Regression Final Assignment

Predictive Modeling of Insurance Charges Using Patient Demographics and Health Indicators

MATH1312 Regression Analysis

PGRD Semester 1 2024 (2410)

Due 2024-06-13

Saurabh Tyagi (3988015)

Adyasaa Mohapatra (3988759)

Amay Iyer (s3970066)

Students' Contributions

Saurabh Tyagi, Adyasaa Mohapatra and Amay Iyer conceived the ideas and designed the methodology; Saurabh Tyagi, Adyasaa Mohapatra and Amay Iyer collected the data from reputable sources as outlined in the project requirements; Saurabh Tyagi, Adyasaa Mohapatra and Amay Iyer conducted the exploratory data analysis and built the regression models using R and RStudio; Saurabh Tyagi, Adyasaa Mohapatra and Amay Iyer assessed the model performance and performed diagnostic checks. Saurabh Tyagi, Adyasaa Mohapatra and Amay Iyer led the writing of the manuscript, ensuring all components specified in the project guidelines were thoroughly covered. All students contributed critically to the report and gave final approval for submission.

We agree and acknowledge that:

1. We have read and understood the Declaration and Statement of Authorship.

2. If we do not agree to the Declaration and Statement of Authorship in this context and a signature is not included below, the assessment outcome is not valid for assessment purposes and will not be included in my final result for this course.

Styagi Adyasaa Mohapatra

Exploratory Data Analysis (EDA)

This dataset contains information on individuals' demographics, lifestyle choices, and medical charges. It includes variables such as age, sex, BMI, number of children, smoking status, region, and medical charges. Medical Charges is our target variable.

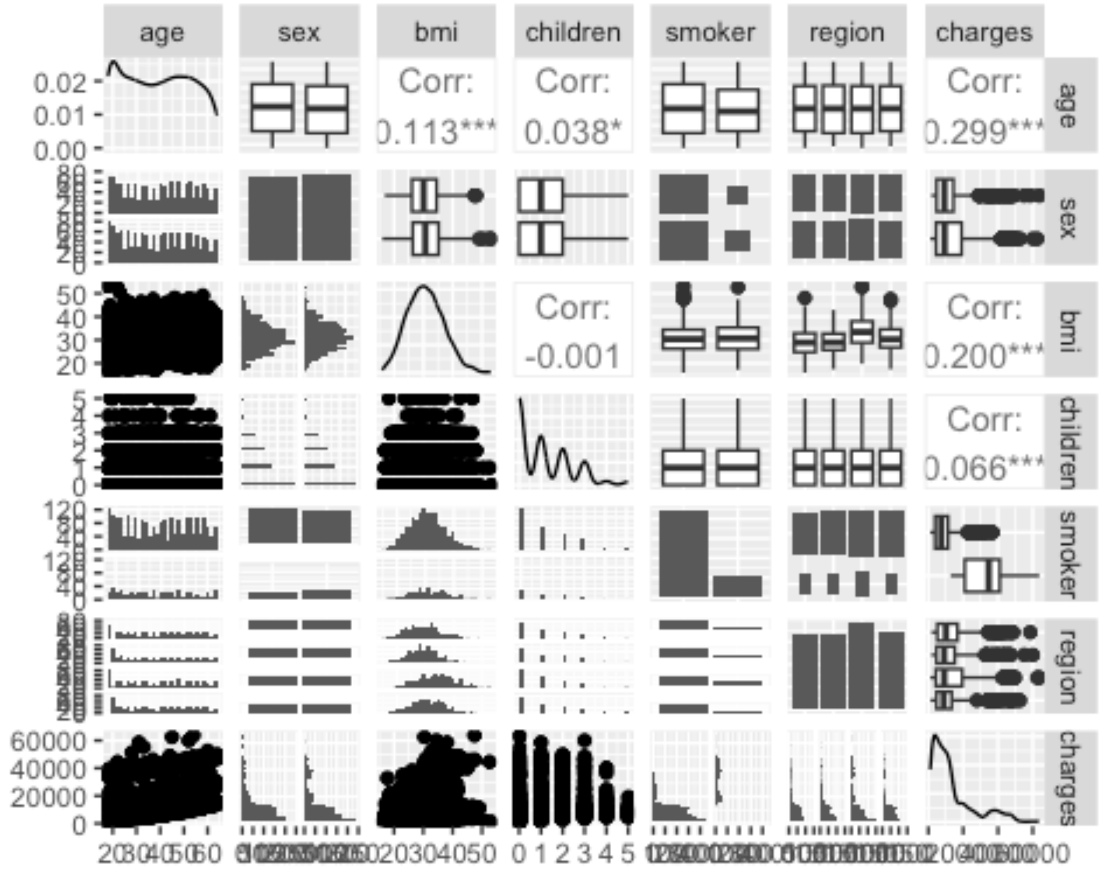
```
# Loading dataset
Insurance <- read_csv("/Users/saurabhtyagi/Downloads/medical_insurance.csv")
head(Insurance)

## # A tibble: 6 × 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    19 female  27.9        0 yes    southwest  16885.
## 2    18 male    33.8        1 no     southeast   1726.
## 3    28 male    33          3 no     southeast   4449.
## 4    33 male    22.7        0 no     northwest  21984.
## 5    32 male    28.9        0 no     northwest   3867.
## 6    31 female  25.7        0 no     southeast   3757.

#Cheking for missing values
sum(is.na(Insurance))

## [1] 0
```

No missing values.

```
summary(Insurance)

##       age            sex                 bmi           children    
##  Min.   :18.00   Length:2772        Min.   :15.96   Min.   :0.000  
##  1st Qu.:26.00   Class :character   1st Qu.:26.22   1st Qu.:0.000  
##  Median :39.00   Mode  :character   Median :30.45   Median :1.000  
##  Mean   :39.11                      Mean   :30.70   Mean   :1.102  
##  3rd Qu.:51.00                      3rd Qu.:34.77   3rd Qu.:2.000  
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000  
##     smoker             region             charges     
##  Length:2772        Length:2772        Min.   : 1122  
##  Class :character   Class :character   1st Qu.: 4688  
##  Mode  :character   Mode  :character   Median : 9333  
##                                        Mean   :13261  
##                                        3rd Qu.:16578  
##                                        Max.   :63770  
```

The dataset consists of 2772 observations with variables including age (ranging from 18 to 64), sex, BMI (15.96 to 53.13), number of children (0 to 5), smoking status, region, and medical charges (1122 to 63770), with mean age, BMI, and charges being 39.11, 30.70, and 13261, respectively.

```
# Generate pairwise plot matrix
ggpairs(Insurance)
```



Interpretation of the figure above:

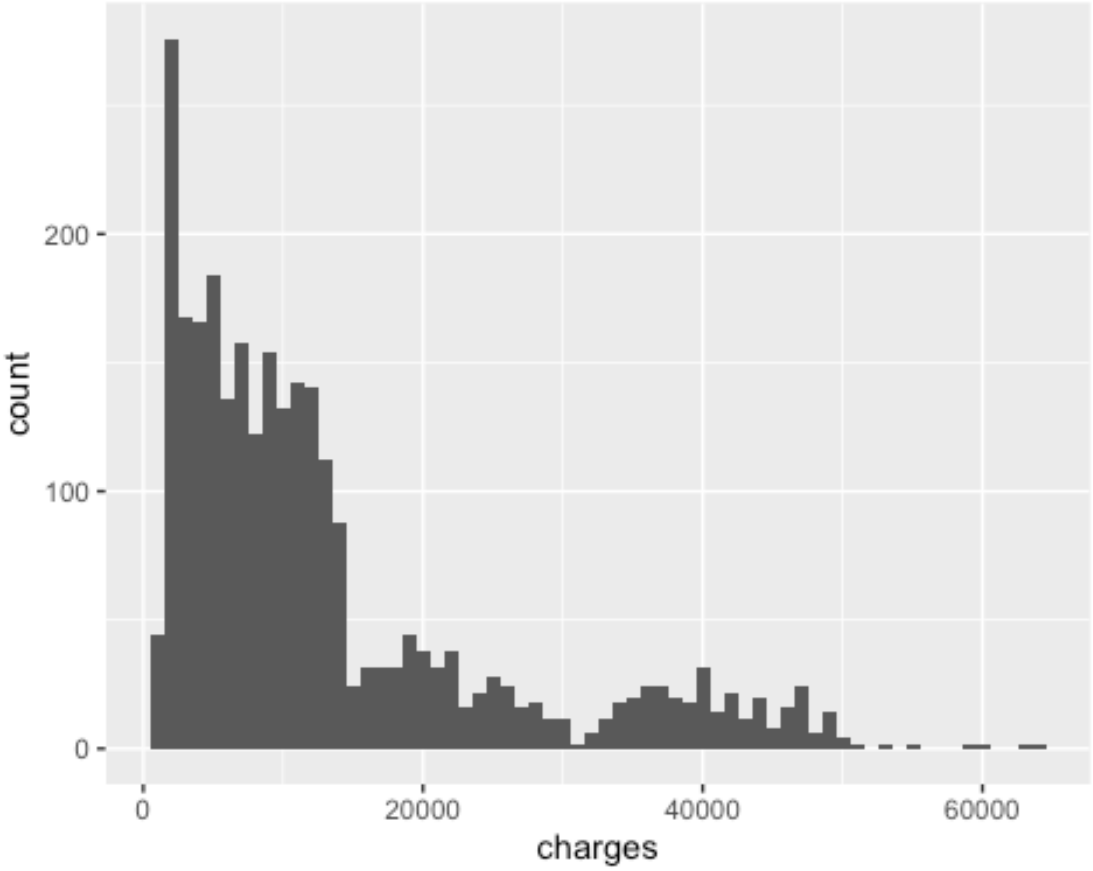
1. **Correlation Analysis**: Key correlations are highlighted, such as `age` (0.299), `bmi` (0.200), and `smoker` status with `charges`, indicating these variables have significant relationships with medical charges.

2. **Distribution of Variables**: The diagonal plots show the distribution of each variable, with `bmi` being approximately normally distributed and `charges` having a right-skewed distribution.

3. **Box Plots and Scatter Plots**: The plot includes box plots for categorical variables (e.g., `sex`, `smoker`, `region`) and scatter plots for continuous variables (e.g., `age`, `bmi`, `charges`), revealing data spread and potential outliers.

4. **Significance Levels**: Stars next to correlation coefficients indicate significance levels, with `age`, `bmi`, `children`, and `smoker` showing significant correlations with `charges`, guiding further analysis.

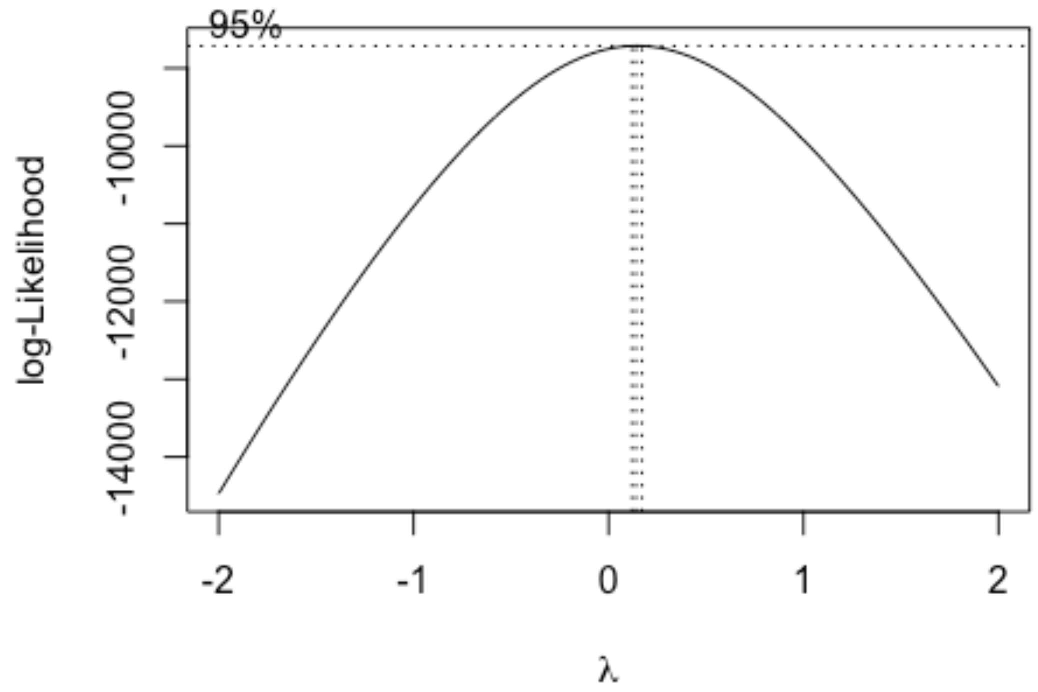
```
# Histograms for target variable
ggplot(Insurance, aes(x = charges)) + geom_histogram(binwidth = 1000)
```



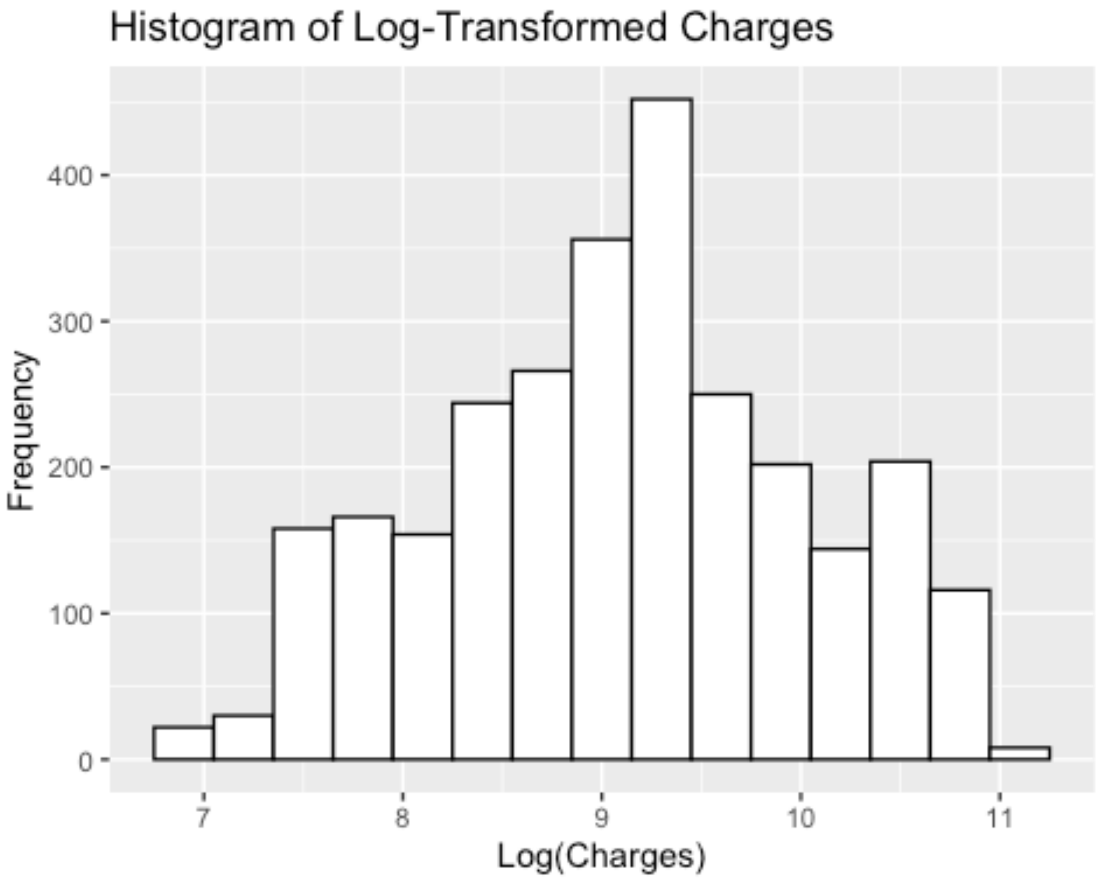
The histogram above shows that medical charges are right-skewed.

Multiple Regression Estimation

```
# Apply Box-Cox transformation
bc <- boxcox(Insurance$charges ~ ., data = Insurance)
```



```
lambda <- bc$x[which.max(bc$y)]
lambda

## [1] 0.1414141

# Apply the logarithmic transformation to the 'charges' variable
Insurance$log_charges <- log(Insurance$charges)

# Plot the histogram of the transformed 'charges' variable
ggplot(Insurance, aes(x = log_charges)) +
  geom_histogram(binwidth = 0.3, color = "black", fill = "white") +
  ggtitle("Histogram of Log-Transformed Charges") +
  xlab("Log(Charges)") +
  ylab("Frequency")
```



```r
# Convert categorical variables to factors
Insurance$sex <- as.factor(Insurance$sex)
Insurance$smoker <- as.factor(Insurance$smoker)
Insurance$region <- as.factor(Insurance$region)

# Split data into training and testing sets
set.seed(123)
train_index <- sample(seq_len(nrow(Insurance)), size = 0.7*nrow(Insurance))
Insurance <- Insurance[train_index, ]
Insurance_test <- Insurance[-train_index, ]
```

The Box-Cox transformation plot displayed suggests a lambda value close to zero, indicating that a logarithmic transformation is appropriate for stabilizing variance and addressing skewness in the charges variable. The histogram of the log-transformed charges displayed above confirms a more symmetric, bell-shaped distribution, which better meets linear regression assumptions. The regression model, fitted on the log-transformed charges, will now provide more accurate and reliable predictions.

Model Assessment

```
# Fit the multiple regression model
model <- lm(log_charges ~ age + sex + bmi + children + smoker + region, data
= Insurance)

# Summary of the model
summary(model)

##
## Call:
## lm(formula = log_charges ~ age + sex + bmi + children + smoker +
##     region, data = Insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94463 -0.20705 -0.05135  0.06592  2.16090
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.0418022  0.0604557 116.479  < 2e-16 ***
## age              0.0346885  0.0007265  47.748  < 2e-16 ***
## sexmale         -0.0754720  0.0203116  -3.716 0.000208 ***
## bmi              0.0126901  0.0017475   7.262 5.51e-13 ***
## children         0.1058834  0.0083928  12.616  < 2e-16 ***
## smokeryes        1.5374562  0.0253036  60.760  < 2e-16 ***
## regionnorthwest -0.0656032  0.0293941  -2.232 0.025739 *
## regionsoutheast -0.1449050  0.0293059  -4.945 8.29e-07 ***
## regionsouthwest -0.1348790  0.0293834  -4.590 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4456 on 1931 degrees of freedom
## Multiple R-squared:  0.7682, Adjusted R-squared:  0.7673
## F-statistic:   800 on 8 and 1931 DF,  p-value: < 2.2e-16
```

The regression model summary above shows that age, BMI, number of children, and being a smoker significantly increase log-transformed medical charges, with smoking having the largest effect. Males, and individuals in the Northwest, Southeast, and Southwest regions, tend to have lower log-transformed charges. The model explains about 76.8% of the variance in charges (R-squared = 0.7682), indicating a good fit, with all predictors being statistically significant ($p < 0.05$).

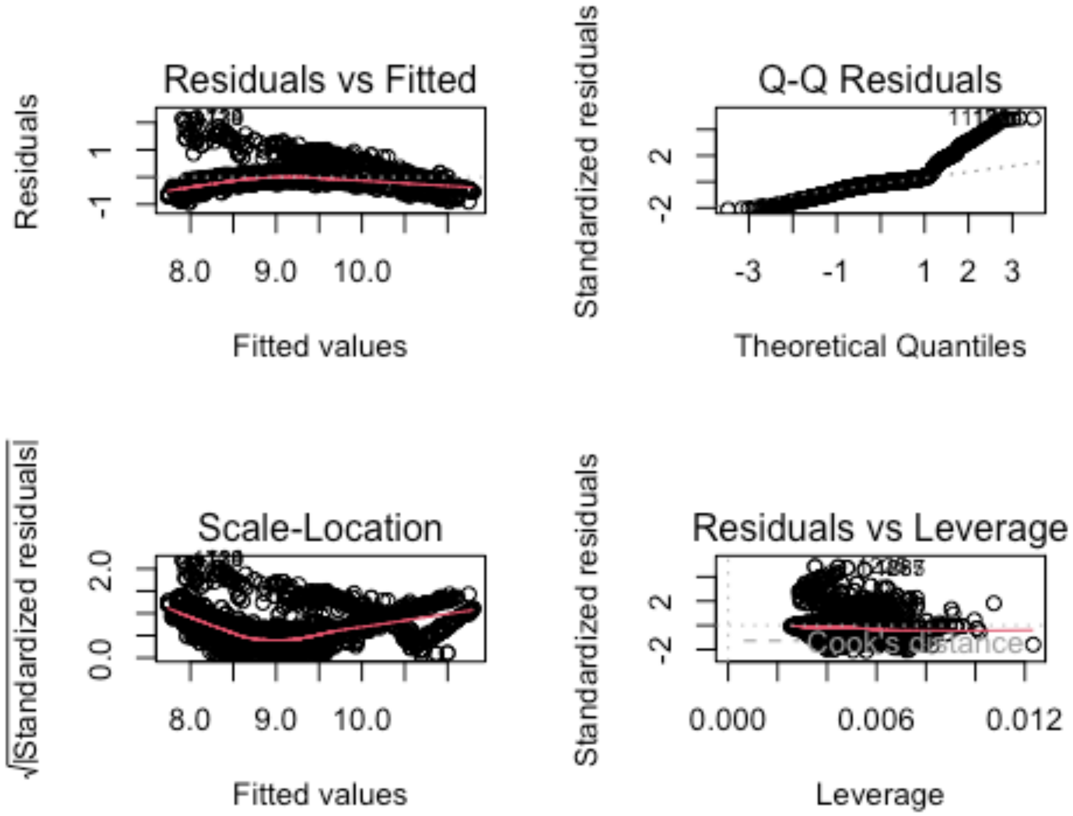
```
#Anova test
anova(model)
```

```
## Analysis of Variance Table
##
## Response: log_charges
##             Df Sum Sq Mean Sq   F value    Pr(>F)    
## age          1 488.82  488.82 2461.5626 < 2.2e-16 ***
## sex          1   0.02    0.02    0.0885    0.7661    
## bmi          1   9.12    9.12   45.9321 1.621e-11 ***
## children     1  32.04   32.04  161.3593 < 2.2e-16 ***
## smoker       1 734.72  734.72 3699.8204 < 2.2e-16 ***
## region       3   6.20    2.07   10.3993 8.681e-07 ***
## Residuals 1931 383.46    0.20                        
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table above for the log-transformed charges shows that age, BMI, number of children, smoking status, and region significantly contribute to the model, with p-values well below 0.05. Smoking status has the largest impact, explaining a significant portion of the variance in charges (F = 3699.8204). Age also has a substantial effect (F = 2461.5626), followed by the number of children (F = 161.3593) and BMI (F = 45.9321). Region also contributes significantly, though to a lesser extent (F = 10.3993).

Model Adequacy Check

```
par(mfrow = c(2, 2))
plot(model)
```



Interpretation pf the plots above:

1. **Residuals vs Fitted**: The plot shows a non-random pattern, indicating potential non-linearity or heteroscedasticity in the model.
2. **Q-Q Plot**: The standardized residuals deviate from the theoretical quantiles, suggesting that the residuals are not normally distributed.
3. **Scale-Location**: The plot shows a pattern indicating non-constant variance (heteroscedasticity) in the residuals.
4. **Residuals vs Leverage**: A few points with high leverage and standardized residuals suggest potential influential data points affecting the model's stability.

```
#Test: Independence
durbinWatsonTest(model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    -0.002612371      2.004517   0.912
##  Alternative hypothesis: rho != 0
```

Null Hypothesis (H0): Errors are uncorrelated. **Alternative Hypothesis (H1):** Errors are correlated.

The Durbin-Watson statistic of 2.004517 and p-value of 0.912 indicate no significant autocorrelation in the residuals of the regression model.

```
#Test: Normality
shapiro.test(residuals(model))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.82924, p-value < 2.2e-16
```

Null Hypothesis (H0): Errors are normally distributed. **Alternative Hypothesis (H1):** Errors are not normally distributed.

The Shapiro-Wilk test result (W = 0.82924, p-value < 2.2e-16) indicates that the residuals of the regression model significantly deviate from a normal distribution.

```
#Test: Homoscedasticity
bptest(model)

##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 118.35, df = 8, p-value < 2.2e-16
```

Null Hypothesis (H0): Errors have a constant variance (homoscedasticity). **Alternative Hypothesis (H1):** Errors have a non-constant variance (heteroscedasticity).

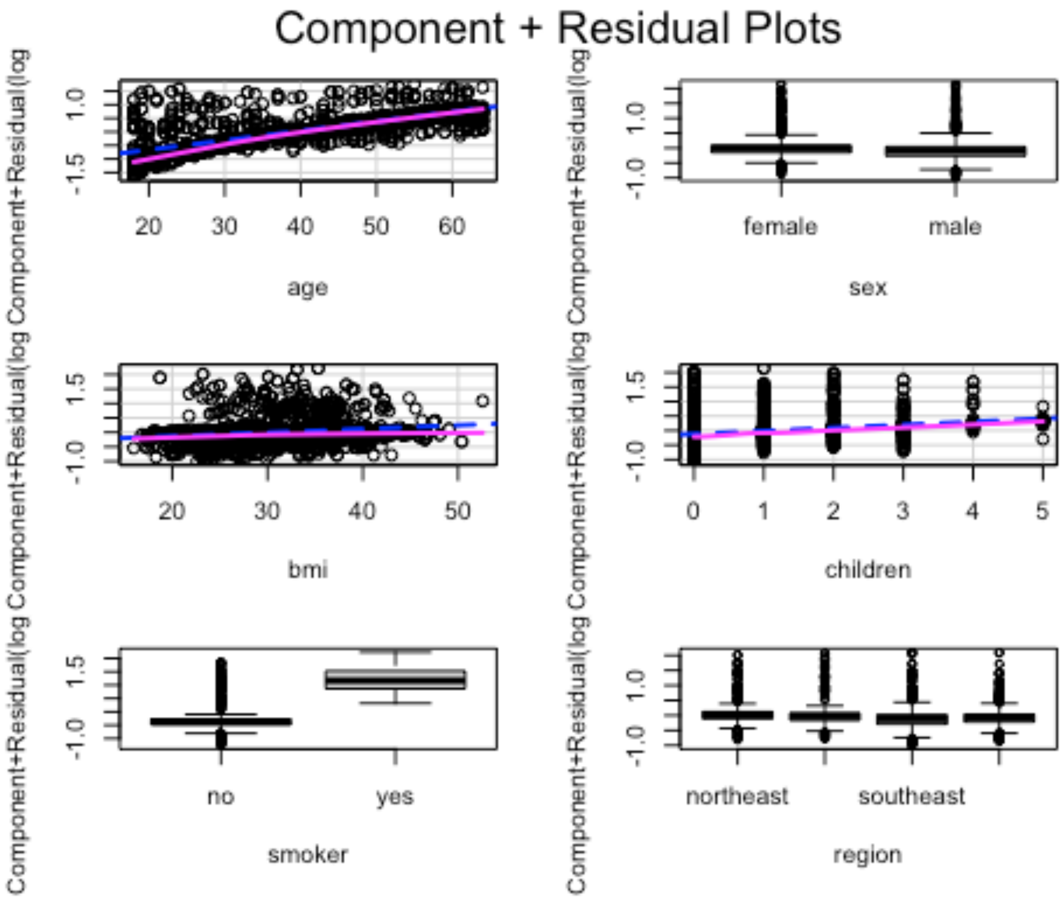
The studentized Breusch-Pagan test result (BP = 118.35, df = 8, p-value < 2.2e-16) indicates the presence of significant heteroscedasticity in the residuals of the regression model.

Model Diagnostic Check

```
#Test: Multicollinearity
vif <- vif(model)
print(vif)

##             age         sexmale             bmi        children
smokeryes
##          1.0143          1.0069          1.1098          1.0066
1.0104
## regionnorthwest regionsoutheast regionsouthwest
##          1.5291          1.7005          1.5529
```

The VIF values, all being below 2, indicate that there is no significant multicollinearity among the predictor variables in the regression model.

```
#Test: CR Plot
crPlots(model)
```



The component and residual plots above indicate mostly linear relationships between predictors and log-transformed charges, with significant effects for age, BMI, children, and smoker status, and some regional differences.

```
#Test: Outlier Test
outlierTest(model)

##      rstudent unadjusted p-value Bonferroni p
## 130  4.886374         1.1119e-06    0.0021570
## 724  4.886374         1.1119e-06    0.0021570
## 1119 4.832297         1.4562e-06    0.0028251
## 1792 4.832297         1.4562e-06    0.0028251
## 1225 4.763753         2.0420e-06    0.0039615
## 1485 4.763753         2.0420e-06    0.0039615
## 1287 4.624130         4.0112e-06    0.0077817
## 1865 4.624130         4.0112e-06    0.0077817
## 1272 4.338495         1.5087e-05    0.0292690
## 1493 4.338495         1.5087e-05    0.0292690
```

The rstudent and Bonferroni p-values indicated in the table above indicate that observations 130, 724, 1119, 1792, 1225, 1485, 1287, 1865, 1272, and 1493 are significant outliers in the regression model, even after adjusting for multiple comparisons.

```
influence_measures <- influence.measures(model)
# Extract Cook's distance from influence measures
cooks_distance <- influence_measures$infmat[, "cook.d"]
# Order the Cook's distances in decreasing order and get the top 10 indices
top_influential_indices <- order(cooks_distance, decreasing = TRUE)[1:10]
# Extract the influence measures for the top 10 influential points
top_influence_measures <- influence_measures$infmat[top_influential_indices,
]
# Print the top 10 influential points
print(top_influence_measures)
```

```
##          dfb.1_     dfb.age    dfb.sxml     dfb.bmi     dfb.chld
dfb.smkr
## 1287 0.19967180 -0.09262605  0.11520019 -0.18578299 -0.103505968 -
0.05441143
## 1865 0.19967180 -0.09262605  0.11520019 -0.18578299 -0.103505968 -
0.05441143
## 1225 0.22622972 -0.09443541 -0.10463548 -0.17472408 -0.084860035 -
0.06243449
## 1485 0.22622972 -0.09443541 -0.10463548 -0.17472408 -0.084860035 -
0.06243449
## 1119 0.02923075 -0.16099349  0.10863885  0.05804621 -0.107714771 -
0.04906967
## 1792 0.02923075 -0.16099349  0.10863885  0.05804621 -0.107714771 -
0.04906967
## 130  0.01781613 -0.15100037  0.10226430  0.04969942  0.002272592 -
0.07512977
## 724  0.01781613 -0.15100037  0.10226430  0.04969942  0.002272592 -
0.07512977
## 1272 0.14120391 -0.15193049 -0.09881775  0.03893844 -0.081993236 -
0.04026663
## 1493 0.14120391 -0.15193049 -0.09881775  0.03893844 -0.081993236 -
0.04026663
##          dfb.rgnn dfb.rgnsths  dfb.rgnsthw     dffit     cov.r      cook.d
## 1287  0.156986863  0.04570842  0.021011807 0.3444461 0.9148749 0.013044872
## 1865  0.156986863  0.04570842  0.021011807 0.3444461 0.9148749 0.013044872
## 1225  0.004364231  0.18083921  0.020016166 0.3313077 0.9087190 0.012060599
## 1485  0.004364231  0.18083921  0.020016166 0.3313077 0.9087190 0.012060599
## 1119 -0.001518746 -0.01959927  0.152665148 0.3229203 0.9056341 0.011453818
## 1792 -0.001518746 -0.01959927  0.152665148 0.3229203 0.9056341 0.011453818
## 130  -0.002335374  0.12371184 -0.005673464 0.2902883 0.9026019 0.009253408
## 724  -0.002335374  0.12371184 -0.005673464 0.2902883 0.9026019 0.009253408
## 1272 -0.144542378 -0.15390692 -0.144488764 0.2896615 0.9247438 0.009237382
## 1493 -0.144542378 -0.15390692 -0.144488764 0.2896615 0.9247438 0.009237382
##              hat
## 1287 0.005517970
## 1865 0.005517970
## 1225 0.004813590
## 1485 0.004813590
## 1119 0.004445786
## 1792 0.004445786
## 130  0.003516864
## 724  0.003516864
## 1272 0.004437840
## 1493 0.004437840
```

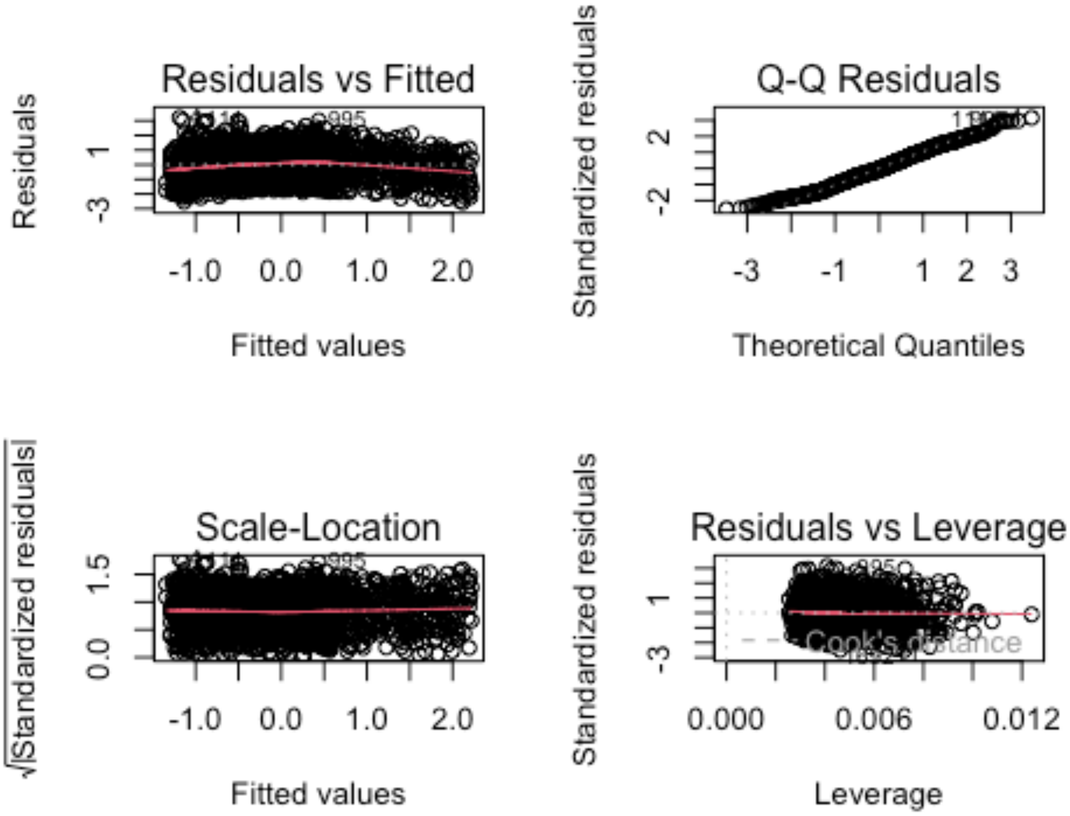
The dfbeta, dffit, cook's distance, and hat values displayed in the table above indicate that observations 1287, 1865, 1225, 1485, 1119, 1792, 130, 724, 1272, and 1493 have substantial influence on the regression model's coefficients, highlighting them as influential points.

Implementation of suitable corrective methods

```
# Differencing the log-transformed charges
Insurance$charges_diff <- c(NA, diff(Insurance$log_charges))
# Calculate studentized residuals
studentized_residuals <- rstudent(model)
# Identify influential points (these are indices of influential points
provided)
influential_points <- c(267, 997, 560, 17, 1181, 171, 1233, 960, 533, 1061)
# Remove influential points from the dataset
Insurance_clean <- Insurance[-influential_points, ]

# Refit the linear model without influential points
model_clean <- lm(charges_diff ~ age + sex + bmi + children + smoker +
region, data = Insurance_clean)
# Summary of the new model
summary(model_clean)

##
## Call:
## lm(formula = charges_diff ~ age + sex + bmi + children + smoker +
##     region, data = Insurance_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5692 -0.7158 -0.0672  0.7189  3.2047
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.102396   0.138697 -15.158  < 2e-16 ***
## age              0.037595   0.001665  22.584  < 2e-16 ***
## sexmale         -0.038847   0.046586  -0.834 0.404452
## bmi              0.011041   0.004009   2.754 0.005935 **
## children         0.072682   0.019250   3.776 0.000164 ***
## smokeryes        1.511952   0.057960  26.086  < 2e-16 ***
## regionnorthwest -0.055572   0.067441  -0.824 0.410038
## regionsoutheast -0.135066   0.067126  -2.012 0.044345 *
## regionsouthwest -0.086956   0.067382  -1.290 0.197033
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 1920 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3941, Adjusted R-squared:  0.3915
## F-statistic: 156.1 on 8 and 1920 DF,  p-value: < 2.2e-16
```

The regression model indicates that age, BMI, number of children, and smoker status significantly impact the difference in charges, with smoker status having the largest effect, while sex and region have lesser or no significant influence.

```
# Plot the new model diagnostics
par(mfrow = c(2, 2))
plot(model_clean)
```



Interpretation of the plots above:

1. **Residuals vs Fitted**: The residuals are randomly scattered around the horizontal axis, suggesting no major non-linearity but potential heteroscedasticity.
2. **Q-Q Plot**: The residuals follow the theoretical quantiles closely, indicating they are approximately normally distributed.

3. **Scale-Location**: The residuals display homoscedasticity with no clear pattern, suggesting constant variance across fitted values.
4. **Residuals vs Leverage**: A few points with higher leverage indicate potential influential observations, but most points have low leverage and standardized residuals.

```
outlierTest(model_clean)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##   rstudent unadjusted p-value Bonferroni p
## 4 3.158555          0.0016102           NA
```

The Bonferroni-adjusted p-value indicates that the observation with the rstudent value of 3.158555 is a significant outlier in the regression model, even after adjusting for multiple comparisons.

```
ncvTest(model_clean)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.09442879, Df = 1, p = 0.75862
```

Null Hypothesis (H0): Errors have a constant variance (homoscedasticity). **Alternative Hypothesis (H1):** Errors have a non-constant variance (heteroscedasticity).

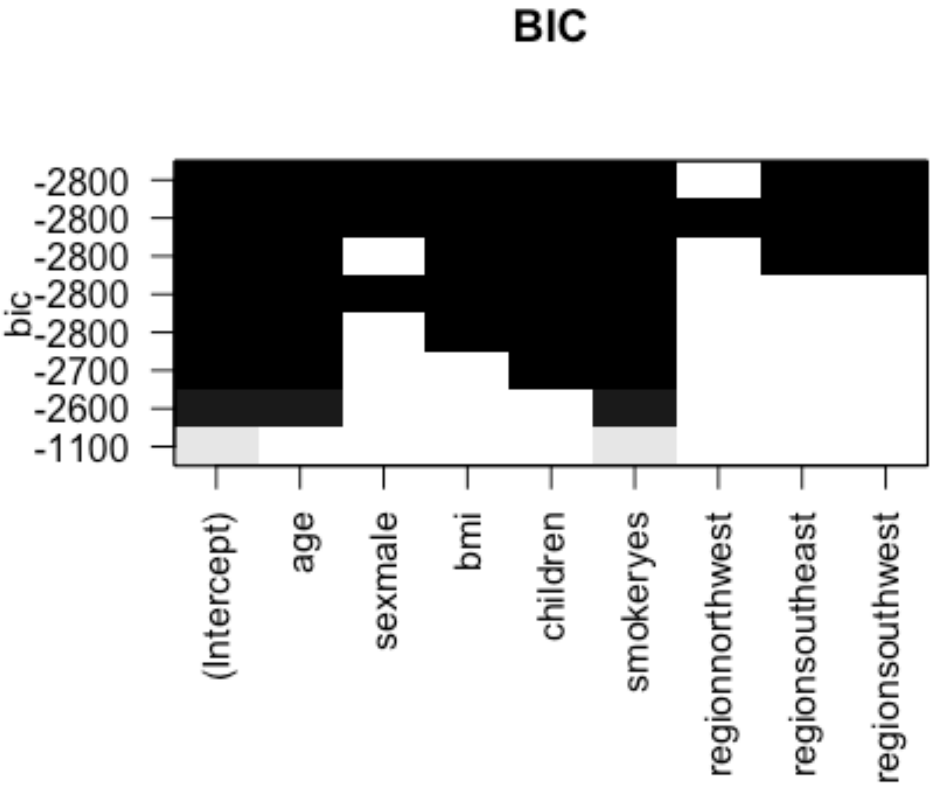
The Non-constant Variance Score Test indicates no significant heteroscedasticity in the residuals of the regression model (Chisquare = 0.09442879, p = 0.75862).

Variable Selection

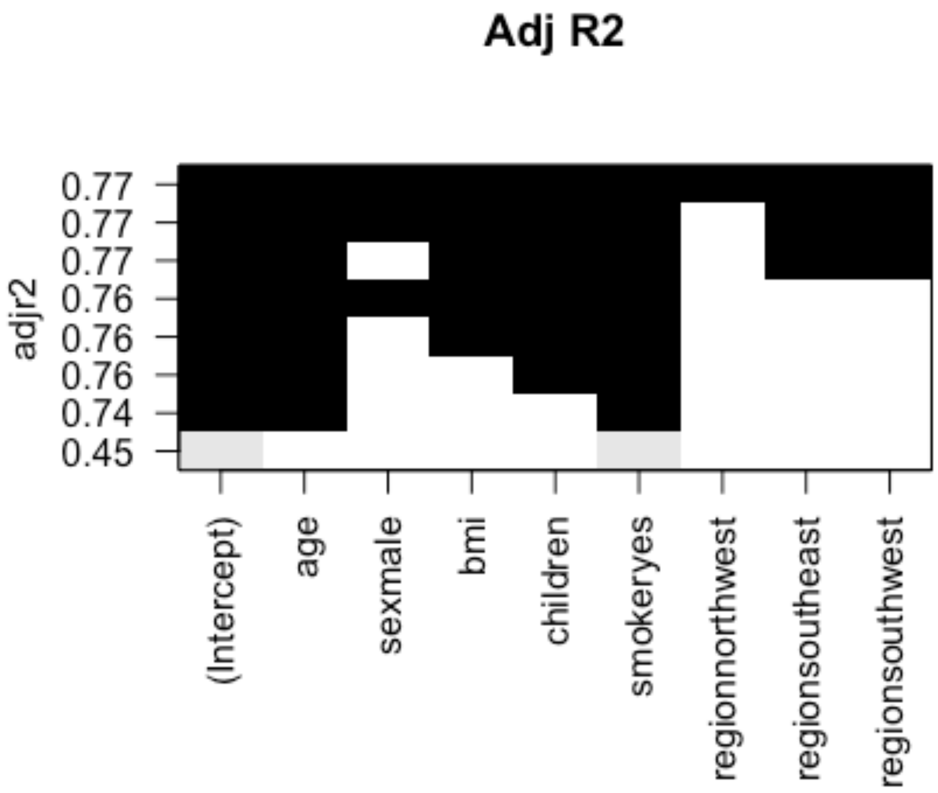
```
# All possible subsets
subsets <- leaps::regsubsets(log_charges ~ age + sex + bmi + children +
smoker + region, data = Insurance, nbest = 1)
# Summary
subsets_summary <- summary(subsets)
subsets_summary

## Subset selection object
## Call: regsubsets.formula(log_charges ~ age + sex + bmi + children +
##     smoker + region, data = Insurance, nbest = 1)
## 8 Variables  (and intercept)
##                 Forced in Forced out
## age                 FALSE      FALSE
## sexmale             FALSE      FALSE
## bmi                 FALSE      FALSE
## children            FALSE      FALSE
## smokeryes           FALSE      FALSE
## regionnorthwest     FALSE      FALSE
## regionsoutheast     FALSE      FALSE
## regionsouthwest     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          age sexmale bmi children smokeryes regionnorthwest
regionsoutheast
## 1  ( 1 ) " " " "     " " " "      "*"       " "             " "
## 2  ( 1 ) "*" " "     " " " "      "*"       " "             " "
## 3  ( 1 ) "*" " "     " " "*"      "*"       " "             " "
## 4  ( 1 ) "*" " "     "*" "*"      "*"       " "             " "
## 5  ( 1 ) "*" "*"     "*" "*"      "*"       " "             " "
## 6  ( 1 ) "*" " "     "*" "*"      "*"       " "             "*"
## 7  ( 1 ) "*" "*"     "*" "*"      "*"       " "             "*"
## 8  ( 1 ) "*" "*"     "*" "*"      "*"       "*"             "*"
##          regionsouthwest
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
```

The subset selection results displayed above from the exhaustive algorithm indicate the best models with up to 8 predictors for predicting log-transformed charges. Smoking status (smokeryes) is included in all models, highlighting its strong predictive power. Age and BMI are consistently included from the 2-variable model onwards, indicating their importance. The full model with all predictors is also considered, showing that including all variables provides the most comprehensive fit.

```
# BIC plot
plot(subsets, scale = "bic", main = "BIC")
```



```r
# Adj_r2 plot
plot(subsets, scale = "adjr2", main = "Adj R2")
```



Interpretation of the plots above:

1. **BIC Plot**: The model including age, BMI, number of children, and smoker status minimizes the Bayesian Information Criterion (BIC), suggesting it is the best subset of predictors.
2. **Adjusted R^2 Plot**: The model with age, BMI, number of children, and smoker status maximizes the adjusted R-squared, indicating it explains the most variance in the log-transformed charges while accounting for the number of predictors.

```
model_new <- lm(log_charges ~ age + sex + bmi + children + smoker, data =
Insurance)
summary(model_new)

##
## Call:
## lm(formula = log_charges ~ age + sex + bmi + children + smoker,
##     data = Insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96942 -0.20846 -0.05255  0.06894  2.11759
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.0211718  0.0588322 119.342  < 2e-16 ***
## age          0.0347870  0.0007311  47.579  < 2e-16 ***
## sexmale     -0.0761232  0.0204581  -3.721 0.000204 ***
## bmi          0.0103509  0.0016827   6.151 9.31e-10 ***
## children     0.1061965  0.0084386  12.585  < 2e-16 ***
## smokeryes    1.5336750  0.0253972  60.388  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4489 on 1934 degrees of freedom
## Multiple R-squared:  0.7645, Adjusted R-squared:  0.7639
## F-statistic:  1255 on 5 and 1934 DF,  p-value: < 2.2e-16
```

Interpretation of the table above:

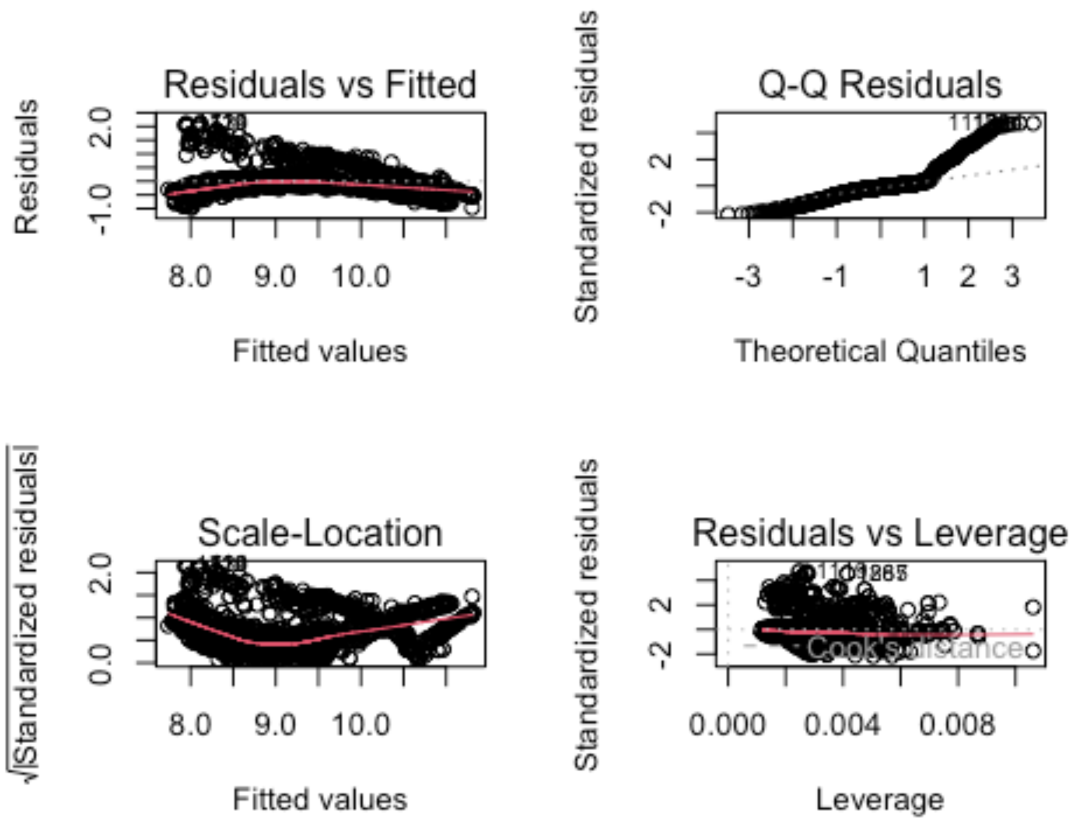
The regression model shows that age, sex, BMI, number of children, and smoker status are significant predictors of log-transformed charges, explaining 76.45% of the variance with all coefficients being highly significant (p < 0.001).

```
anova(model_new)
```

```
## Analysis of Variance Table
##
## Response: log_charges
##             Df Sum Sq Mean Sq   F value    Pr(>F)
## age          1 488.82  488.82 2426.1884 < 2.2e-16 ***
## sex          1   0.02    0.02    0.0872    0.7678
## bmi          1   9.12    9.12   45.2720 2.252e-11 ***
## children     1  32.04   32.04  159.0405 < 2.2e-16 ***
## smoker       1 734.72  734.72 3646.6517 < 2.2e-16 ***
## Residuals 1934 389.66    0.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table above indicates that age, BMI, number of children, and smoking status are significant predictors of log-transformed charges (p < 0.001), with smoking status having the largest effect, while sex is not a significant predictor (p = 0.7678).

```
par(mfrow = c(2, 2))
plot(model_new)
```



Interpretation of the plots above:

1. **Residuals vs Fitted**: The residuals show a slight curve, indicating potential non-linearity and heteroscedasticity in the model.
2. **Q-Q Plot**: The residuals deviate from the diagonal line, especially at the tails, suggesting that they are not normally distributed.
3. **Scale-Location**: The residuals show a funnel shape, indicating heteroscedasticity, as the spread increases with fitted values.
4. **Residuals vs Leverage**: A few points with high leverage and standardized residuals indicate potential influential observations affecting the model's stability.

```
#Test: Independence
durbinWatsonTest(model_new)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    0.0006154744      1.997864   0.972
##  Alternative hypothesis: rho != 0
```

Null Hypothesis (H0): Errors are uncorrelated. **Alternative Hypothesis (H1):** Errors are correlated.

The Durbin-Watson statistic of 1.997864 and p-value of 0.972 indicate no significant autocorrelation in the residuals of the regression model, as the test fails to reject the null hypothesis of no autocorrelation.

```
#Test: Normality
shapiro.test(residuals(model_new))

## 
##  Shapiro-Wilk normality test
## 
## data:  residuals(model_new)
## W = 0.83203, p-value < 2.2e-16
```

Null Hypothesis (H0): Errors are normally distributed. **Alternative Hypothesis (H1):** Errors are not normally distributed.

The Shapiro-Wilk test result (W = 0.83203, p-value < 2.2e-16) indicates that the residuals of the new regression model significantly deviate from a normal distribution.

```
#Test: Homoscedasticity
bptest(model_new)

## 
##  studentized Breusch-Pagan test
## 
## data:  model_new
## BP = 118.74, df = 5, p-value < 2.2e-16
```

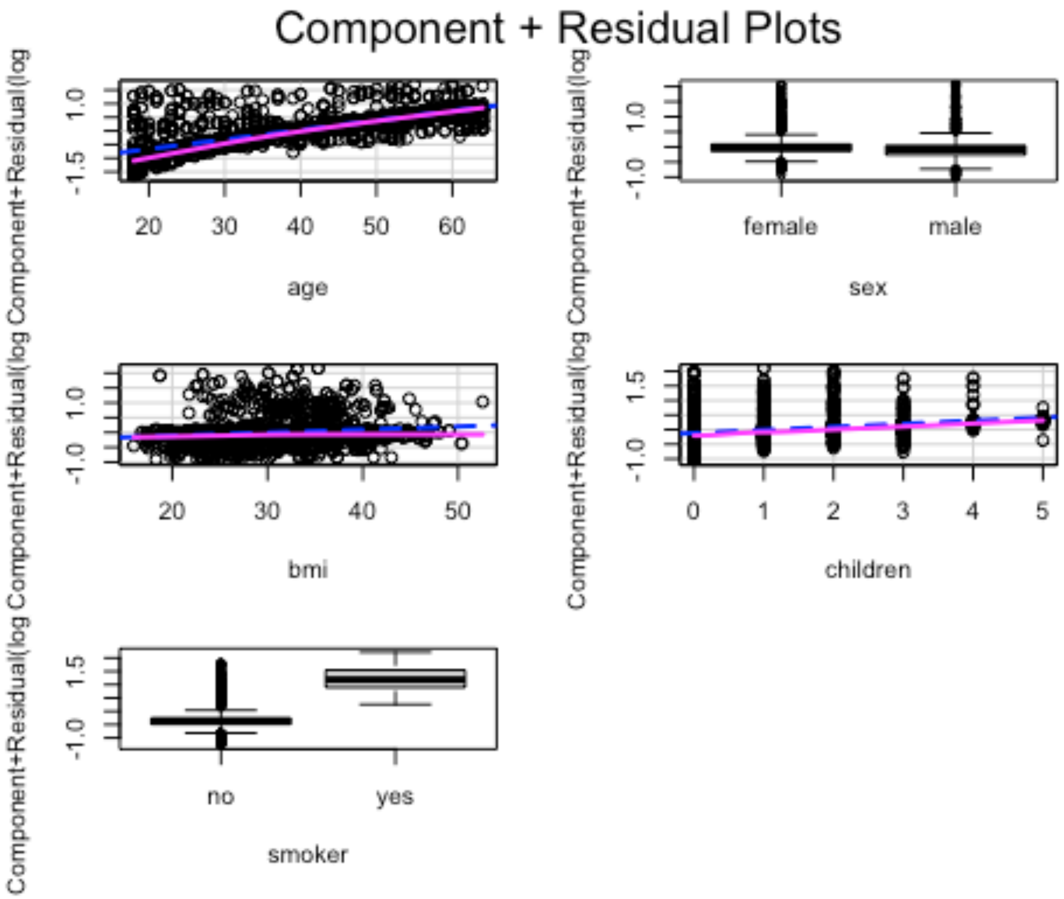
Null Hypothesis (H0): Errors have a constant variance (homoscedasticity). **Alternative Hypothesis (H1):** Errors have a non-constant variance (heteroscedasticity).

The studentized Breusch-Pagan test result (BP = 118.74, df = 5, p-value < 2.2e-16) indicates significant heteroscedasticity in the residuals of the new regression model.

```
#Test: Multicollinearity
vif1 <- vif(model_new)
print(vif1)

##       age   sexmale       bmi  children smokeryes 
##    1.0125    1.0068    1.0142    1.0030    1.0033
```

The VIF values, all being close to 1, indicate that there is no significant multicollinearity among the predictor variables in the regression model.

```
#Test: CR Plot
crPlots(model_new)
```



Interpretation of the plots above:

The component and residual plots show mostly linear relationships between predictors and log-transformed charges, with significant positive effects for age, BMI, number of children, and smoker status.

```
#Test: Outlier Test
outlierTest(model_new)

##       rstudent unadjusted p-value Bonferroni p
## 130   4.749784         2.1862e-06    0.0042413
## 724   4.749784         2.1862e-06    0.0042413
## 1119 4.706480         2.6992e-06    0.0052365
## 1792 4.706480         2.6992e-06    0.0052365
## 1287 4.580816         4.9271e-06    0.0095587
## 1865 4.580816         4.9271e-06    0.0095587
## 1225 4.559996         5.4360e-06    0.0105460
## 1485 4.559996         5.4360e-06    0.0105460
## 1272 4.505315         7.0235e-06    0.0136260
## 1493 4.505315         7.0235e-06    0.0136260
```

The Bonferroni-adjusted p-values indicate that observations 130, 724, 1119, 1792, 1287, 1865, 1225, 1485, 1272, and 1493 are significant outliers in the regression model.

```
influence_measures_new <- influence.measures(model_new)
# Extract Cook's distance from influence measures
cooks_distance_new <- influence_measures_new$infmat[, "cook.d"]
# Order the Cook's distances in decreasing order and get the top 10 indices
top_influential_indices_new <- order(cooks_distance_new, decreasing =
TRUE)[1:10]
# Extract the influence measures for the top 10 influential points
top_influence_measures_new <-
influence_measures_new$infmat[top_influential_indices_new, ]
# Print the top 10 influential points
print(top_influence_measures_new)
```

```
##             dfb.1_     dfb.age    dfb.sxml      dfb.bmi     dfb.chld
dfb.smkr
## 1287  0.239355843 -0.09257781  0.11392389 -0.200236210 -0.102084906 -
0.05408569
## 1865  0.239355843 -0.09257781  0.11392389 -0.200236210 -0.102084906 -
0.05408569
## 1119  0.048796998 -0.15262501  0.10480291  0.051813466 -0.095606730 -
0.05951962
## 1792  0.048796998 -0.15262501  0.10480291  0.051813466 -0.095606730 -
0.05951962
## 1225  0.201598674 -0.09868950 -0.09800308 -0.113781166 -0.091513352 -
0.04389657
## 1485  0.201598674 -0.09868950 -0.09800308 -0.113781166 -0.091513352 -
0.04389657
## 130  -0.001036918 -0.15304600  0.10089723  0.093786046 -0.006006477 -
0.06078012
## 724  -0.001036918 -0.15304600  0.10089723  0.093786046 -0.006006477 -
0.06078012
## 1272  0.120460956 -0.15394135 -0.10326430  0.009028062 -0.085306723 -
0.04505202
## 1493  0.120460956 -0.15394135 -0.10326430  0.009028062 -0.085306723 -
0.04505202
##           dffit     cov.r      cook.d         hat
## 1287 0.2956976 0.9441019 0.014423804 0.004149579
## 1865 0.2956976 0.9441019 0.014423804 0.004149579
## 1119 0.2463612 0.9393888 0.010006207 0.002732523
## 1792 0.2463612 0.9393888 0.010006207 0.002732523
## 1225 0.2404874 0.9433504 0.009541378 0.002773636
## 1485 0.2404874 0.9433504 0.009541378 0.002773636
## 130  0.2366729 0.9379686 0.009232749 0.002476694
## 724  0.2366729 0.9379686 0.009232749 0.002476694
## 1272 0.2355121 0.9447417 0.009152992 0.002725152
## 1493 0.2355121 0.9447417 0.009152992 0.002725152
```

Interpretation of the table above:

The dfbeta, dffit, Cook's distance, and hat values indicate that observations 1287, 1865, 1119, 1792, 1225, 1485, 130, 724, 1272, and 1493 have substantial influence on the regression model's coefficients, highlighting them as influential points.

Model Validation

```
compare <-anova(model,model_new)
compare

## Analysis of Variance Table
##
## Model 1: log_charges ~ age + sex + bmi + children + smoker + region
## Model 2: log_charges ~ age + sex + bmi + children + smoker
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1931 383.46
## 2   1934 389.66 -3   -6.1954 10.399 8.681e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows that including the region variable in the model significantly improves the fit (F = 10.399, p < 0.001). This indicates that region has a significant impact on log-transformed charges beyond the effects of age, sex, BMI, number of children, and smoker status.

```
# PRESS
DAAG::press(model)

## [1] 387.1274

DAAG::press(model_new)

## [1] 392.1664
```

The predicted residual sum of squares (PRESS) values indicate that the original model (PRESS = 387.1274) has a better predictive performance compared to the new model (PRESS = 392.1664), suggesting that the original model provides more accurate predictions for log-transformed charges.

```
# Predict on the testing set
predictions <- predict(model, Insurance)
# Calculate performance metrics
mse <- mean((predictions - Insurance_test$log_charges)^2)
rmse <- sqrt(mse)
mae <- mean(abs(predictions - Insurance_test$log_charges))
# Print performance metrics
cat("MSE:", mse, "\n")

## MSE: 1.551862
```

```r
cat("RMSE:", rmse, "\n")

## RMSE: 1.245737

cat("MAE:", mae, "\n")

## MAE: 1.008911

# Predict on the testing set
predictions2 <- predict(model_new, Insurance)
# Calculate performance metrics
mse2 <- mean((predictions2 - Insurance_test$log_charges)^2)
rmse2 <- sqrt(mse2)
mae2 <- mean(abs(predictions2 - Insurance_test$log_charges))
# Print performance metrics
cat("MSE:", mse2, "\n")

## MSE: 1.549877

cat("RMSE:", rmse2, "\n")

## RMSE: 1.24494

cat("MAE:", mae2, "\n")

## MAE: 1.008502
```

Considering all the evaluation metrics—ANOVA, PRESS, MSE, RMSE, and MAE—let's summarize:

1. **ANOVA**: The original model (with region) significantly improves the fit compared to the new model (without region) (F = 10.399, $p < 0.001$).
2. **PRESS**: The original model has a lower PRESS value (387.1274) compared to the new model (392.1664), indicating better predictive performance.
3. **MSE, RMSE, and MAE**: The new model has slightly better values, but the differences are very marginal (MSE: 1.549877 vs. 1.551862, RMSE: 1.24494 vs. 1.245737, MAE: 1.008502 vs. 1.008911).

Given these points: - The ANOVA and PRESS values strongly favor the original model. - The differences in MSE, RMSE, and MAE are minimal and do not outweigh the significant improvement seen in the ANOVA and PRESS. The original model, which includes the region variable, is considered the best overall model due to its significantly better fit and predictive performance as indicated by the ANOVA and PRESS values.

References

Wang, Y. (2024). *Regression analysis* [Lecture and lab notes]. RMIT University.

Yasm, R. (2024). *Medical insurance cost prediction* [Data set]. Kaggle. https://www.kaggle.com/datasets/rahulvyasm/medical-insurance-cost-prediction/data

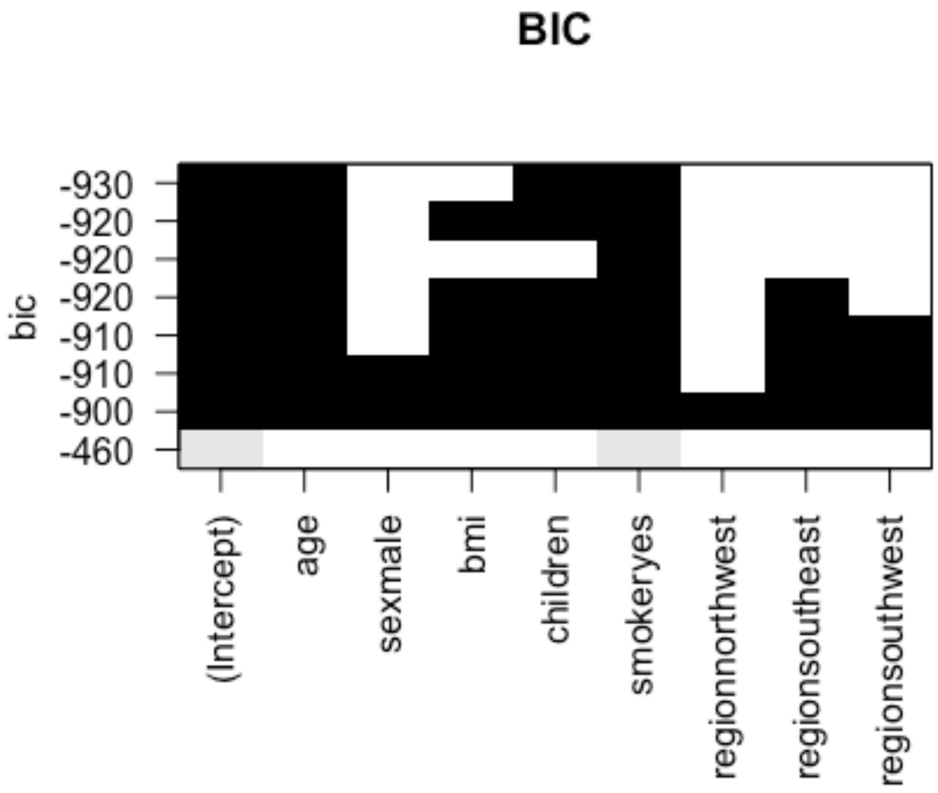
Appendix

We have fitted a model using the corrections made but it gave us new models but their Press values were really high. hence we did not move forward with them. However we did perform necessary analysis.

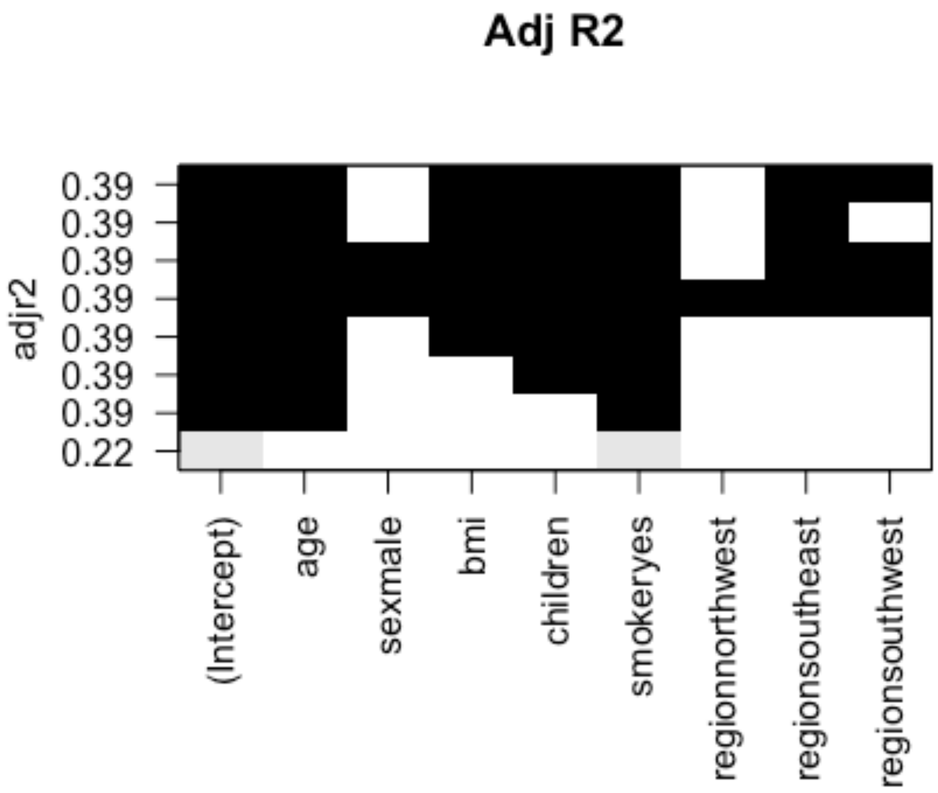
```
# All possible subsets
subsets_clean <- leaps::regsubsets(charges_diff ~ age + sex + bmi + children
+ smoker + region, data = Insurance_clean, nbest = 1)
# Summary
subsets_summary_clean <- summary(subsets_clean)
subsets_summary_clean

## Subset selection object
## Call: regsubsets.formula(charges_diff ~ age + sex + bmi + children +
##     smoker + region, data = Insurance_clean, nbest = 1)
## 8 Variables  (and intercept)
##                 Forced in Forced out
## age                 FALSE      FALSE
## sexmale             FALSE      FALSE
## bmi                 FALSE      FALSE
## children            FALSE      FALSE
## smokeryes           FALSE      FALSE
## regionnorthwest     FALSE      FALSE
## regionsoutheast     FALSE      FALSE
## regionsouthwest     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          age sexmale bmi children smokeryes regionnorthwest
regionsoutheast
## 1  ( 1 ) " " " "     " " " "      "*"       " "             " "
## 2  ( 1 ) "*" " "     " " " "      "*"       " "             " "
## 3  ( 1 ) "*" " "     " " "*"      "*"       " "             " "
## 4  ( 1 ) "*" " "     "*" "*"      "*"       " "             " "
## 5  ( 1 ) "*" " "     "*" "*"      "*"       " "             "*"
## 6  ( 1 ) "*" " "     "*" "*"      "*"       " "             "*"
## 7  ( 1 ) "*" "*"     "*" "*"      "*"       " "             "*"
## 8  ( 1 ) "*" "*"     "*" "*"      "*"       "*"             "*"
##          regionsouthwest
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
```

The exhaustive subset selection algorithm indicates that all variables, starting from the most significant ones (smoker, age, bmi, etc.), should be included sequentially, with the best model including all variables.

```
# BIC plot
plot(subsets_clean, scale = "bic", main = "BIC")
```



```
# Adj_r2 plot
plot(subsets_clean, scale = "adjr2", main = "Adj R2")
```



Interpretation of the plots above:

Plot 1 (BIC): The model including `smokeryes`, `age`, `bmi`, `children`, `regionsoutheast`, and `regionsouthwest` has the lowest Bayesian Information Criterion (BIC), indicating the best model among the subsets evaluated. **Plot 2 (Adjusted R^2):** The model including `smokeryes`, `age`, `bmi`, `children`, and `regionsoutheast` has the highest adjusted R^2 value, indicating the best fit among the subsets evaluated.

```
model2 <- lm(charges_diff ~ age + children + smoker, data = Insurance_clean)
summary(model2)

## 
## Call:
## lm(formula = charges_diff ~ age + children + smoker, data = 
Insurance_clean)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.6249 -0.7350 -0.0551  0.7077  3.2059 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.874316   0.072317 -25.918   <2e-16 ***
## age          0.038131   0.001656  23.025   <2e-16 ***
## children     0.071597   0.019217   3.726    2e-04 ***
## smokeryes    1.505382   0.057737  26.073   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.02 on 1925 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3909, Adjusted R-squared:  0.3899
## F-statistic: 411.8 on 3 and 1925 DF,  p-value: < 2.2e-16
```

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: charges_diff
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## age           1  563.87  563.87  541.56 < 2.2e-16 ***
## children      1   14.50   14.50   13.93 0.0001953 ***
## smoker        1  707.80  707.80  679.80 < 2.2e-16 ***
## Residuals 1925 2004.29    1.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation of the table above:

The regression model shows that age, `children`, and `smokeryes` are significant predictors of `charges_diff`, with all coefficients highly significant ($p < 0.001$), and the model explains approximately 39% of the variability in `charges_diff` (Adjusted R-squared = 0.3899).

```
model3 <- lm(charges_diff ~ age + bmi + children + smoker, data =
Insurance_clean)
summary(model3)
```

```
##
## Call:
## lm(formula = charges_diff ~ age + bmi + children + smoker, data =
Insurance_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6104 -0.7293 -0.0610  0.7240  3.1714
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.125637   0.132711 -16.017  < 2e-16 ***
## age          0.037747   0.001663  22.698  < 2e-16 ***
## bmi          0.008642   0.003828   2.258 0.024088 *
## children     0.072897   0.019206   3.796 0.000152 ***
## smokeryes    1.503187   0.057684  26.059  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.019 on 1924 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3925, Adjusted R-squared:  0.3912
## F-statistic: 310.8 on 4 and 1924 DF,  p-value: < 2.2e-16

anova(model3)

## Analysis of Variance Table
##
## Response: charges_diff
##             Df  Sum Sq Mean Sq  F value    Pr(>F)
## age          1  563.87  563.87 542.7101 < 2.2e-16 ***
## bmi          1    6.94    6.94   6.6765 0.0098423 **
## children     1   15.12   15.12  14.5574 0.0001402 ***
## smoker       1  705.53  705.53 679.0637 < 2.2e-16 ***
## Residuals 1924 1999.00    1.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation of the table above:

The regression model indicates that `age`, `bmi`, `children`, and `smokeryes` are significant predictors of `charges_diff`, with all coefficients highly significant (p < 0.05), and the model explains approximately 39% of the variability in `charges_diff` (Adjusted R-squared = 0.3912).

```
# PRESS
DAAG::press(model2)

## [1] 2013.026

DAAG::press(model3)

## [1] 2009.608
```

Previous models had lower PRESS values, it indicates those models were better at predicting the target variable.

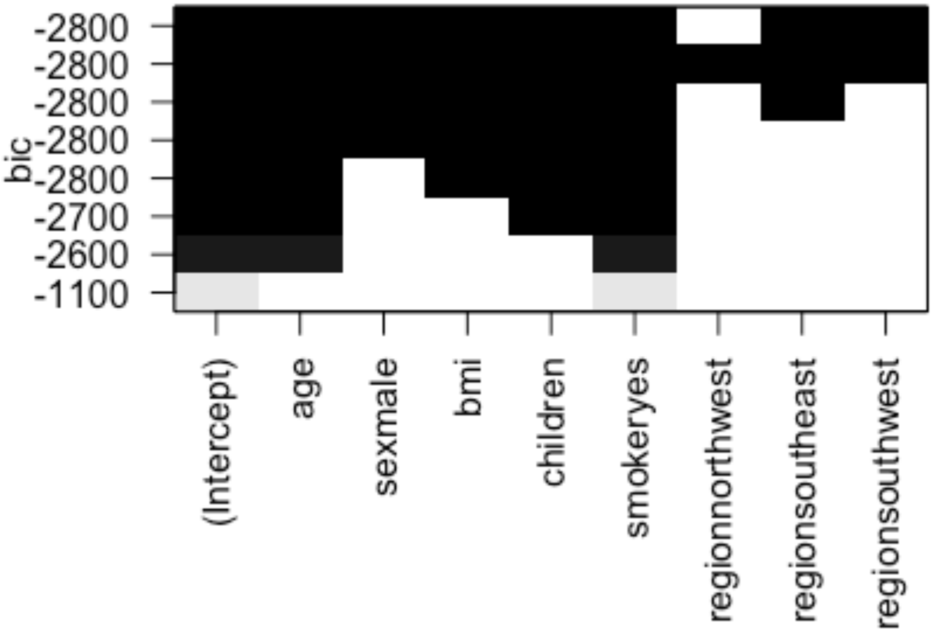
We did stepwise regression for model and model new but we did not find any new models. However we did the necessary analysis.

```
## Forward stepwise regression
model_forward_og <- regsubsets(log_charges ~ age + sex + bmi + children +
smoker + region, method = "forward", data = Insurance)
summary(model_forward_og)

## Subset selection object
## Call: regsubsets.formula(log_charges ~ age + sex + bmi + children +
##     smoker + region, method = "forward", data = Insurance)
## 8 Variables  (and intercept)
##                 Forced in Forced out
## age                   FALSE      FALSE
## sexmale               FALSE      FALSE
## bmi                   FALSE      FALSE
## children              FALSE      FALSE
## smokeryes             FALSE      FALSE
## regionnorthwest       FALSE      FALSE
## regionsoutheast       FALSE      FALSE
## regionsouthwest       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##          age sexmale bmi children smokeryes regionnorthwest
regionsoutheast
## 1  ( 1 ) " " " "     " " " "      "*"       " "             " "
## 2  ( 1 ) "*" " "     " " " "      "*"       " "             " "
## 3  ( 1 ) "*" " "     " " "*"      "*"       " "             " "
## 4  ( 1 ) "*" " "     "*" "*"      "*"       " "             " "
## 5  ( 1 ) "*" "*"     "*" "*"      "*"       " "             " "
## 6  ( 1 ) "*" "*"     "*" "*"      "*"       " "             "*"
## 7  ( 1 ) "*" "*"     "*" "*"      "*"       " "             "*"
## 8  ( 1 ) "*" "*"     "*" "*"      "*"       "*"             "*"
##          regionsouthwest
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
```

Interpretation of the table above:

The forward subset selection algorithm indicates that all variables, starting from the most significant ones (smokeryes, age, bmi, children, etc.), should be included sequentially, with the best model including all variables.

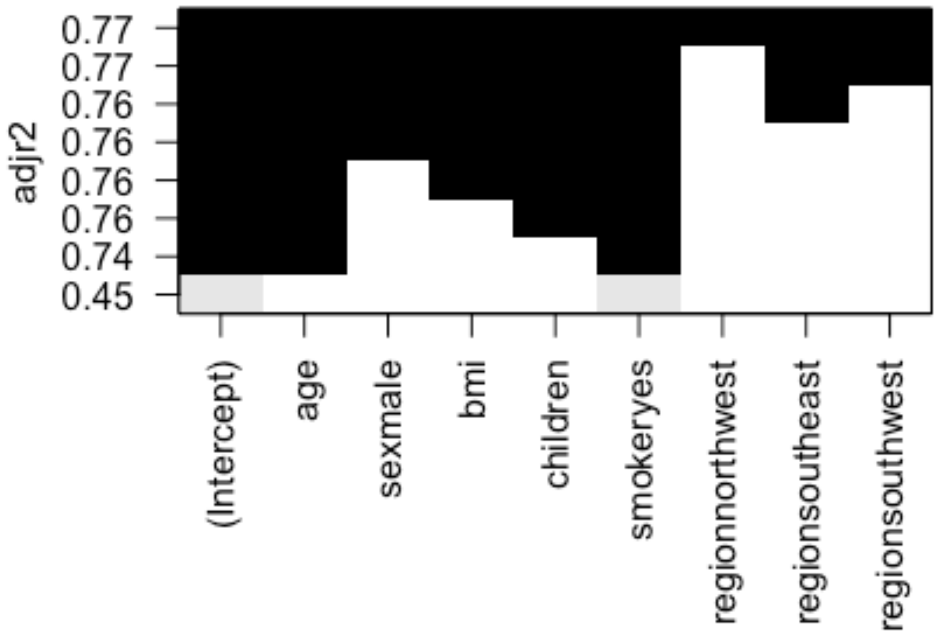
```
plot(model_forward_og,scale="bic")
```



Interpretation of the plot above:

The Bayesian Information Criterion (BIC) plot suggests that the model including `smokeryes`, `age`, `bmi`, `children`, and `regionsoutheast` has the lowest BIC value, indicating it is the best-fitting model among the evaluated subsets.

```
plot(model_forward_og,scale="adjr2")
```



Interpretation of the plot above:

The adjusted R^2 plot indicates that the model including `smokeryes`, `age`, `bmi`, `children`, and `regionsoutheast` achieves the highest adjusted R^2 value, signifying it provides the best fit among the evaluated subsets.

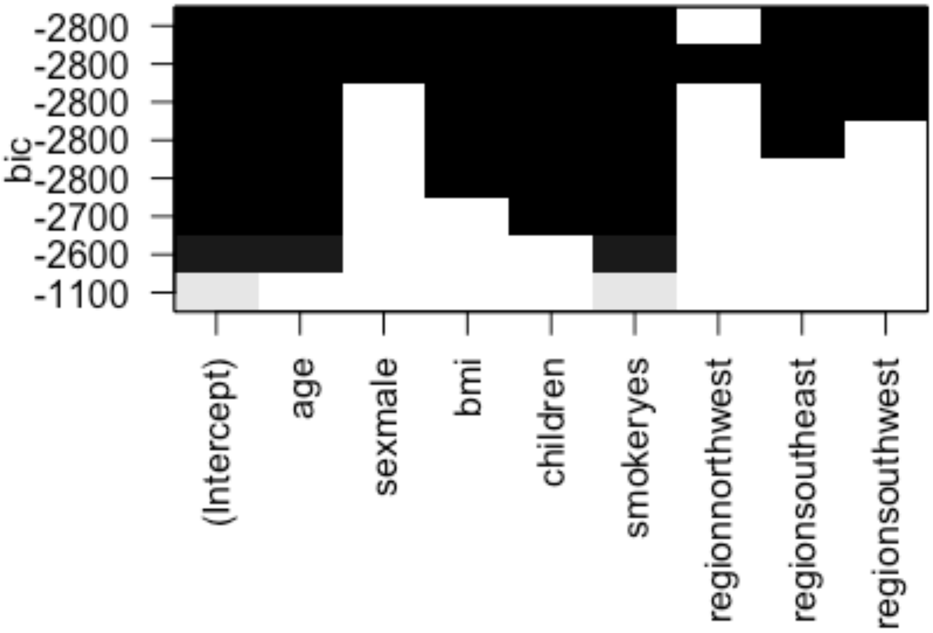
```
## Backward stepwise regression
model_backward_og <- regsubsets(log_charges ~ age + sex + bmi + children +
smoker + region, method = "backward", data = Insurance)
summary(model_backward_og)
```

```
## Subset selection object
## Call: regsubsets.formula(log_charges ~ age + sex + bmi + children +
##     smoker + region, method = "backward", data = Insurance)
## 8 Variables  (and intercept)
##                  Forced in Forced out
## age                  FALSE      FALSE
## sexmale              FALSE      FALSE
## bmi                  FALSE      FALSE
## children             FALSE      FALSE
## smokeryes            FALSE      FALSE
## regionnorthwest      FALSE      FALSE
## regionsoutheast      FALSE      FALSE
## regionsouthwest      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: backward
##          age sexmale bmi children smokeryes regionnorthwest
regionsoutheast
## 1  ( 1 ) " " " "     " " " "      "*"       " "             " "
## 2  ( 1 ) "*" " "     " " " "      "*"       " "             " "
## 3  ( 1 ) "*" " "     " " "*"      "*"       " "             " "
## 4  ( 1 ) "*" " "     "*" "*"      "*"       " "             " "
## 5  ( 1 ) "*" " "     "*" "*"      "*"       " "             "*"
## 6  ( 1 ) "*" " "     "*" "*"      "*"       " "             "*"
## 7  ( 1 ) "*" "*"     "*" "*"      "*"       " "             "*"
## 8  ( 1 ) "*" "*"     "*" "*"      "*"       "*"             "*"
##          regionsouthwest
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
```

Interpretation of the table above:

The backward subset selection algorithm indicates that all variables, starting from the least significant ones, should be excluded sequentially, with the best model including `smokeryes`, `age`, `bmi`, `children`, and `regionsoutheast`.

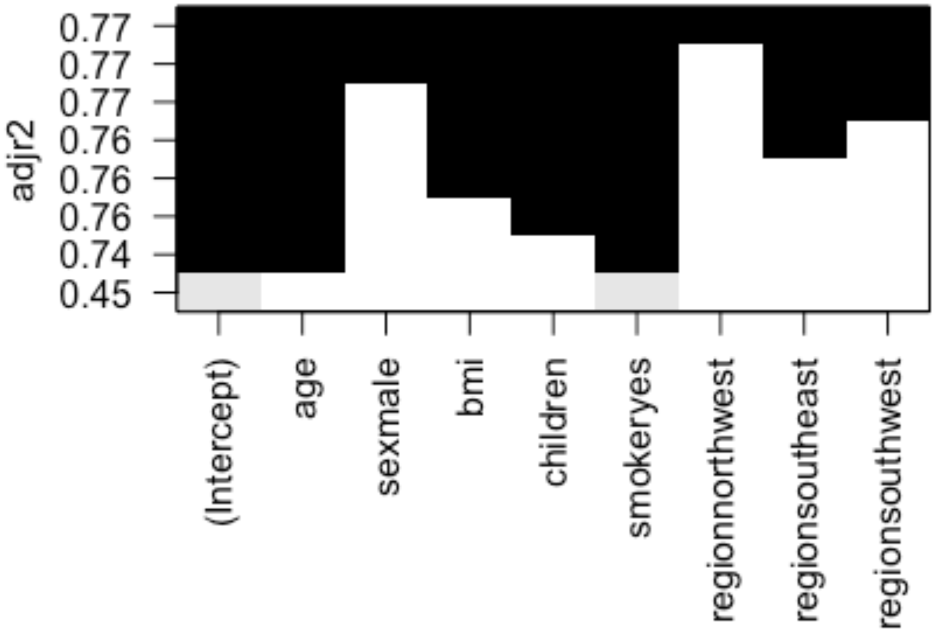
```
plot(model_backward_og,scale="bic")
```



Interpretation of the table above:

The Bayesian Information Criterion (BIC) plot for the backward selection indicates that the model including `smokeryes`, `age`, `bmi`, `children`, and `regionsoutheast` has the lowest BIC value, signifying it is the best-fitting model among the subsets evaluated.

```
plot(model_backward_og,scale="adjr2")
```



Interpretation of the table above:

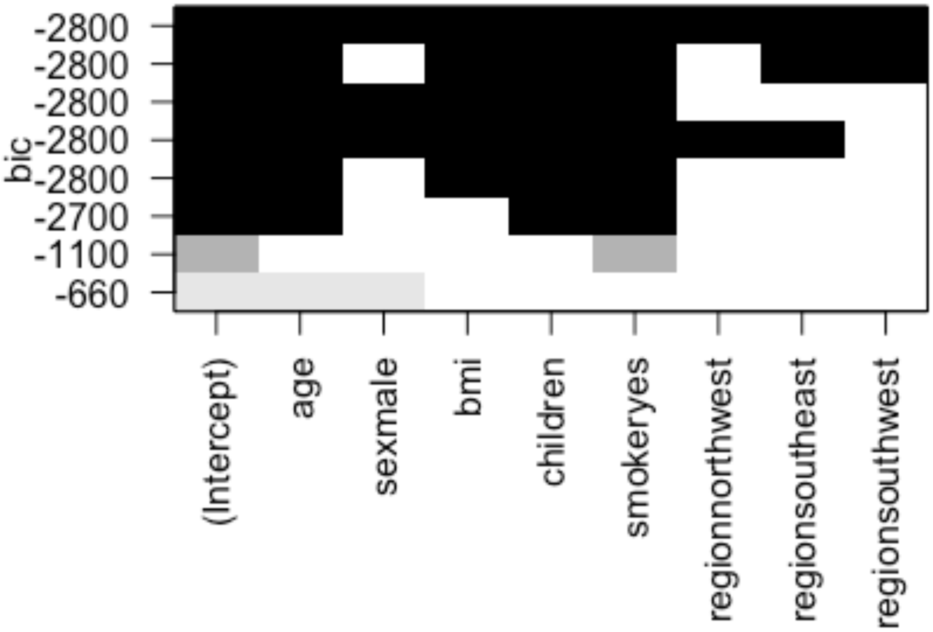
The adjusted R^2 plot for the backward selection indicates that the model including `smokeryes`, `age`, `bmi`, `children`, and `regionsoutheast` achieves the highest adjusted R^2 value, signifying it provides the best fit among the evaluated subsets.

```
## Seqrep stepwise regression
model_seqrep_og <- regsubsets(log_charges ~ age + sex + bmi + children +
smoker + region, method = "seqrep", data = Insurance)
summary(model_seqrep_og)

## Subset selection object
## Call: regsubsets.formula(log_charges ~ age + sex + bmi + children +
##     smoker + region, method = "seqrep", data = Insurance)
## 8 Variables  (and intercept)
##                   Forced in Forced out
## age                   FALSE      FALSE
## sexmale               FALSE      FALSE
## bmi                   FALSE      FALSE
## children              FALSE      FALSE
## smokeryes             FALSE      FALSE
## regionnorthwest       FALSE      FALSE
## regionsoutheast       FALSE      FALSE
## regionsouthwest       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: 'sequential replacement'
##          age sexmale bmi children smokeryes regionnorthwest
regionsoutheast
## 1  ( 1 ) " " " "     " " " "      "*"       " "             " "
## 2  ( 1 ) "*" "*"     " " " "      " "       " "             " "
## 3  ( 1 ) "*" " "     " " "*"      "*"       " "             " "
## 4  ( 1 ) "*" " "     "*" "*"      "*"       " "             " "
## 5  ( 1 ) "*" "*"     "*" "*"      "*"       " "             " "
## 6  ( 1 ) "*" " "     "*" "*"      "*"       " "             "*"
## 7  ( 1 ) "*" "*"     "*" "*"      "*"       "*"             "*"
## 8  ( 1 ) "*" "*"     "*" "*"      "*"       "*"             "*"
##          regionsouthwest
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) "*"
## 7  ( 1 ) " "
## 8  ( 1 ) "*"
```

Interpretation of the table above:

The sequential replacement subset selection algorithm indicates that all variables, starting from the most significant ones (smokeryes, age, bmi, children, etc.), should be included sequentially, with the best model including all variables.

```
plot(model_seqrep_og,scale="bic")
```



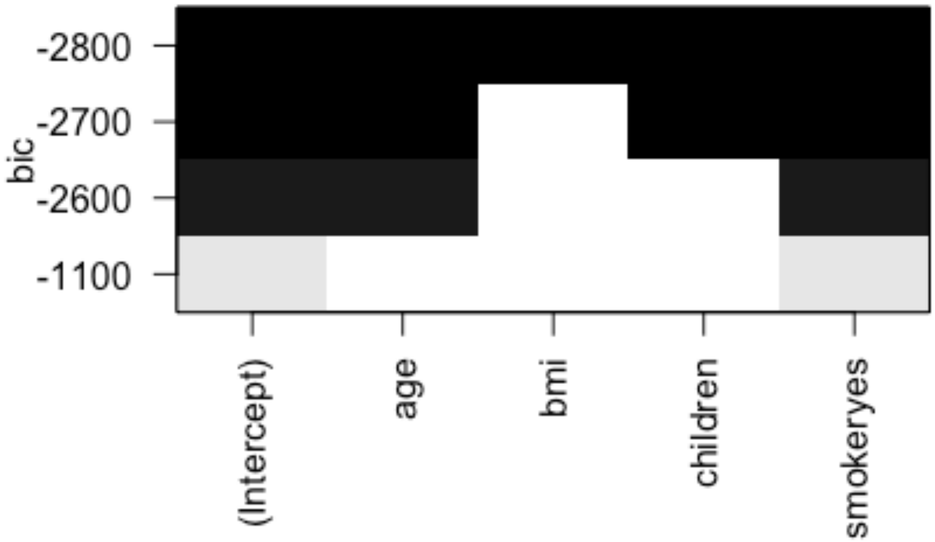
Interpretation of the plot above:

The Bayesian Information Criterion (BIC) plot for the sequential replacement selection indicates that the model including `smokeryes`, `age`, `bmi`, `children`, and `regionsoutheast` has the lowest BIC value, suggesting it is the best-fitting model among the evaluated subsets.

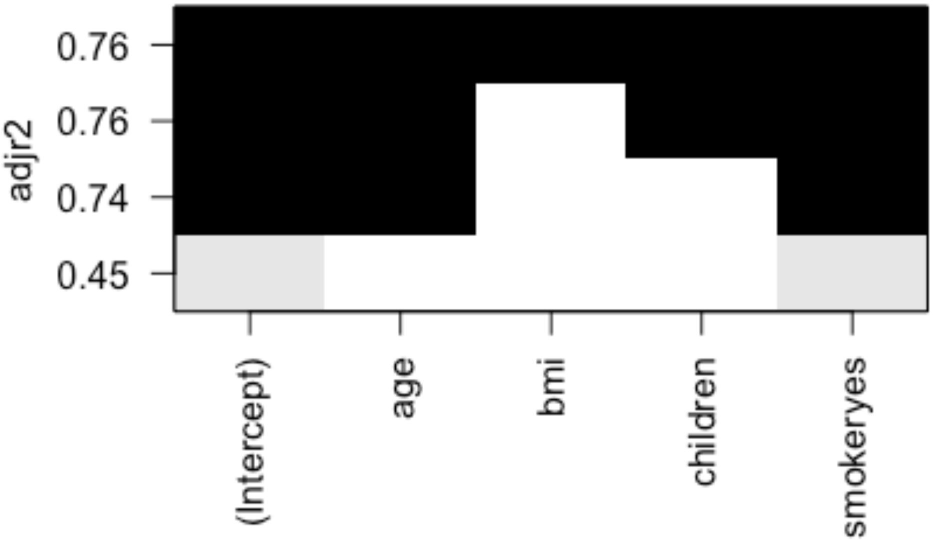
We also did the stepwise regression for the new model but we got the same results as displayed below. We have not interpreted the the results becuse they dont help build our model but here however included to show that the necessary analysis was perfromed.

```
## Forward stepwise regression
model_forward <- regsubsets(log_charges ~ age + bmi + children + smoker,
method = "forward", data = Insurance)
summary(model_forward)

## Subset selection object
## Call: regsubsets.formula(log_charges ~ age + bmi + children + smoker,
##     method = "forward", data = Insurance)
## 4 Variables  (and intercept)
##            Forced in Forced out
## age            FALSE      FALSE
## bmi            FALSE      FALSE
## children       FALSE      FALSE
## smokeryes     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: forward
##          age bmi children smokeryes
## 1  ( 1 ) " " " " " "      "*"
## 2  ( 1 ) "*" " " " "      "*"
## 3  ( 1 ) "*" " " "*"      "*"
## 4  ( 1 ) "*" "*" "*"      "*"
```

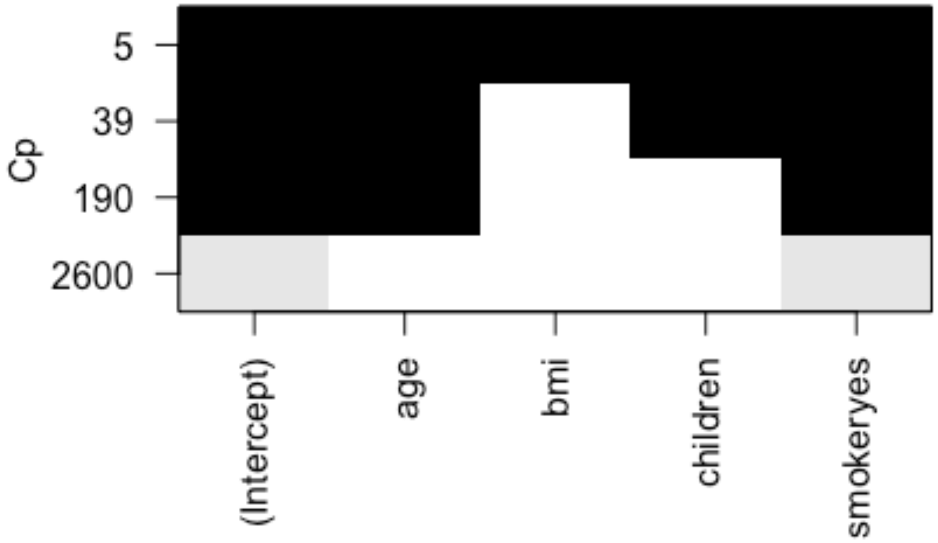
```
plot(model_forward,scale="bic")
```



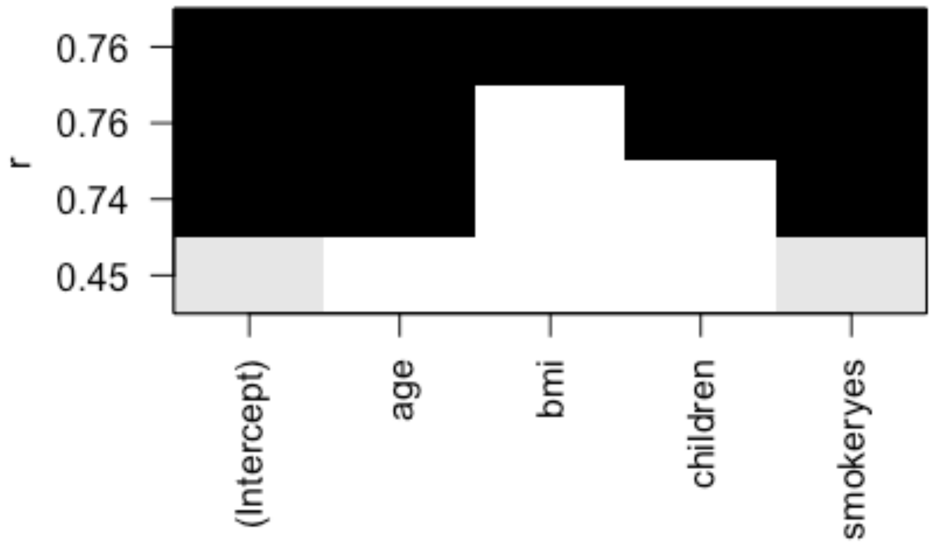
```
plot(model_forward,scale="adjr2")
```



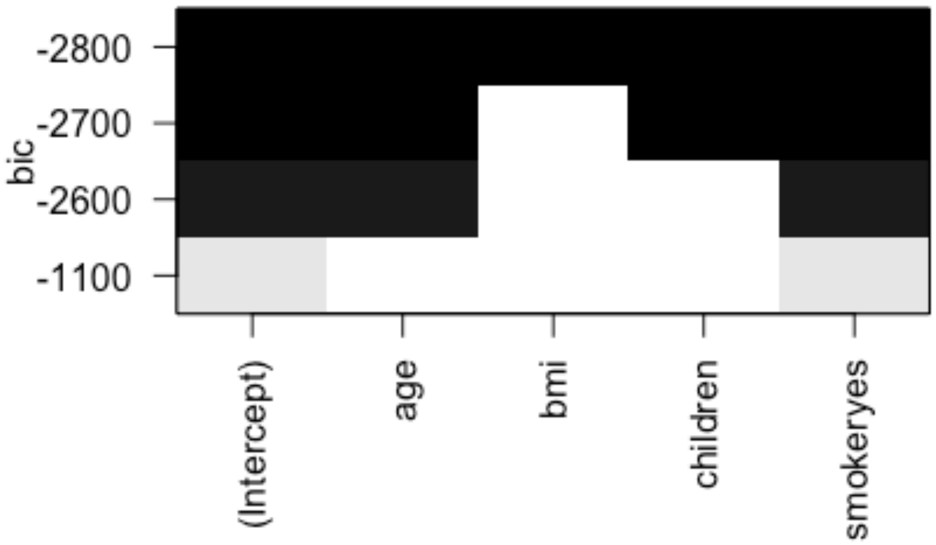
```
plot(model_forward,scale="Cp")
```



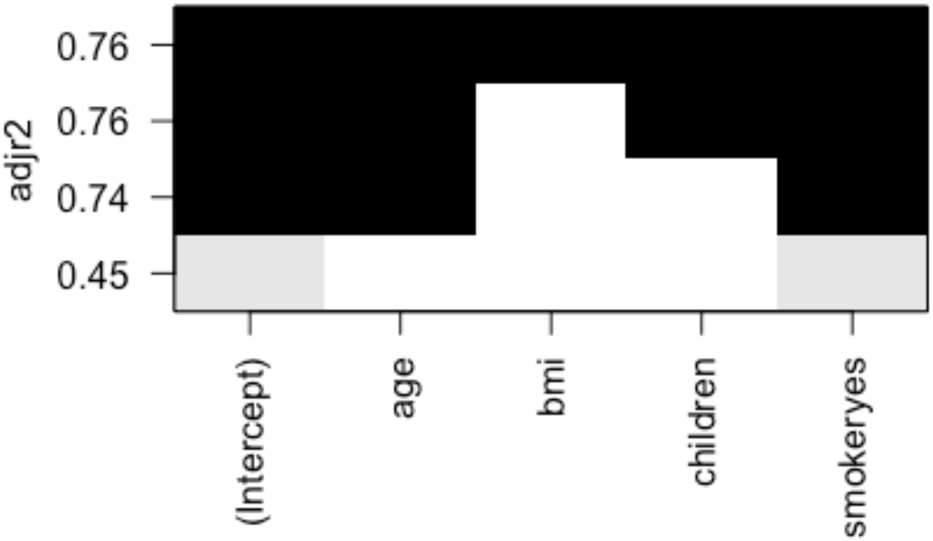
```
plot(model_forward,scale="r")
```



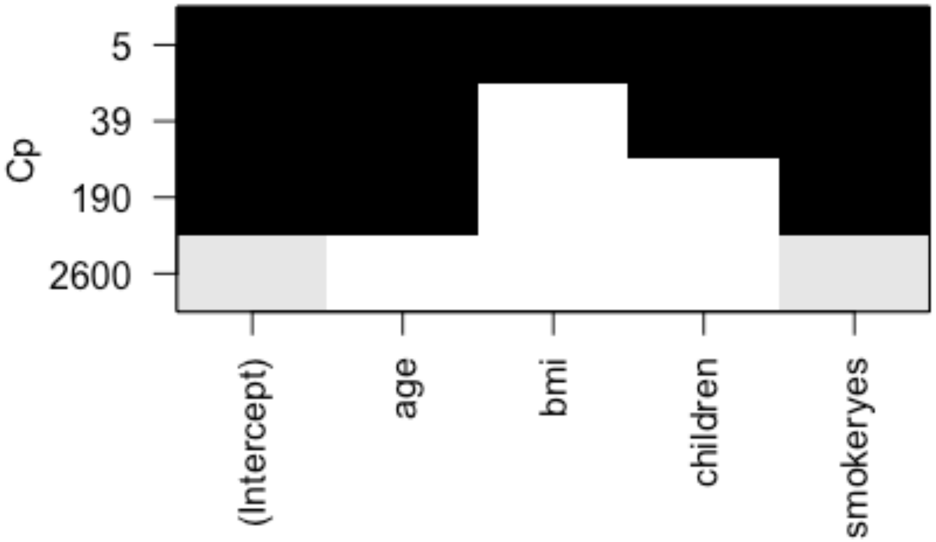
```
## Backward stepwise regression
model_backward <- regsubsets(log_charges ~ age + bmi + children + smoker,
method = "backward", data = Insurance)
summary(model_backward)

## Subset selection object
## Call: regsubsets.formula(log_charges ~ age + bmi + children + smoker,
##     method = "backward", data = Insurance)
## 4 Variables  (and intercept)
##            Forced in Forced out
## age            FALSE      FALSE
## bmi            FALSE      FALSE
## children       FALSE      FALSE
## smokeryes      FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: backward
##          age bmi children smokeryes
## 1  ( 1 ) " " " " " "      "*"
## 2  ( 1 ) "*" " " " "      "*"
## 3  ( 1 ) "*" " " "*"      "*"
## 4  ( 1 ) "*" "*" "*"      "*"

plot(model_backward,scale="bic")
```



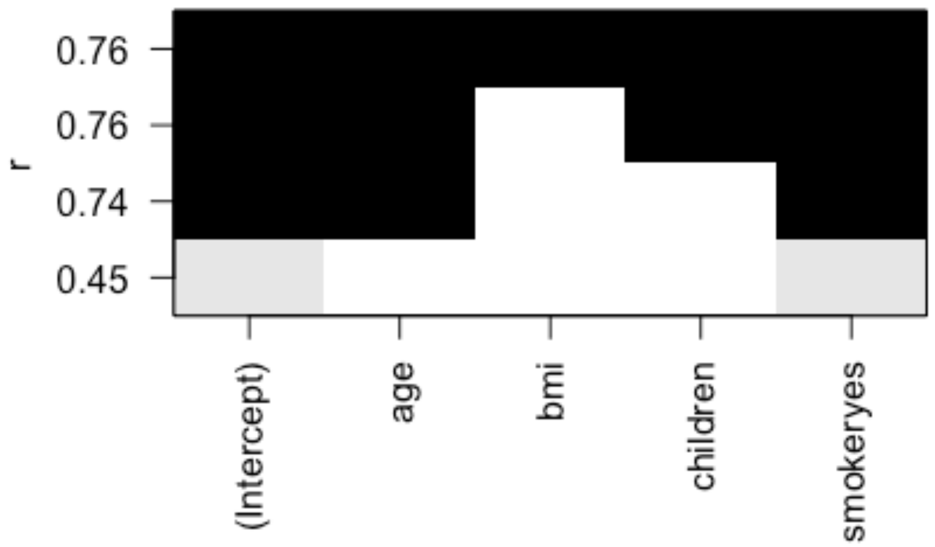
```
plot(model_backward,scale="adjr2")
```



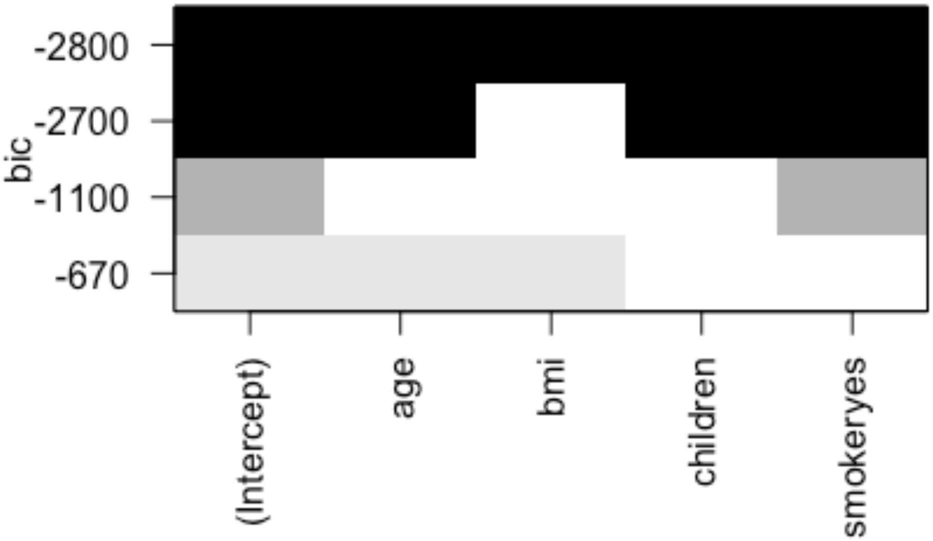
```
plot(model_backward,scale="Cp")
```



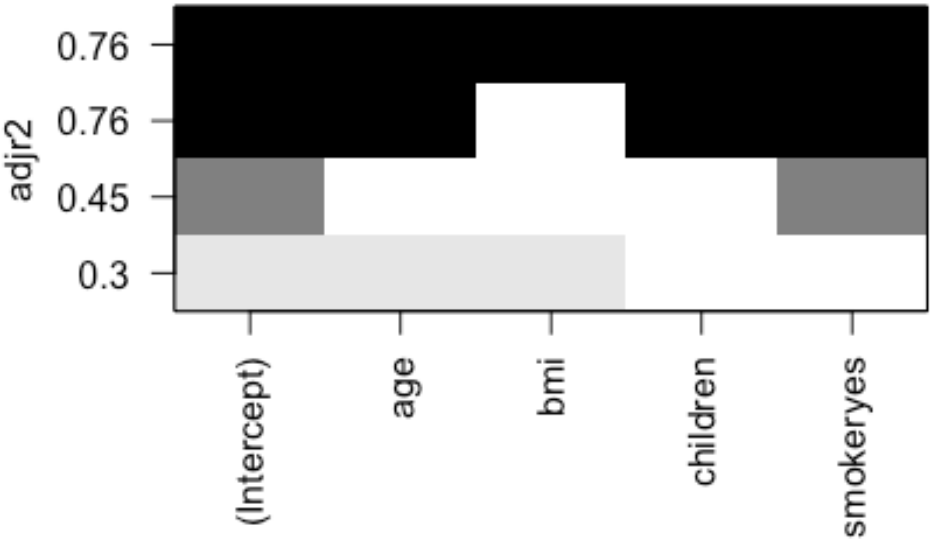
```
plot(model_backward,scale="r")
```



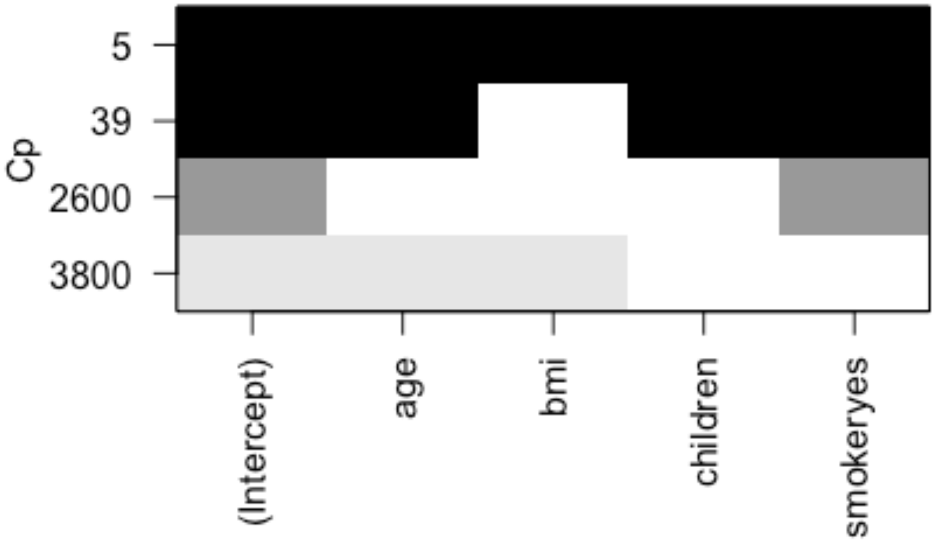
```
## Seqrep stepwise regression
model_seqrep <- regsubsets(log_charges ~ age + bmi + children + smoker,
method = "seqrep", data = Insurance)
summary(model_seqrep)

## Subset selection object
## Call: regsubsets.formula(log_charges ~ age + bmi + children + smoker,
##     method = "seqrep", data = Insurance)
## 4 Variables  (and intercept)
##           Forced in Forced out
## age           FALSE      FALSE
## bmi           FALSE      FALSE
## children      FALSE      FALSE
## smokeryes     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: 'sequential replacement'
##          age bmi children smokeryes
## 1  ( 1 ) " " " " " "      "*"
## 2  ( 1 ) "*" "*" " "      " "
## 3  ( 1 ) "*" " " "*"      "*"
## 4  ( 1 ) "*" "*" "*"      "*"

plot(model_seqrep,scale="bic")
```



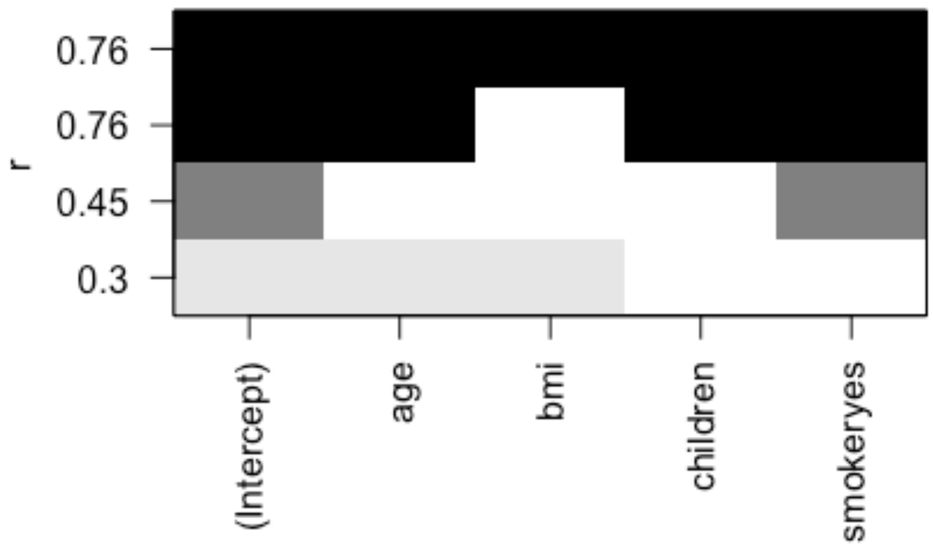
```
plot(model_seqrep,scale="adjr2")
```



```
plot(model_seqrep,scale="Cp")
```



```
plot(model_seqrep,scale="r")
```



0.76
0.76
0.45
0.3
r
(Intercept)
age
bmi
children
smokeryes