Regression Analysis of Medical Insurance Costs

MATH1312 Regression Analysis

Abhishekh Shankar s3652116

Contents

1 Introduction

Medical insurance and the costs associated with it is an important aspect every day life, and as such, it is essential that companies price it accurately to the individual. The objective of this project is to apply a number of regression models, and evaluate the best model to fit to the data set. The modelling takes into account a number of variables such as an individual's age and body mass index (bmi) and outputs a figure which is optimised to be the insurance costs to the individual. The data set is sourced from the online community Kaggle,[1] which takes reference from Lantz's book "Machine Learning With R" (Lantz 2013).

2 Data Exploration

2.1 Data Set

The data set contains 1338 instances with 6 regressor variables and 1 response variable as summarised below.

- `age`: Age of an individual, continuous.
- `sex`: Sex of an individual, categorical – male, female.
- `BMI`: Body Mass Index of an individual, continuous.
- `children`: Number of children an individual has, continuous.
- `smoking status`: Smoking status of an individual, categorical – yes, no.
- `region`: Region where an individual lives, categorical – southwest, southeast, northwest, northeast.
- `charges`: **Response variable**, individual medical insurance costs, continuous.

2.2 Descriptive Statistics

It is assumed that certain variables in the data set contribute to a larger effect than others when determining the insurance charges. According to (Medicare and Services, n.d.), "the oldest adult who uses tobacco may be charged up to 4.5 times more than the youngest adult who does not", so it is assumed that the smoking status and age of individuals contribute significantly on the insurance charges. We explore this through the scatter plot below.

```
insurance <- read.csv("insurance.csv")
ggplot(insurance, aes(x=charges, y=age, color=smoker)) + geom_point() +
  labs(title = 'Plot of Charges Against Age By Smoking Status')
```

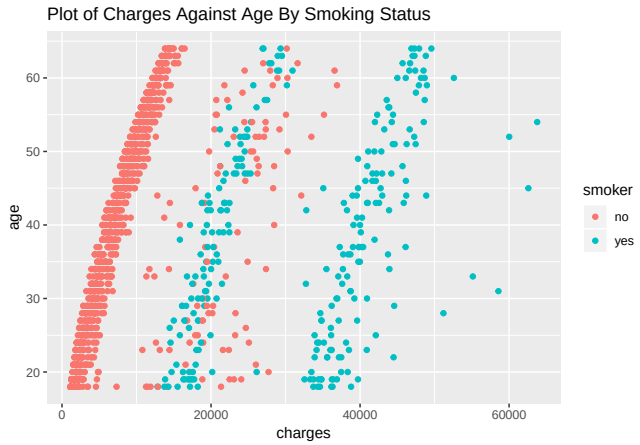


Figure 1: Scatter plot of charges against age by smoking status

[1] https://www.kaggle.com/mirichoi0218/insurance

It can be observed that the insurance costs tend to be much larger for individuals who smoke, even for those that are younger. Furthermore, the cluster of non-smokers shown in the left-most portion of the plot shows that as age increases, the costs to the individual tend to increase also. This suggests that both the smoking status and age of the individual are useful predictors to their insurance costs, as hypothesised.

It is also hypothesised that the `region` variable in the data does not play a large role in determining an individual's medical insurance charges and it can be seen from the scatter plot below that there does not seem to be a correlation as the charges are seemingly distributed equally amongst the different regions.

```
ggplot(insurance, aes(x=charges, y=region)) + geom_point() +
  labs(title = 'Plot of Charges By Region')
```

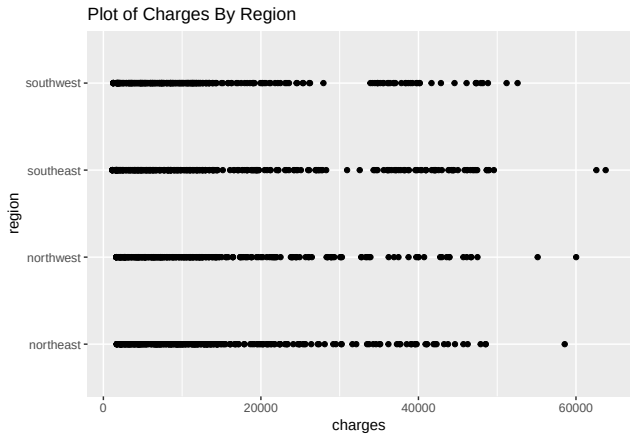


Figure 2: Scatter plot of charges by region

3 Modelling

First, we define the following variables:

$$\begin{aligned} \text{age} &= x_1 \\ \text{sex} &= x_2 \\ \text{bmi} &= x_3 \\ \text{children} &= x_4 \\ \text{smoker} &= x_5 \\ \text{region} &= x_6 \end{aligned}$$

3.1 Backward Elimination

The first model to be fit to the data is computed through backward elimination, where a regression model utilising every predictor variable is initially fit, and variable significance through an F-test is completed in order to determine which variable to eliminate from the model. Appendix One shows the entirety of the R code and outputs for the backward elimination process.

We define a critical value $F_{stay} = F_{0.05,1,n-p} = F_{0.05,1,1331} = 3.85$ which represents the value with which a variable is to be removed from the model. The results of the first iteration shown in Appendix One indicate that the `age`, `bmi`, `children`, and `smoker` variables are all highly significant to a 5% level of signifiance. The `sex`, and `region` variables are shown to not be significant since $p > 0.05$ for both of those variables. Since `sex` is shown to have the largest p-value with $p = 0.6933$, and an F value less than F_{stay}, it is removed from the model.

The following iteration shows `age`, `bmi`, `children`, and `smoker` are all still highly significant at the 5% level of significance, and since `region` is insignificant and has the highest p-value with $p = 0.096$. Alongside this, with an F-value of $2.1166 < F_{stay} = 3.85$, we remove this variable from the final model. After these iterations, the remaining variables are all significant and the backward elimination model building process is completed. The final model summary is shown below.

```
back.model <- lm(charges ~ age + bmi + children + factor(smoker), insurance)
summary(back.model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + factor(smoker),
##     data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.9  -2920.8   -986.6   1392.2  29509.6
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -12102.77     941.98 -12.848  < 2e-16 ***
## age                  257.85      11.90  21.675  < 2e-16 ***
## bmi                  321.85      27.38  11.756  < 2e-16 ***
## children             473.50     137.79   3.436 0.000608 ***
## factor(smoker)yes  23811.40     411.22  57.904  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6068 on 1333 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7489
## F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

From this, the final model is as follows: $\hat{y} = -12102.77 + 257.85x_1 + 321.85x_3 + 473.50x_4 + 23811.40x_5$

3.2 Forward Selection

The second method of regression model building is completed via forward selection. A null model is initially defined and the significance of all the variables available to be added is tested. Appendix Two showcases the entire code outputs for the forward selection method and it can be seen in the first iteration that the most significant variable is the `smoker` varable and with the highest F-value that exceeds $F_{in} = 3.85$, it is added to the model.

The next iteration shows that the `age` variable is the most significant variable and is added to the model. Further iterations show that `bmi` and `children` are the last two variables which show significance and an F-value that is greater than F_{in} and so are added to the final model. Iterations beyond this show no more of the remaining variables whose F-values exceed the threshold to be able to be placed in the model. The final summary is shown below:

```
forward.model <- lm(charges ~ age + bmi + children + factor(smoker), insurance)
summary(forward.model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + factor(smoker),
##     data = insurance)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.9  -2920.8   -986.6   1392.2  29509.6
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -12102.77     941.98 -12.848  < 2e-16 ***
## age                  257.85      11.90  21.675  < 2e-16 ***
## bmi                  321.85      27.38  11.756  < 2e-16 ***
## children             473.50     137.79   3.436 0.000608 ***
## factor(smoker)yes  23811.40     411.22  57.904  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6068 on 1333 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7489
## F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

The final model is: $\hat{y} = -12102.77 + 257.85x_1 + 321.85x_3 + 473.50x_4 + 23811.40x_5$

The model found through backward elimination is equal to the model computed through forward selection.

3.3 Best Subsets Regression

The final type of regression model building is completed through best subsets regression which is using the adjusted R^2 value as a performance indicator. With a number of categorical variables present in the data, using the best subsets regression as a model builder presents the possiblilty of certain levels of a categorical variable being significant with the remaining being insignificant. As a result, for this analysis, should the majority of levels in a categorical variable be significant, that variable is considered for the model.

```
r <- leaps::regsubsets(charges ~ age + sex + bmi + children + factor(smoker) + factor(region),
                       data = insurance)
plot(r, scale='adjr2')
```

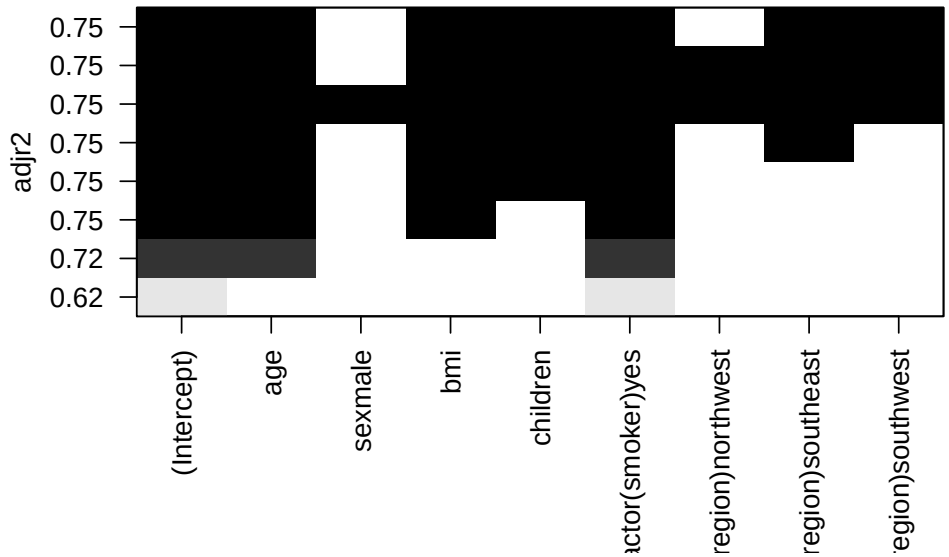


Figure 3: All subsets regression results with adjusted R-squared.

Shown above are six different models which share an adjusted R^2 of 0.75 and thus, the best subsets regression is computed again with C_p as a performance indicator in an attempt to find one discernable model. The plot shown below shows that with a C_p value of 5.7, a model without the `sex` variable is ideal. The categorical `region` variable is shown to be significant in two of its three defined levels and as such, is chosen to be significant in the model.

```
r <- leaps::regsubsets(charges ~ age + sex + bmi + children + factor(smoker) + factor(region),
                       data = insurance)
plot(r, scale='Cp')
```

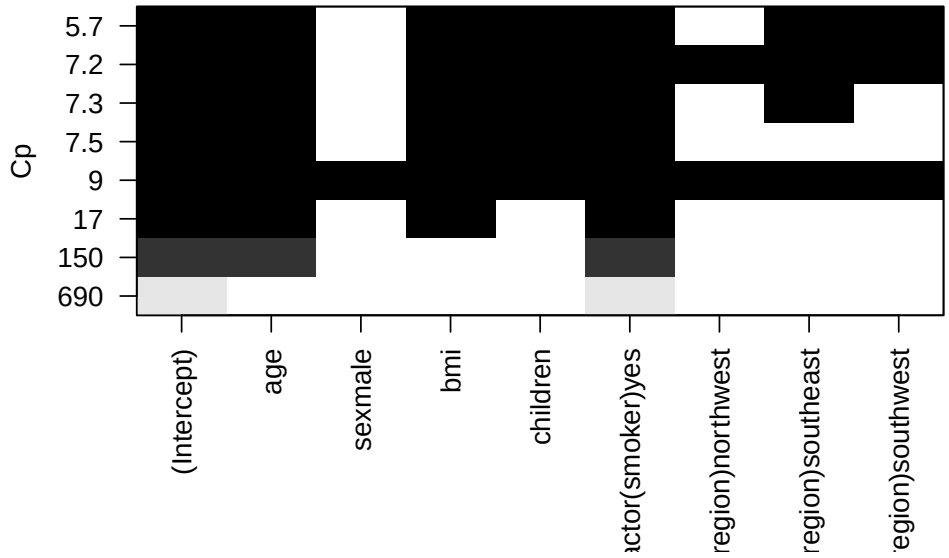


Figure 4: All subsets regression results with Cp

From this, the model summary is shown below.

```
best.sub.model <- lm(charges ~ age + bmi + children + factor(smoker) + factor(region), insurance)
summary(best.sub.model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + factor(smoker) +
##     factor(region), data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11367.2  -2835.4   -979.7   1361.9  29935.5
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -11990.27     978.76 -12.250  < 2e-16 ***
## age                         256.97      11.89  21.610  < 2e-16 ***
## bmi                         338.66      28.56  11.858  < 2e-16 ***
## children                    474.57     137.74   3.445 0.000588 ***
## factor(smoker)yes         23836.30     411.86  57.875  < 2e-16 ***
## factor(region)northwest    -352.18     476.12  -0.740 0.459618
## factor(region)southeast   -1034.36     478.54  -2.162 0.030834 *
## factor(region)southwest    -959.37     477.78  -2.008 0.044846 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7496
## F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

The final model is:

$$\hat{y} = \begin{cases} -11990.27 + 256.97x_1 + 338.66x_2 + 474.57x_3 + 23836.30x_4 - 352.18x_5, & \text{if region} = \text{Northwest} \\ -11990.27 + 256.97x_1 + 338.66x_2 + 474.57x_3 + 23836.30x_4 - 1034.36x_5, & \text{if region} = \text{Southeast} \\ -11990.27 + 256.97x_1 + 338.66x_2 + 474.57x_3 + 23836.30x_4 - 959.37x_5, & \text{if region} = \text{Southwest} \end{cases}$$

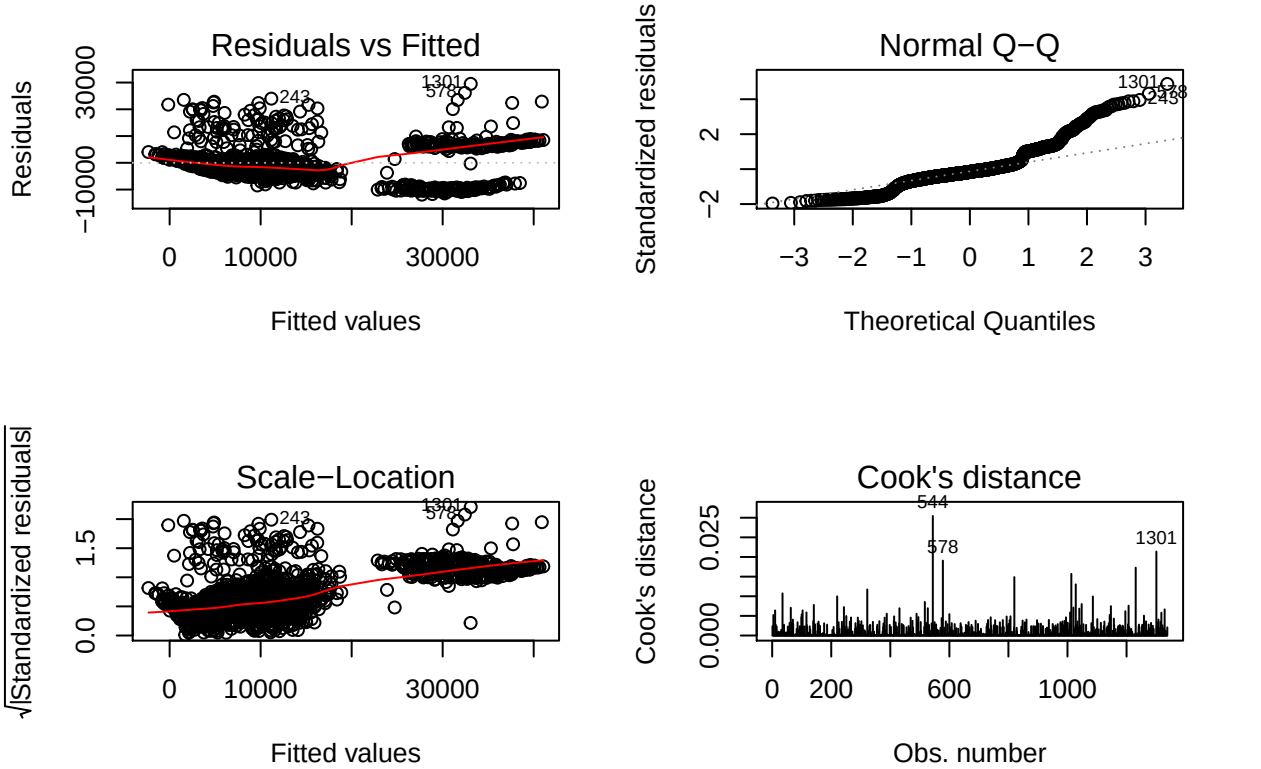
4 Model Evaluation

Considering both the backwards elimination and forwards selection model building resulted in the same model, all of their model evaluation results will be grouped together.

4.1 Backward Elimination/Forward Selection Evaluation

First, an analysis into the residuals is conducted. The plot in the top left shows the residuals amongst the fitted values and it can be observed that there are large numbers of residuals positioned on either side of the zero mean mark. From the plot, there does not appear to be a non-linear relationship amongst the residuals.

```
par(mfrow = c(2,2))
plot(back.model, which = c(1,2,3,4))
```



The QQ-plot in the top right shows a skewed, heavy tailed distribution and there are a large number of points that do not fall onto the line of normality. As a result, it does not appear that the data is normally distributed.

The plot on the bottom left shows the Spread-Location plot and it can be observed that the red line is not horizontal, indicating that the data may not have an equal variance.

Finally, the bottom right plot shows that there are no instances of residuals that lie outside of Cook's Distances and as such, there does not seem to be evidence that outliers in the data are influential to the regression results.

The Durbin-Watson test below shows that with a p-value of $p > 0.05$, there is insufficient evidence to reject the null hypothesis, therefore, this implies that the uncorrelated error assumption has not been violated.

```
car::durbinWatsonTest(back.model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.04506494      2.087394   0.108
##  Alternative hypothesis: rho != 0
```

The Shapiro-Wilk test shows a highly signifiant p-value to a 5% level of significance, therefore, there is sufficient evidence to reject the null hypothesis that the normality error assumption has not been violated.

```
stdres <- rstudent(back.model)
shapiro.test(stdres)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  stdres
## W = 0.89904, p-value < 2.2e-16
```

The non-constant variance test below was seen to be highly significant to a 5% level of significance, indicating that there is sufficient evidence to reject the null hypothesis that the residuals' constant error assumption has not been violated.

```
car::ncvTest(back.model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 230.5625, Df = 1, p = < 2.22e-16
```

Variance Inflation Factor (VIF) values are used to assess multicolinearity and as shown below, all of the variables used in the model show VIF values very close to 1.00, indicating that there does not appear to be any significant multicollinearity amongst the variables.

```
car::vif(back.model)
```

```
##            age            bmi       children factor(smoker)
##       1.014498       1.012194       1.001950       1.000745
```

4.2 Best Subsets Regression Evaluation

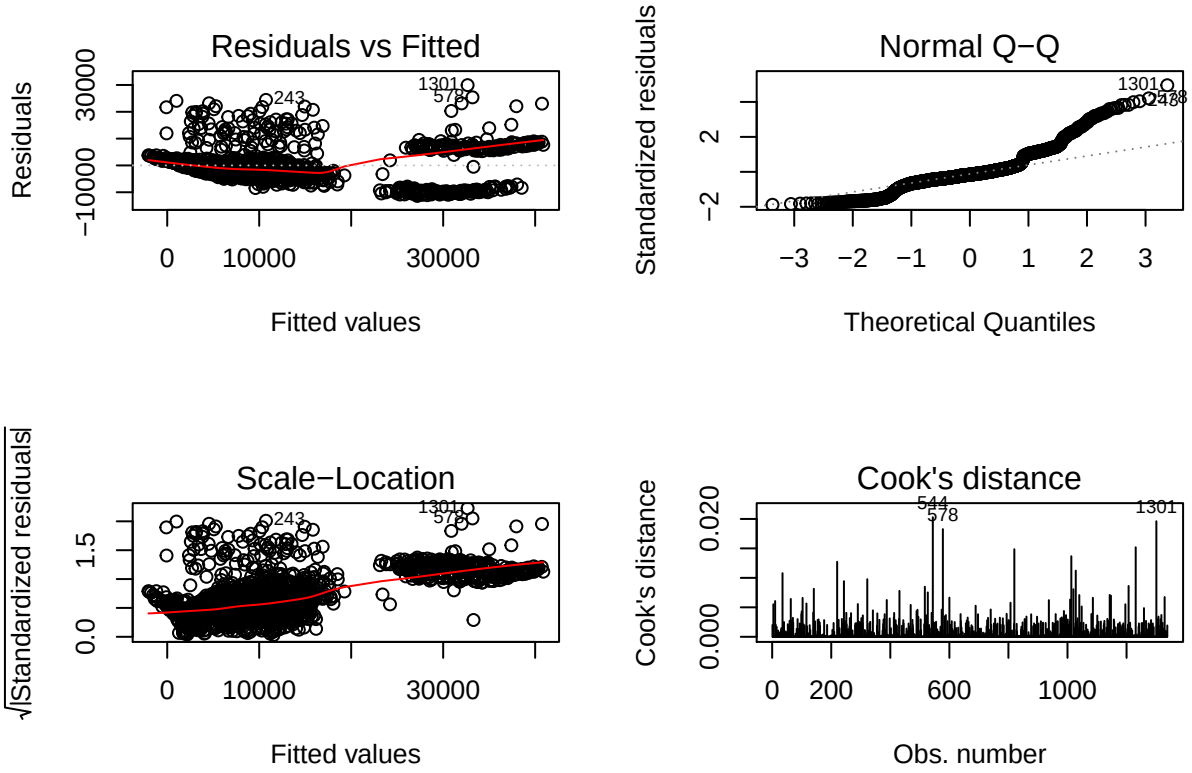
The residuals amongst the fitted values in the top left shows that there are large numbers of residuals positioned on either side of the zero mean mark, indicating that there does not appear to be a non-linear relationship amongst the residuals.

The QQ-plot in the top right shows a skewed, heavy tailed distribution similar to that of the QQ plot computed through the backwards elimination and forward selection. Since there are a large number of points that do not fall onto the line of normality, it appears the data is not normally distributed.

The plot on the bottom left shows the Spread-Location plot and similar to the results shown above with the backward elimination and forward selection methods it can be observed that the red line is not horizontal, indicating that the data may not have an equal variance.

Finally, the bottom right plot shows that there are no instances of residuals that lie outside of Cook's Distances and as such, there does not seem to be evidence that outliers in the data are influential to the regression results.

```
par(mfrow = c(2,2))
plot(best.sub.model, which = c(1,2,3,4))
```



The Durbin-Watson test below shows that with a p-value of $p > 0.05$, there is insufficient evidence to reject the null hypothesis, therefore, this implies that the uncorrelated error assumption has not been violated.

```
car::durbinWatsonTest(best.sub.model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.04582739      2.088964   0.118
##  Alternative hypothesis: rho != 0
```

The Shapiro-Wilk test results were seen to be highly signifiant to a 5% level of significance, therefore, there is sufficient evidence to reject the null hypothesis that the normality error assumption has not been violated.

```
stdres <- rstudent(best.sub.model)
shapiro.test(stdres)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  stdres
## W = 0.89854, p-value < 2.2e-16
```

The non-constant variance test below was seen to be highly significant to a 5% level of significance. This

indicates that there is sufficient evidence to reject the null hypothesis that the residuals' constant error assumption has not been violated.

```
car::ncvTest(best.sub.model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 235.2819, Df = 1, p = < 2.22e-16
```

The VIF values of all of the variables used in the model show values very close to 1.00, indicating that there does not appear to be any significant multicollinearity amongst the variables.

```
car::vif(best.sub.model)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## age             1.016188  1        1.008061
## bmi             1.104197  1        1.050808
## children        1.003714  1        1.001855
## factor(smoker)  1.006369  1        1.003179
## factor(region)  1.098870  3        1.015838
```

4.3 ANOVA Comparisons

The ANOVA comparison of the two potential models is shown below. The p-value is seen to be insignificant to a 5% level of significance, indicating that the addition of the `region` variable does not lead to a significantly improved fit.

```
anova(back.model, best.sub.model)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + bmi + children + factor(smoker)
## Model 2: charges ~ age + bmi + children + factor(smoker) + factor(region)
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1   1333 4.9078e+10
## 2   1330 4.8845e+10  3 233200844 2.1166 0.09631 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5 Discussion

Three different regression model building techniques were implemented: backwards elimination, forward selection, and two separate best subsets regression models using both an adjusted R^2 and a C_p value as performance indicators of each model in an attempt to develop an optimal regression model to the data. Results of the backwards elimination and forwards selection methods resulted in the same final model.

An initial look into the accuracy of these models showed that each model had very similar adjusted R^2 values, with 0.7489 and 0.7496 for the backwards elimination and best subsets regression model respectively, indicating that both of these models account for approximately 75% of the variation in the data. Considering

A linear regression model has the following assumptions:

- A linear relationship between the response variable and the regressors.
- The error term has 0 mean.
- The error term has a constant variance, σ^2.

- The errors are uncorrelated.
- The errors are normally distributed.

These assumptions were tested through a residual analysis which was conducted on the potential models. A look into linearity showed that the residuals for both models are seemingly uncorrelated, indicating a linear relationship. The errors were shown to be uncorrelated through the inability to reject the null hypothesis of the Durbin-Watson test. The QQ-plots of both models was shown to be non-normal since a large number of data points were skewed from the line of normality. This was supported by the results of the Shapiro-Wilk test which was found to be highly significant, indicating that the normality assumption had been violated. The non-constant variance error tests were conducted and in the cases of both models was highly significant, thereby rejecting the null hypothesis that the residuals' constant error variable has not been violated. Finally, the non-constant variance test showed that this assumption had been violated for both models.

Alongside a residual analysis, a look into any potential multicollinearity in the two model was investigated. VIF values for both models were very close 1.00, indicating that there does not appear to be any significant multicollinearity effects in the model amongst the variables.

A final comparison of the two potential models was conducted through an ANOVA test, where the addition of the `region` variable was seen to be insignificant to improving the fit of the model. With this result alongside the residual analysis and the computation of the adjusted R^2 values, it is recommended that the `region` varible be excluded from the final model. As such, the final model is as follows: $\hat{y} = -12102.77 + 257.85x_1 + 321.85x_3 + 473.50x_4 + 23811.40x_5$.

6 Conclusion

An initial look into the research surrounding medical insurance costs suggested that certain variables contribute to higher costs for the individual, such as their smoking status or age, and a preliminary investigation into the data supported these statements.

Fitting an optimal regression model to the data was completed through the utilisation of three different model building techniques: backward elimination, forward selection, and best subsets regression. Result showed that the backwards elimination and forward selection algorithms converged to the same ideal model, utilising `age`, `bmi`, `children`, and `smoker` as independent variables, while the best subsets regression model included all these alongside the `region` variable.

Residual analysis and a look into multicollinearity on the potential models showed very similar results and no obvious optimal model. As such, an ANOVA test was conducted to determine whether the inclusion of the `region` variable was significant. Test results indicated that the addition of the `region` variable was insignificant to the fit of the model and as such, the optimal model was deemed to be the result of the backwards elimination.

Appendix One - Backward Elimination R Code

```
full.model <- lm(charges ~ age + factor(sex) + bmi + children + factor(smoker) +
                   factor(region), insurance)
drop1(full.model, test = "F")
```

```
## Single term deletions
##
## Model:
## charges ~ age + factor(sex) + bmi + children + factor(smoker) +
##     factor(region)
##                 Df  Sum of Sq        RSS   AIC   F value    Pr(>F)
## <none>                        4.8840e+10 23316
## age              1 1.7124e+10 6.5964e+10 23717  465.9837 < 2.2e-16 ***
## factor(sex)      1 5.7164e+06 4.8845e+10 23315    0.1556  0.693348
## bmi              1 5.1692e+09 5.4009e+10 23449  140.6627 < 2.2e-16 ***
## children         1 4.3755e+08 4.9277e+10 23326   11.9063  0.000577 ***
## factor(smoker)   1 1.2245e+11 1.7129e+11 24993 3331.9680 < 2.2e-16 ***
## factor(region)   3 2.3343e+08 4.9073e+10 23317    2.1173  0.096221 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(update(full.model, ~ . -factor(sex)), test = "F")
```

```
## Single term deletions
##
## Model:
## charges ~ age + bmi + children + factor(smoker) + factor(region)
##                 Df  Sum of Sq        RSS   AIC   F value    Pr(>F)
## <none>                        4.8845e+10 23315
## age              1 1.7151e+10 6.5996e+10 23715  466.9969 < 2.2e-16 ***
## bmi              1 5.1645e+09 5.4010e+10 23447  140.6226 < 2.2e-16 ***
## children         1 4.3596e+08 4.9281e+10 23324   11.8706  0.000588 ***
## factor(smoker)   1 1.2301e+11 1.7186e+11 24996 3349.5461 < 2.2e-16 ***
## factor(region)   3 2.3320e+08 4.9078e+10 23315    2.1166  0.096315 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
drop1(update(full.model, ~ . -factor(sex) - factor(region)), test = "F")
```

```
## Single term deletions
##
## Model:
## charges ~ age + bmi + children + factor(smoker)
##                 Df  Sum of Sq        RSS   AIC  F value    Pr(>F)
## <none>                        4.9078e+10 23315
## age              1 1.7297e+10 6.6375e+10 23717  469.789 < 2.2e-16 ***
## bmi              1 5.0884e+09 5.4167e+10 23445  138.203 < 2.2e-16 ***
## children         1 4.3477e+08 4.9513e+10 23325   11.809 0.0006077 ***
## factor(smoker)   1 1.2345e+11 1.7253e+11 24995 3352.911 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix Two - Forward Selection R Code

```
model.null <- lm(charges ~ 1, data = insurance)
add1(model.null, scope = ~ age + factor(sex) + bmi + children + factor(smoker) +
                   factor(region), test = "F")
```

```
## Single term additions
##
## Model:
## charges ~ 1
##                 Df  Sum of Sq        RSS   AIC   F value    Pr(>F)
## <none>                       1.9607e+11 25160
## age              1 1.7530e+10 1.7854e+11 25037  131.1740 < 2.2e-16 ***
## factor(sex)      1 6.4359e+08 1.9543e+11 25158    4.3997   0.03613 *
## bmi              1 7.7134e+09 1.8836e+11 25108   54.7093 2.459e-13 ***
## children         1 9.0660e+08 1.9517e+11 25156    6.2060   0.01285 *
## factor(smoker)   1 1.2152e+11 7.4554e+10 23868 2177.6149 < 2.2e-16 ***
## factor(region)   3 1.3008e+09 1.9477e+11 25157    2.9696   0.03089 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(update(model.null, ~ . +factor(smoker)), scope = ~ age + factor(sex) + bmi + children +
       factor(smoker) + factor(region), test = "F")
```

```
## Single term additions
##
## Model:
## charges ~ factor(smoker)
##                 Df  Sum of Sq        RSS   AIC  F value    Pr(>F)
## <none>                       7.4554e+10 23868
## age              1 1.9928e+10 5.4626e+10 23454 487.0225 < 2.2e-16 ***
## factor(sex)      1 1.4213e+06 7.4553e+10 23870   0.0255 0.8732723
## bmi              1 7.4856e+09 6.7069e+10 23729 148.9997 < 2.2e-16 ***
## children         1 7.5272e+08 7.3802e+10 23857  13.6160 0.0002333 ***
## factor(region)   3 1.0752e+08 7.4447e+10 23872   0.6417 0.5882074
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
add1(update(model.null, ~ . +factor(smoker) + age), scope = ~ age + factor(sex) + bmi +
       children + factor(smoker) + factor(region), test = "F")
```

```
## Single term additions
##
## Model:
## charges ~ factor(smoker) + age
##                 Df  Sum of Sq        RSS   AIC  F value    Pr(>F)
## <none>                       5.4626e+10 23454
## factor(sex)      1    2225509 5.4624e+10 23456   0.0544 0.8156949
## bmi              1 5112896646 4.9513e+10 23325 137.7532 < 2.2e-16 ***
## children         1  459283727 5.4167e+10 23445  11.3111 0.0007923 ***
## factor(region)   3  138426748 5.4488e+10 23457   1.1280 0.3365350
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
add1(update(model.null, ~ . +factor(smoker) + age + bmi), scope = ~ age + factor(sex) + bmi +
       children + factor(smoker) + factor(region), test = "F")
```

```
## Single term additions
##
## Model:
## charges ~ factor(smoker) + age + bmi
##                Df Sum of Sq        RSS   AIC F value    Pr(>F)
## <none>                      4.9513e+10 23325
## factor(sex)     1   3942912 4.9509e+10 23327  0.1062 0.7446102
## children        1 434769398 4.9078e+10 23315 11.8086 0.0006077 ***
## factor(region)  3 232012208 4.9281e+10 23324  2.0887 0.0998905 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
add1(update(model.null, ~ . +factor(smoker) + age + bmi + children), scope = ~ age + factor(sex) +
       bmi + children + factor(smoker) + factor(region), test = "F")
```

```
## Single term additions
##
## Model:
## charges ~ factor(smoker) + age + bmi + children
##                Df Sum of Sq        RSS   AIC F value  Pr(>F)
## <none>                      4.9078e+10 23315
## factor(sex)     1   5486063 4.9073e+10 23317  0.1489 0.69964
## factor(region)  3 233200844 4.8845e+10 23315  2.1166 0.09631 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix Three - R Code Session

```
# Import necessary libraries.
library(ggplot2)

# Read in the data.
insurance <- read.csv("insurance.csv")

# Plot the charges by age and smoking status.
ggplot(insurance, aes(x=charges, y=age, color=smoker)) + geom_point() +
  labs(title = 'Plot of Charges Against Age By Smoking Status')

# Plot the charges against the region.
ggplot(insurance, aes(x=charges, y=region)) + geom_point() +
  labs(title = 'Plot of Charges By Region')

# Fit the backwards elimination model.
full.model <- lm(charges ~ age + factor(sex) + bmi + children + factor(smoker) +
                   factor(region), insurance)
drop1(full.model, test = "F")
drop1(update(full.model, ~ . -factor(sex)), test = "F")
drop1(update(full.model, ~ . -factor(sex) - factor(region)), test = "F")

back.model <- lm(charges ~ age + bmi + children + factor(smoker), insurance)
summary(back.model)

# Fit the forwards selection model.
model.null <- lm(charges ~ 1, data = insurance)
add1(model.null, scope = ~ age + factor(sex) + bmi + children + factor(smoker) +
                   factor(region), test = "F")

add1(update(model.null, ~ . +factor(smoker)), scope = ~ age + factor(sex) + bmi + children +
       factor(smoker) + factor(region), test = "F")
add1(update(model.null, ~ . +factor(smoker) + age), scope = ~ age + factor(sex) + bmi +
       children + factor(smoker) + factor(region), test = "F")
add1(update(model.null, ~ . +factor(smoker) + age + bmi), scope = ~ age + factor(sex) + bmi +
       children + factor(smoker) + factor(region), test = "F")
add1(update(model.null, ~ . +factor(smoker) + age + bmi + children), scope = ~ age + factor(sex) +
       bmi + children + factor(smoker) + factor(region), test = "F")

forward.model <- lm(charges ~ age + bmi + children + factor(smoker), insurance)
summary(forward.model)

# Fit a best subsets regression model.
# adjusted R^2
r <- leaps::regsubsets(charges ~ age + sex + bmi + children + factor(smoker) + factor(region),
                       data = insurance)
plot(r, scale='adjr2')

# Cp
r <- leaps::regsubsets(charges ~ age + sex + bmi + children + factor(smoker) + factor(region),
                       data = insurance)
plot(r, scale='Cp')
```

```r
# Model Diagnostics
par(mfrow = c(2,2))
plot(back.model, which = c(1,2,3,4))

# Durbin-Watson Test.
car::durbinWatsonTest(back.model)

# Shapiro-Wilk Test.
stdres <- rstudent(best.sub.model)
shapiro.test(stdres)

# Non-constant variance test.
car::ncvTest(best.sub.model)

#VIF Values.
car::vif(best.sub.model)

# ANOVA comparisons.
anova(back.model, best.sub.model)
```

References

Lantz, Brett. 2013. *Machine Learning with R.* Packt Publishing Ltd.

Medicare, Centres for, and Medicaid Services. n.d. "Overview: Final Rule for Health Insurance Market Reforms." Available at https://www.cms.gov/CCIIO/Resources/Files/Downloads/market-rules-technical-summary-2-27-2013.pdf.