path = temp, overwrite = TRUE
```

```r
)

out = unzip(temp, exdir = tempdir())

df = read.csv(out[1], sep = ",")
```

Cleaning dataset

```r
# taking a look at the original data
head(df)
```

```
##   territory gender  birthdate ypc current_premium indicated_premium
## 1       601      M  10/5/1947   0          863.97            830.58
## 2       601      F   7/6/1953   0          828.63            611.14
## 3       601      M  4/18/1956   0         1000.59            593.99
## 4       601      F  8/16/1956   0          700.42            547.95
## 5       601      F  1/23/1957   0          505.92            448.33
## 6       601      F 12/31/1960   0         1674.34            932.74
##   selected_premium underlying_premium fixed_expenses underlying_total_premium
## 1           862.57             673.06         175.98                   849.04
## 2           826.43             612.75         175.98                   788.73
## 3           996.60             858.20         175.98                  1034.18
## 4           697.84             571.49         180.48                   751.97
## 5           504.56             333.71         152.08                   485.79
## 6          1671.47            1505.90         180.48                  1686.38
##   cgr_factor cgr
## 1       1.02 ZHK
## 2       1.06 6NS
## 3       0.96 Z2D
## 4       0.91 D7G
## 5       1.06 3YN
## 6       0.99 Z20
```

```r
# Removing unnecessary variables indicated_premium, selected_premium,
# underlying_premium, underlying_total_premium, fixed_expenses and cgr
df1 = df %>%
  select(-indicated_premium, -selected_premium, -underlying_premium,
         -underlying_total_premium, -fixed_expenses, -cgr)

head(df1)
```

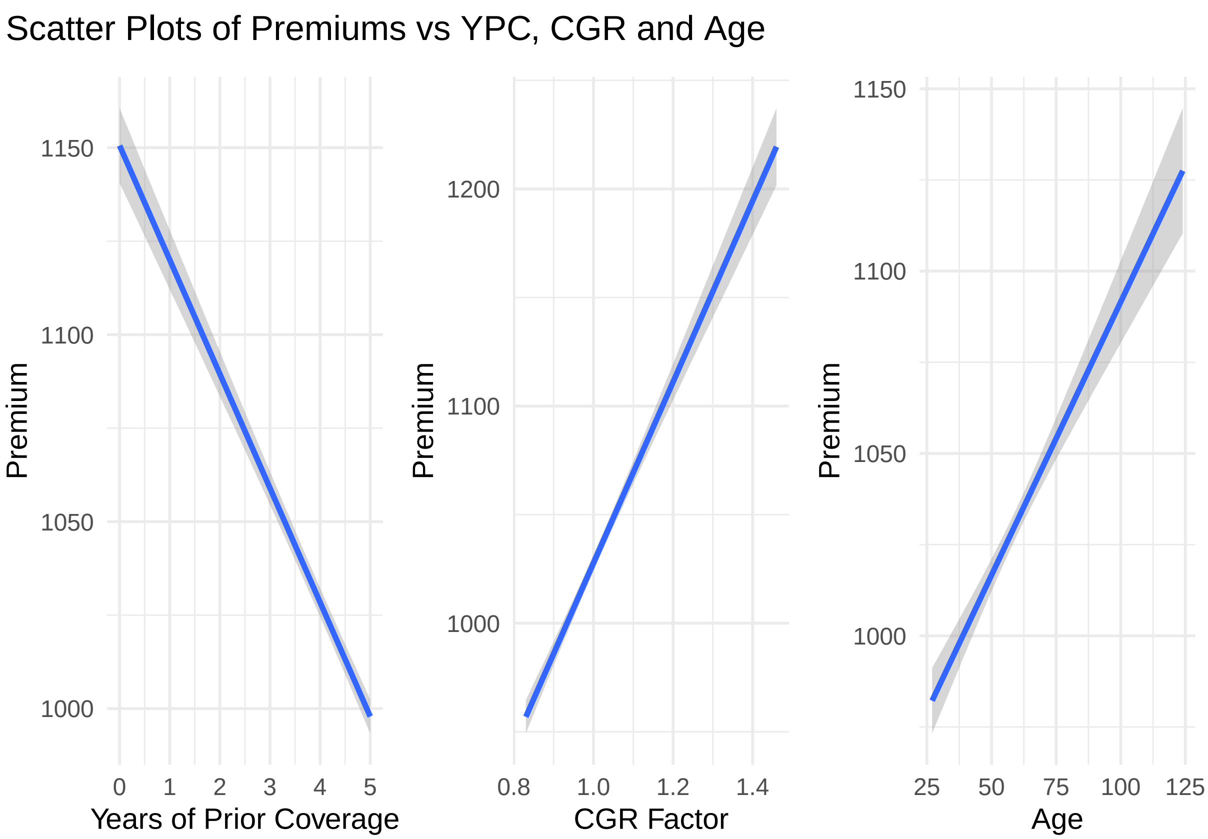
```
##   territory gender  birthdate ypc current_premium cgr_factor
## 1       601      M  10/5/1947   0          863.97       1.02
## 2       601      F   7/6/1953   0          828.63       1.06
## 3       601      M  4/18/1956   0         1000.59       0.96
## 4       601      F  8/16/1956   0          700.42       0.91
## 5       601      F  1/23/1957   0          505.92       1.06
## 6       601      F 12/31/1960   0         1674.34       0.99
```

```r
# changing gender to factor

df2 = df1%>%
  mutate(gender = as.factor(gender),
         birthdate = mdy(birthdate),
         territory = as.factor(territory)) %>%
  mutate(age = interval(birthdate, today()) / years(1)) %>%
  mutate(age = floor(age)) %>%
  select(-birthdate) %>%
  select(current_premium, everything())

head(df2)
```

```
##   current_premium territory gender ypc cgr_factor age
## 1          863.97       601      M   0       1.02  76
## 2          828.63       601      F   0       1.06  71
## 3         1000.59       601      M   0       0.96  68
## 4          700.42       601      F   0       0.91  68
## 5          505.92       601      F   0       1.06  67
## 6         1674.34       601      F   0       0.99  63
```

Correlation between variables

```r
# current_premium vs ypc
ypc_xy = ggplot(df2, aes(x = ypc, y = current_premium)) +
  geom_smooth(method = "lm") +
  labs(x = "Years of Prior Coverage",
       y = "Premium") +
  theme_minimal()

# current_premium vs cgr_factor
cgr_xy = ggplot(df2, aes(x = cgr_factor, y = current_premium)) +
  geom_smooth(method = "lm") +
  labs(x = "CGR Factor",
       y = "Premium") +
  theme_minimal()

# current_premium vs age
age_xy = ggplot(df2, aes(x = age, y = current_premium)) +
  geom_smooth(method = "lm") +
  labs(x = "Age",
       y = "Premium") +
  theme_minimal()

(ypc_xy + cgr_xy + age_xy) +
  plot_annotation(title = "Scatter Plots of Premiums vs YPC, CGR and Age")
```




```
# Since there is a general positive or negative linear correlation between
# premiums and the 3 numerical variables, will keep all 3 variables for now
```

Simplifying datasets

```
# territory has too many factors
# reduce number of categories for territory to top 4, based on count of entries
territory_top4 = df2 %>%
  count(territory, sort = TRUE) %>%
  head(5)

territory_top4_names = territory_top4$territory

df3 = df2 %>%
  filter(territory %in% territory_top4_names)

head(df3)
```

```
##   current_premium territory gender ypc cgr_factor age
## 1           98.89      1122      F   0       0.90  66
## 2          695.26      1122      M   0       0.91  66
## 3          656.93      1122      M   0       0.96  66
## 4         2136.57      1122      M   0       1.06  65
```

```
## 5          814.66      1122      F   0        1.06  65
## 6         1673.14      1122      F   0        1.05  65
```

Linear regression modelling

```r
# model 1 with all variables
model_1 = lm(current_premium ~ territory + gender + ypc + cgr_factor + age, data = df3)
summary_1 = summary(model_1)
summary_1
```

```
##
## Call:
## lm(formula = current_premium ~ territory + gender + ypc + cgr_factor +
##     age, data = df3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1382.8  -386.3  -165.6   200.0  5971.6
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   460.39993  106.15667   4.337 1.47e-05 ***
## territory1206 172.17729   25.64228   6.715 2.05e-11 ***
## territory1207 109.08704   27.09072   4.027 5.72e-05 ***
## territory1215  50.29052   26.78344   1.878  0.06047 .
## territory1234  68.85882   25.19556   2.733  0.00629 **
## genderM        98.45556   15.58330   6.318 2.83e-10 ***
## ypc           -36.43239    4.85794  -7.500 7.27e-14 ***
## cgr_factor    697.81273  100.50647   6.943 4.22e-12 ***
## age            -0.07626    0.51635  -0.148  0.88259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 620.3 on 6419 degrees of freedom
## Multiple R-squared:  0.03444,    Adjusted R-squared:  0.03323
## F-statistic: 28.62 on 8 and 6419 DF,  p-value: < 2.2e-16
```

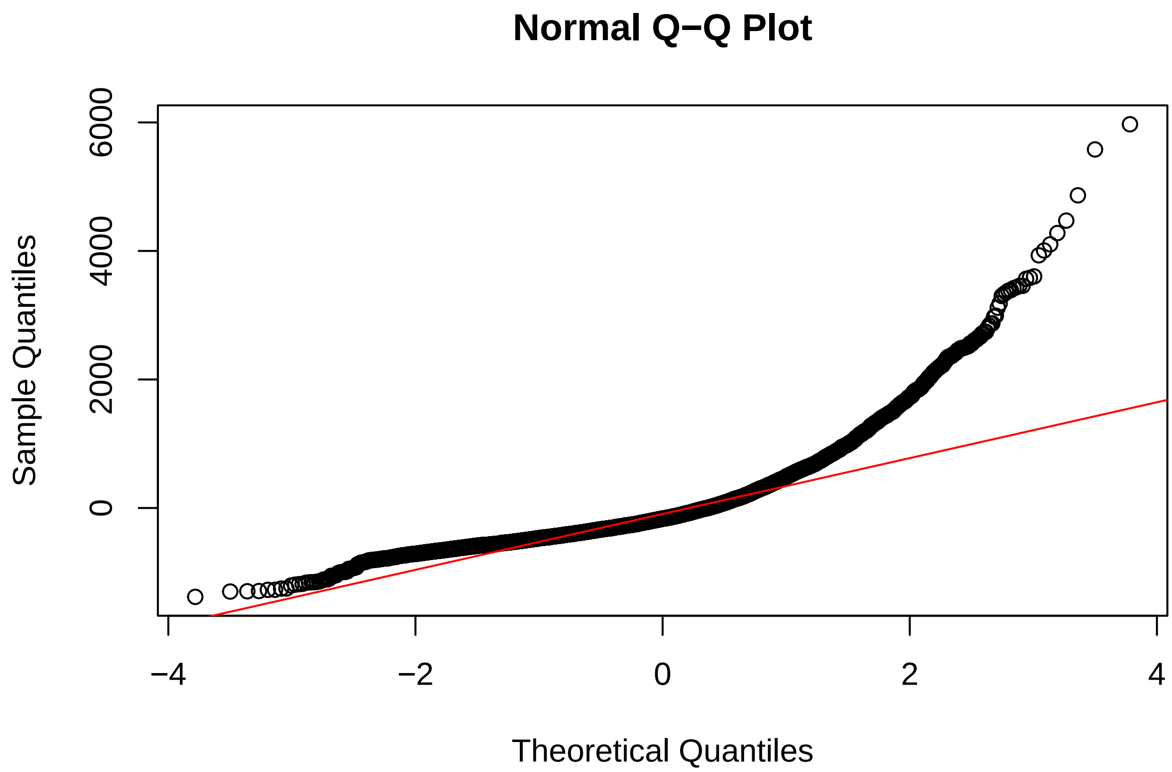
```r
# AIC and BIC
AIC(model_1)
```

```
## [1] 100919.3
```

```r
BIC(model_1)
```

```
## [1] 100987
```

```r
qqnorm(residuals(model_1))
qqline(residuals(model_1), col = "red")
```




```r
# removing territory1215 as p-value > 0.05
df4 = df3 %>%
  filter(territory != "1215")

head(df4)
```

```
##   current_premium territory gender ypc cgr_factor age
## 1           98.89      1122      F   0       0.90  66
## 2          695.26      1122      M   0       0.91  66
## 3          656.93      1122      M   0       0.96  66
## 4         2136.57      1122      M   0       1.06  65
## 5          814.66      1122      F   0       1.06  65
## 6         1673.14      1122      F   0       1.05  65
```

```r
model_2 = lm(current_premium ~ territory + gender + ypc + cgr_factor, data = df4)
summary_2 = summary(model_2)
summary_2
```

```
##
## Call:
## lm(formula = current_premium ~ territory + gender + ypc + cgr_factor,
##     data = df4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1361.6  -387.1  -170.2   196.1  5557.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    531.073    118.592   4.478 7.70e-06 ***
## territory1206  171.648     25.890   6.630 3.71e-11 ***
## territory1207  125.410     27.425   4.573 4.93e-06 ***
## territory1234   66.339     25.418   2.610  0.00908 **
## genderM        113.798     17.737   6.416 1.53e-10 ***
## ypc            -30.147      5.569  -5.414 6.46e-08 ***
## cgr_factor     581.945    115.761   5.027 5.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 622 on 5003 degrees of freedom
## Multiple R-squared:  0.03277,    Adjusted R-squared:  0.03161
## F-statistic: 28.25 on 6 and 5003 DF,  p-value: < 2.2e-16
```

```r
# AIC and BIC
AIC(model_2)
```

```
## [1] 78685.48
```

```r
BIC(model_2)
```

```
## [1] 78737.63
```

```r
# adjusted R squared decreases slightly for model_2
# qq plots are similar with slight deviation from normality
# (removed as no comparison insights)
# however AIC and BIC is significantly smaller for model_2
# hence will keep model_2 as a better fit
```

Addition of interactive terms

```r
# analysed variables and gathered possible interactions between variables
# adding interactive terms gender*cgr_factor, ypc*gender,
# territory*age and territory*cgr_factor
model_3 = lm(current_premium ~ territory + gender + ypc + cgr_factor
             + gender*cgr_factor +  ypc*gender + territory*age + territory*cgr_factor
             , data = df4)
summary_3 = summary(model_3)
summary_3
```

```
##
## Call:
## lm(formula = current_premium ~ territory + gender + ypc + cgr_factor +
##     gender * cgr_factor + ypc * gender + territory * age + territory *
##     cgr_factor, data = df4)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1383.8  -385.3  -166.6   199.1  5521.9
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1524.996    286.776   5.318 1.10e-07 ***
## territory1206                -898.970    342.414  -2.625  0.00868 **
## territory1207               -1423.631    351.852  -4.046 5.29e-05 ***
## territory1234                -432.020    343.406  -1.258  0.20843
## genderM                      -214.932    197.799  -1.087  0.27726
## ypc                           -42.263      7.886  -5.359 8.73e-08 ***
## cgr_factor                   -551.677    278.806  -1.979  0.04790 *
## age                             2.639      1.199   2.202  0.02773 *
## genderM:cgr_factor            252.425    193.900   1.302  0.19303
## genderM:ypc                    21.518     10.950   1.965  0.04946 *
## territory1206:age              -4.788      1.686  -2.840  0.00453 **
## territory1207:age              -3.226      1.661  -1.943  0.05212 .
## territory1234:age              -1.424      1.594  -0.893  0.37175
## territory1206:cgr_factor  1414.257    337.015   4.196 2.76e-05 ***
## territory1207:cgr_factor  1732.074    330.453   5.242 1.66e-07 ***
## territory1234:cgr_factor   577.615    342.826   1.685  0.09208 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 619.5 on 4994 degrees of freedom
## Multiple R-squared:  0.04241,    Adjusted R-squared:  0.03953
## F-statistic: 14.74 on 15 and 4994 DF,  p-value: < 2.2e-16
```

```r
# AIC and BIC
AIC(model_3)
```

```
## [1] 78653.33
```

```r
BIC(model_3)
```

```
## [1] 78764.15
```

```r
# some interactive terms have p-value > 0.05

# will make an educated decision to exclude gender*cgr_factor interaction
# as it has high p-value and will not improve the model significantly
# also removing territory1234 as its term and interactions have high p-values > 0.05

# age has p-value < 0.05 and will be kept back into model

# AIC and BIC decreased and increased slightly respectively, will continue to monitor

# overall the adjusted p-value increased which is an improvement
```

```r
# removing territory1234 as p-value > 0.05
df5 = df4 %>%
  filter(territory != "1234")

head(df5)
```

```
##   current_premium territory gender ypc cgr_factor age
## 1           98.89      1122      F   0       0.90  66
## 2          695.26      1122      M   0       0.91  66
## 3          656.93      1122      M   0       0.96  66
## 4         2136.57      1122      M   0       1.06  65
## 5          814.66      1122      F   0       1.06  65
## 6         1673.14      1122      F   0       1.05  65
```

```r
model_4 = lm(current_premium ~ territory + ypc + cgr_factor + gender
             + ypc*gender + territory * age + territory*cgr_factor
             , data = df5)
summary_4 = summary(model_4)
summary_4
```

```
##
## Call:
## lm(formula = current_premium ~ territory + ypc + cgr_factor +
##     gender + ypc * gender + territory * age + territory * cgr_factor,
##     data = df5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1374.1  -395.3  -160.3   218.3  4894.1
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1355.312    256.424   5.285 1.33e-07 ***
## territory1206             -869.569    337.232  -2.579  0.00996 **
## territory1207            -1379.420    346.033  -3.986 6.84e-05 ***
## ypc                        -37.729      8.856  -4.260 2.09e-05 ***
## cgr_factor                -387.592    245.865  -1.576  0.11501
## genderM                     48.371     52.416   0.923  0.35616
## age                          2.648      1.184   2.237  0.02535 *
## ypc:genderM                 13.791     12.231   1.128  0.25960
## territory1206:age           -4.789      1.664  -2.878  0.00402 **
## territory1207:age           -3.321      1.639  -2.026  0.04282 *
## territory1206:cgr_factor  1382.962    331.744   4.169 3.13e-05 ***
## territory1207:cgr_factor  1690.352    324.681   5.206 2.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 611.1 on 3669 degrees of freedom
## Multiple R-squared:  0.04724,    Adjusted R-squared:  0.04438
## F-statistic: 16.54 on 11 and 3669 DF,  p-value: < 2.2e-16
```

```r
# AIC and BIC
AIC(model_4)
```

```
## [1] 57689.82
```

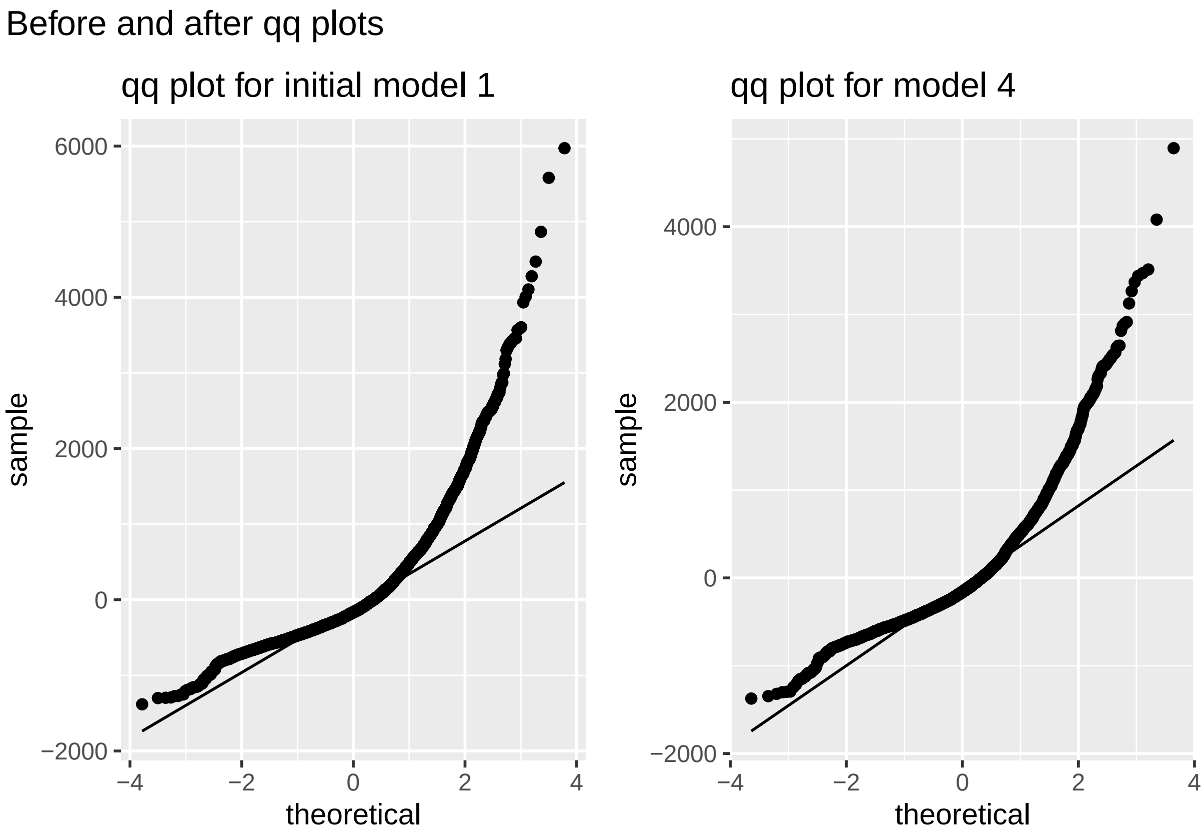
```r
BIC(model_4)
```

```
## [1] 57770.56
```

```r
# adjusted r-squared value has increased
# AIC and BIC has decreased significantly
# overall, model has improved
```

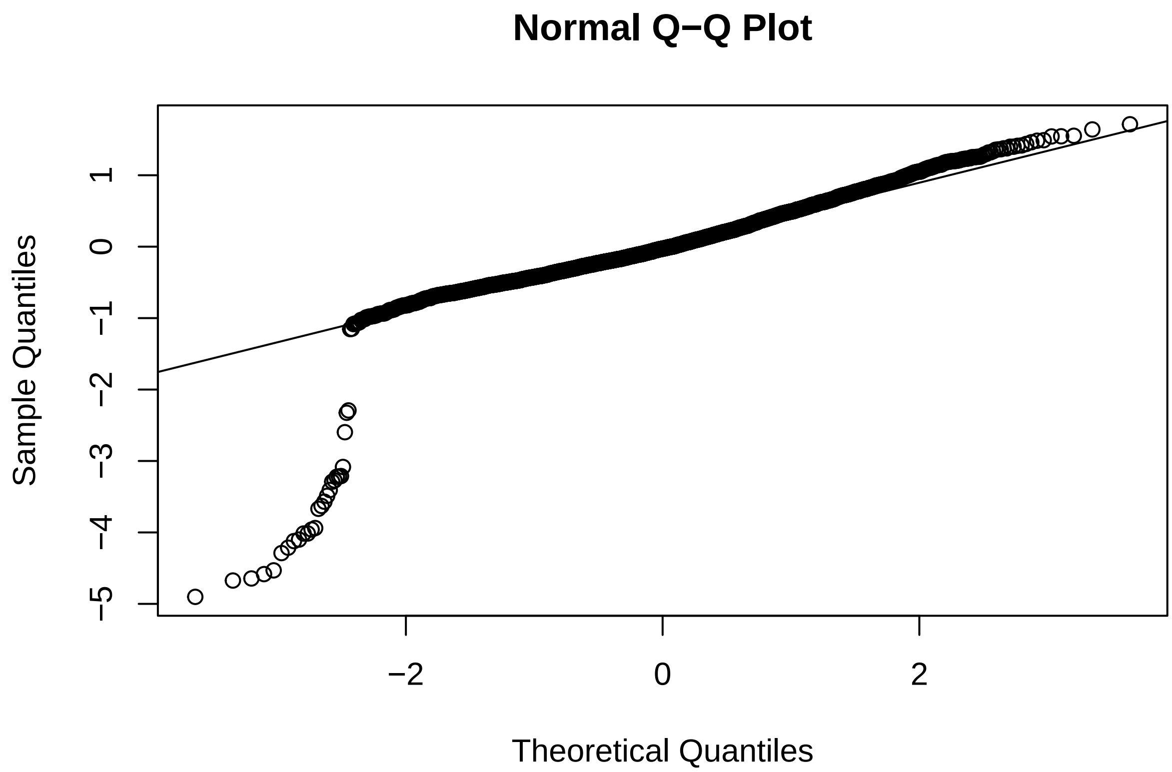
Checking for normality

```r
residuals_model_1 = residuals(model_1)
residuals_model_4 = residuals(model_4)

qq_model_1 = ggplot(data = data.frame(residuals = residuals_model_1),
                    aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "qq plot for initial model 1",
       x = "theoretical",
       y = "sample")

qq_model_4 = ggplot(data = data.frame(residuals = residuals_model_4),
                    aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "qq plot for model 4",
       x = "theoretical",
       y = "sample")

combined_qqs = qq_model_1 + qq_model_4 +
  plot_annotation(title = "Before and after qq plots")

combined_qqs
```




```
# quite similar, some deviation from normal line
# transformation of response variable premiums may improve normality
```

```
df5$transformed_response = log(df5$current_premium)
model_5 <- lm(transformed_response ~ territory + ypc + cgr_factor + gender
              + ypc*gender + territory * age + territory*cgr_factor, data = df5)

qqnorm(residuals(model_5))
qqline(residuals(model_5))
```




```r
# for majority off plot, has improved normality
# except for left tail that is lower than normal line
```

```r
summary(model_5)
```

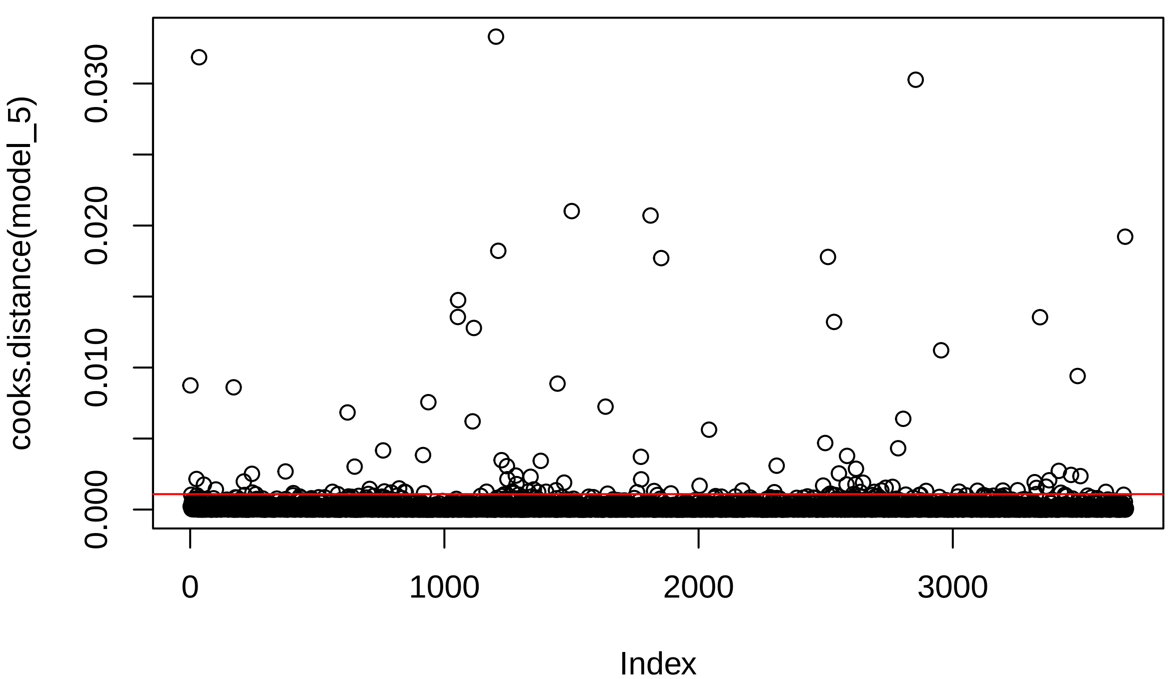
```
##
## Call:
## lm(formula = transformed_response ~ territory + ypc + cgr_factor +
##     gender + ypc * gender + territory * age + territory * cgr_factor,
##     data = df5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9018 -0.2994 -0.0274  0.3032  1.7133
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                6.768682   0.233222  29.023  < 2e-16 ***
## territory1206             -0.218070   0.306718  -0.711 0.477143
## territory1207             -0.503120   0.314722  -1.599 0.109992
## ypc                       -0.028520   0.008055  -3.541 0.000404 ***
## cgr_factor                -0.039978   0.223618  -0.179 0.858121
## genderM                    0.013546   0.047674   0.284 0.776325
## age                        0.002328   0.001077   2.162 0.030690 *
## ypc:genderM                0.017106   0.011124   1.538 0.124200
```

```
## territory1206:age        -0.005196   0.001513  -3.434 0.000602 ***
## territory1207:age        -0.004063   0.001491  -2.725 0.006453 **
## territory1206:cgr_factor  0.737556   0.301726   2.444 0.014554 *
## territory1207:cgr_factor  0.875355   0.295303   2.964 0.003054 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5558 on 3669 degrees of freedom
## Multiple R-squared:  0.04106,    Adjusted R-squared:  0.03818
## F-statistic: 14.28 on 11 and 3669 DF,  p-value: < 2.2e-16
```

```
# however, adjusted r square has decreased significantly
# might not be best method to improve normality
```

Checking for outliers

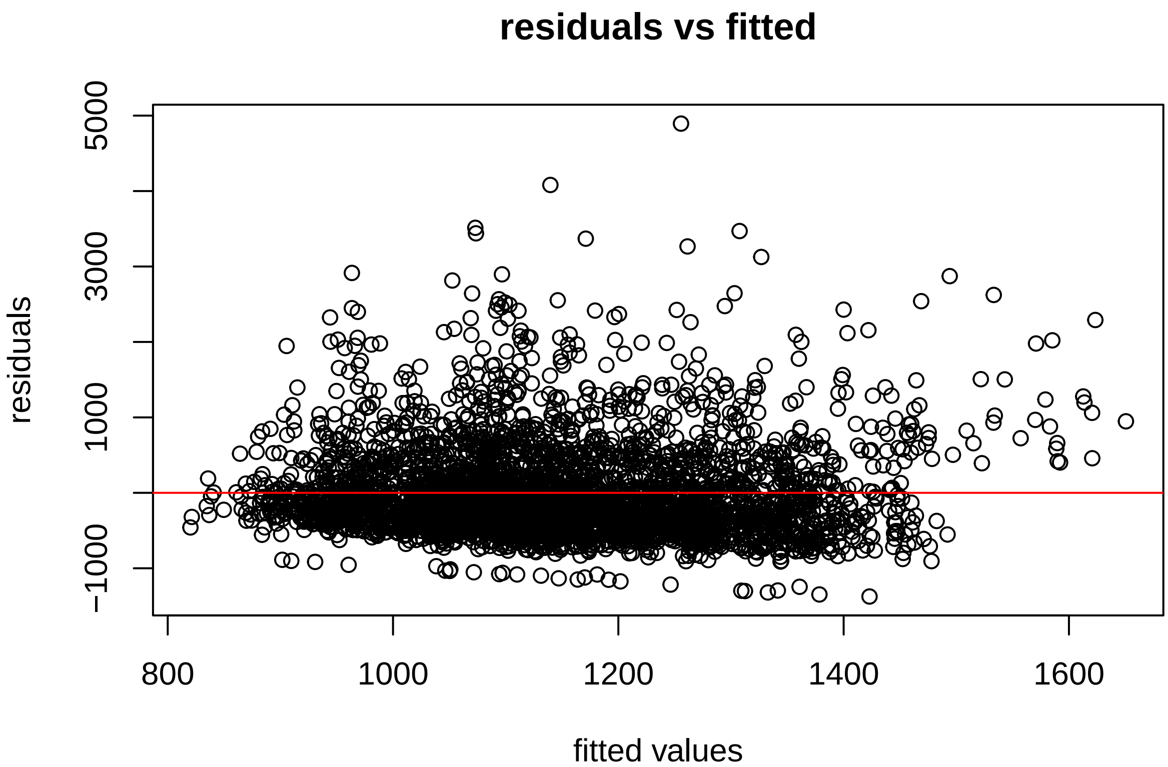
```
# cook's distance to check for influence points
plot(cooks.distance(model_5))
abline(h = 4 / length(df5$transformed_response), col = "red")
```




```
# does not show extreme influence points in model data
```

Checking for homoscedasticity

```r
# residuals vs fitted plot
plot(fitted(model_4), residuals(model_4),
     xlab = "fitted values",
     ylab = "residuals",
     main = "residuals vs fitted")
abline(h = 0, col = "red")
```




```r
# does not show signs of deviating from homoscedasticity
```

Checking for multicollinearity

```r
# vif(model_4)
vif(model_4, type = "predictor")
```

```
##                    GVIF Df GVIF^(1/(2*Df))  Interacts With
## territory      1.152908  8        1.008933 age, cgr_factor
## ypc            1.152908  3        1.023998          gender
## cgr_factor   295.556215  5        1.766298       territory
## gender         1.152908  3        1.023998             ypc
## age        28893.057130  5        2.793054       territory
```

```
##                         Other Predictors
## territory                    ypc, gender
## ypc         territory, cgr_factor, age
## cgr_factor            ypc, gender, age
## gender      territory, cgr_factor, age
## age            ypc, cgr_factor, gender
```

```r
# shows high collinearity for cgr_factor and age
```

```r
# scaling age and cgr_factor to fix collinearity
df6 = df5
df6$age = scale(df6$age, center = TRUE, scale = FALSE)
df6$cgr_factor = scale(df6$cgr_factor, center = TRUE, scale = FALSE)

model_6 = lm(current_premium ~ territory + ypc + cgr_factor + gender
             + ypc*gender + territory * age + territory*cgr_factor
             , data = df6)
```

```r
vif(model_6, type = "predictor")
```

```
##                 GVIF Df GVIF^(1/(2*Df))  Interacts With
## territory  1.152908  8        1.008933 age, cgr_factor
## ypc        1.152908  3        1.023998          gender
## cgr_factor 1.117088  5        1.011134       territory
## gender     1.152908  3        1.023998             ypc
## age        2.264791  5        1.085183       territory
##                         Other Predictors
## territory                    ypc, gender
## ypc         territory, cgr_factor, age
## cgr_factor            ypc, gender, age
## gender      territory, cgr_factor, age
## age            ypc, cgr_factor, gender
```

```r
# improved collinearity issue, gvif values are smaller now
```

```r
summary(model_6)
```

```
##
## Call:
## lm(formula = current_premium ~ territory + ypc + cgr_factor +
##     gender + ypc * gender + territory * age + territory * cgr_factor,
##     data = df6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1374.1  -395.3  -160.3   218.3  4894.1
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1134.597     42.613  26.626  < 2e-16 ***
## territory1206                197.810     28.246   7.003 2.96e-12 ***
## territory1207                 77.253     30.108   2.566  0.01033 *
```

```
## ypc                        -37.729      8.856  -4.260 2.09e-05 ***
## cgr_factor                -387.592    245.865  -1.576  0.11501
## genderM                     48.371     52.416   0.923  0.35616
## age                          2.648      1.184   2.237  0.02535 *
## ypc:genderM                 13.791     12.231   1.128  0.25960
## territory1206:age           -4.789      1.664  -2.878  0.00402 **
## territory1207:age           -3.321      1.639  -2.026  0.04282 *
## territory1206:cgr_factor 1382.962    331.744   4.169 3.13e-05 ***
## territory1207:cgr_factor 1690.352    324.681   5.206 2.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 611.1 on 3669 degrees of freedom
## Multiple R-squared:  0.04724,    Adjusted R-squared:  0.04438
## F-statistic: 16.54 on 11 and 3669 DF,  p-value: < 2.2e-16
```

```r
# AIC and BIC
AIC(model_4)
```

```
## [1] 57689.82
```

```r
BIC(model_4)
```

```
## [1] 57770.56
```

From the coefficients of the final model, significant variables are `territory`, `ypc`, and the interactive variables between `territory` and `age`/`cgr_factor`.

For instance, individuals living in `territory1206` are expected to pay $197 more in premium, which could be due to the location being more prone to car accidents due to poor traffic.

Whereas for `ypc`, for each year of an individual's years of prior coverage they are expected to pay $37 less in premium, likely as they have proven to be reliable and less likely to be at risk of car accidents from their history.

Model has improved AIC from 100919.3 to 57689.78 and BIC from 100987 to 57770.52 which is a significant improvement from initial model with all variables. Adjusted r-squared has also improved from 0.03323 to 0.04439.

Final thoughts: If I were to do it again, I would definitely try to transform the response variable from the start, since it deviated from normality at the extremes. With that, the model may have fit better and the variables chosen in the model may have changes. Just something I've learnt which is the order in which I should take to output more optimal results! :)