project

December 18, 2022

1 The Personal Attributes of A Patient that Influences One's Medical Costs

1.1 Introduction

Society has come a long way with the development in medical technology and the establishment of healthcare systems, and, without a doubt, there is a high cost associated with an individual's health and wellness. Healthcare costs are on a dramatic rise across numerous countries, and a study by the Journal of the American Medical Association has investigated some key factors associated with the increase such as the aging population and the disease prevalence in the region. Thus, an individual's medical costs can vary greatly from individual to individual and our project proposes, out of our interest, to build a model which can accurately predict one's medical costs. Through building a model of some number of input variables, which pertain to an individual's personal characteristics, which accurately predicts the response variable, the individual's medical cost and answer our question:

"What features would most accurately predict the medical costs, and what are the affects of those attributes?"

To assess this topic and to build our desired model, our group has selected a simulated dataset on the basis of demographic statistics from the US census bureau. The dataset presents a variety of attributes such as one response variable, "charges"; and six explanatory variables, which are "age", "sex", "bmi"(Body mass index), "children", "smoker", and "region", repectively.

1.2 Preliminary Results

1.2.1 Importing Modules

[1]:
```
library(latex2exp)
library(repr)
library(digest)
library(gridExtra)
library(mltools)
library(cowplot)
library(infer)
library(AER)
install.packages("glmtoolbox")
library(glmtoolbox)
suppressWarnings(suppressMessages(library(tidyverse)))
```

```
suppressWarnings(suppressMessages(library(dplyr)))
suppressWarnings(suppressMessages(library(broom)))
suppressWarnings(suppressMessages(library(GGally)))
suppressWarnings(suppressMessages(library(leaps)))
suppressWarnings(suppressMessages(library(glmnet)))
suppressWarnings(suppressMessages(library(faraway)))
suppressWarnings(suppressMessages(library(base)))
```

```
Loading required package: car

Loading required package: carData

Loading required package: lmtest

Loading required package: zoo


Attaching package: ‘zoo’


The following objects are masked from ‘package:base’:

    as.Date, as.Date.numeric


Loading required package: sandwich

Loading required package: survival

Updating HTML index of packages in '.Library'

Making 'packages.html' …
 done
```

1.2.2 Downloading Dataset

[2]:
```
insurance <- read_csv("https://raw.githubusercontent.com/LukeXboy/
  ↪STAT-301-Group-Project/main/insurance.csv")
head(insurance)
```

```
Rows: 1338 Columns: 7
  Column specification

Delimiter: ","
chr (3): sex, smoker, region
dbl (4): age, bmi, children, charges
```

```
  Use `spec()` to retrieve the full column specification for this
data.
  Specify the column types or set `show_col_types = FALSE` to quiet
this message.
```

A tibble: 6 × 7

age <dbl>	sex <chr>	bmi <dbl>	children <dbl>	smoker <chr>	region <chr>	charges <dbl>
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
31	female	25.740	0	no	southeast	3756.622

Table 1.1: Preview of Orginial Insurance Dataset

1.2.3 Feature Descriptions and Wrangling the Dataset

We look into the features of the dataset to see if we can wrangle the raw data into more comprehensable and useable forms.

The dataset contains the following variables relating to the patient: - age - sex - bmi: Represents the Body Mass Index - children: Number of children or dependents covered by the health insurance - smoker: If the patient is a smoker - region: Beneficiary's residentual area in the United States of America - charges: Individual medical costs billed by the health insurance company **(response)**

[3]:
```
#Produces further information about the feature values within the dataset
summary(insurance)
```

```
      age            sex                 bmi           children
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
 Median :39.00   Mode  :character   Median :30.40   Median :1.000
 Mean   :39.21                      Mean   :30.66   Mean   :1.095
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
 Max.   :64.00                      Max.   :53.13   Max.   :5.000
    smoker             region             charges
 Length:1338        Length:1338        Min.   : 1122
 Class :character   Class :character   1st Qu.: 4740
 Mode  :character   Mode  :character   Median : 9382
                                       Mean   :13270
                                       3rd Qu.:16640
                                       Max.   :63770
```

The summary results suggests that some of the variables in the dataset could be better represented as categorical variables for the purposes of our proposal; thus, we apply the changes for the feature values of sex, smoker, and region variables.

[4]:
```
#Combining similar levels of region for future feature selection
insurance_factor <- insurance %>%
    mutate(region = ifelse(insurance$region == "northwest" | insurance$region␣
 ↪== "northeast", "north", "south"))
head(insurance_factor)

#Mutating the features to be categorical variables for further analysis
insurance_factor <- insurance_factor %>%
    mutate(sex = as.factor(sex)) %>%
    mutate(smoker = as.factor(smoker)) %>%
    mutate(region = as.factor(region))

#To check the feature values after mutating the variables
summary(insurance_factor)
```

A tibble: 6 × 7

age <dbl>	sex <chr>	bmi <dbl>	children <dbl>	smoker <chr>	region <chr>	charges <dbl>
19	female	27.900	0	yes	south	16884.924
18	male	33.770	1	no	south	1725.552
28	male	33.000	3	no	south	4449.462
33	male	22.705	0	no	north	21984.471
32	male	28.880	0	no	north	3866.855
31	female	25.740	0	no	south	3756.622

```
      age            sex           bmi           children     smoker
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
 Median :39.00                Median :30.40   Median :1.000
 Mean   :39.21                Mean   :30.66   Mean   :1.095
 3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000
 Max.   :64.00                Max.   :53.13   Max.   :5.000
   region       charges
 north:649   Min.   : 1122
 south:689   1st Qu.: 4740
             Median : 9382
             Mean   :13270
             3rd Qu.:16640
             Max.   :63770
```

Table 1.2 (top): Preview of Insurance Dataset with Mutated Categorical Variables

Through Table 1.2, it is clear that there are discrete values or levels for the changed variables. We confirm our speculations and implement the changes of wrangling the data set by mutating the variables to be categorical.

[5]:
```
#Implementing changes
insurance <- insurance_factor

#Printing number of rows related to the database
```

```
cat("Number of instances in the insurance dataset: ", nrow(insurance), "\n")
cat("Number of instances with missing values in the dataset: ", sum(is.
 ↪na(insurance)), "\n")
```

```
Number of instances in the insurance dataset:  1338
Number of instances with missing values in the dataset:  0
```

Our dataset is now wrangled to accurately represent our data and contains no missing values, and ready to be split into the training and testing sets.

1.2.4 Variable Distributions and Correlations

[6]:
```
#Historgram to visualize response variable's, charges, distribution
options(repr.plot.width = 10, repr.plot.height = 8)
charges_hist <- insurance %>%
    ggplot(aes(x = charges)) +
    geom_histogram(bins = 25, boundary = 0, colour = 'white') +
    labs(title = "Training Set Distribution of Patient's Medical Costs",
         x = "Charges (USD)") +
    theme(text = element_text(size = 16)) +
    geom_vline(xintercept = mean(insurance$charges), colour = 'red', size = 3)
charges_hist
```

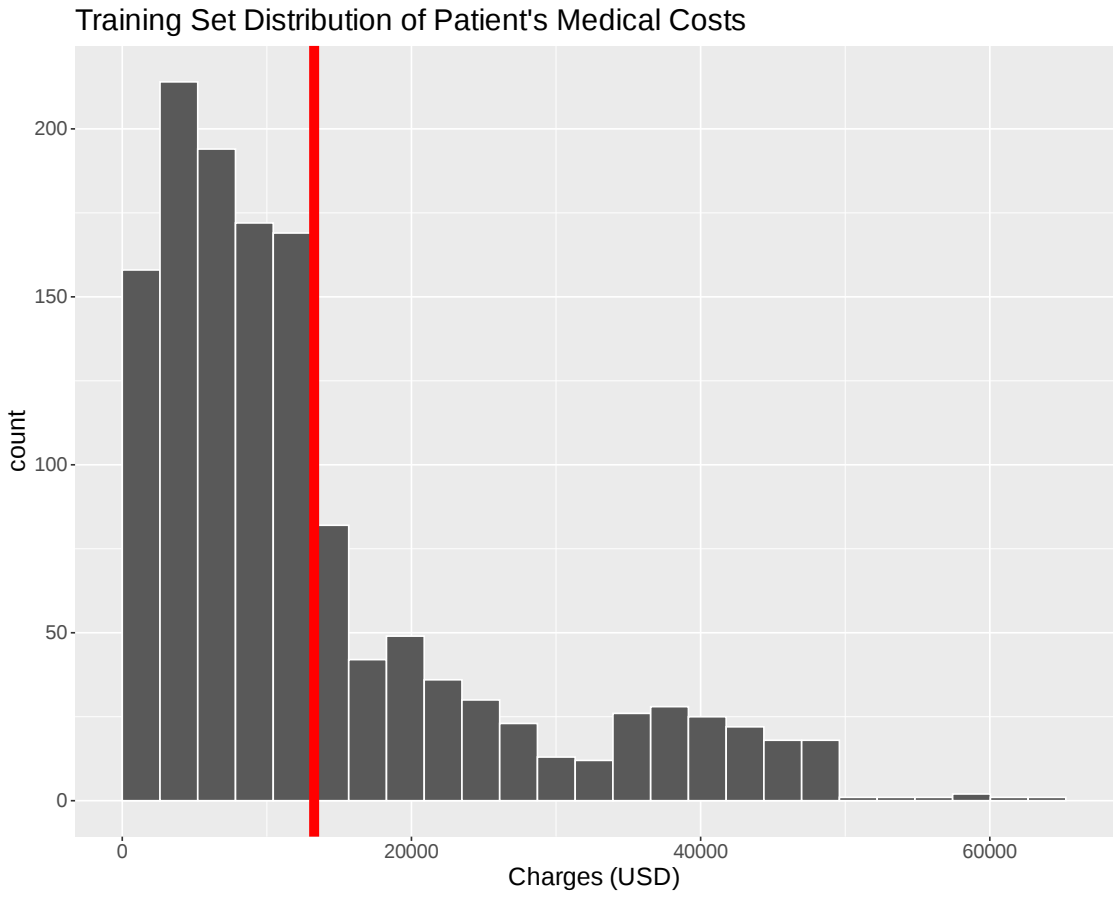

Figure 1.1: Histogram Displaying Training Set Distribution of Response Variable within Training , Charges, which represents the Medical Costs of the Patient. The Red Vertical Line Indicates the Variable Mean. We notice that Charges has a Left-skewed Distribution.

[7]:
```
#Comparing correlation across all variables with a GGPairs plot
options(repr.plot.width = 15, repr.plot.height = 15)
all_corr_plot <- insurance %>%
    ggpairs() +
  labs(title = "Distribution and Correlation across Variables",
       x = "Variable",
       y = "Variable") +
  theme(
    text = element_text(size = 15),
    plot.title = element_text(face = "bold"),
    axis.title = element_text(face = "bold"),
    axis.text.x = element_text(angle = 70, vjust = 1, hjust=1) #To ensure␣
 ↪labels do not overlap
  )
all_corr_plot
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

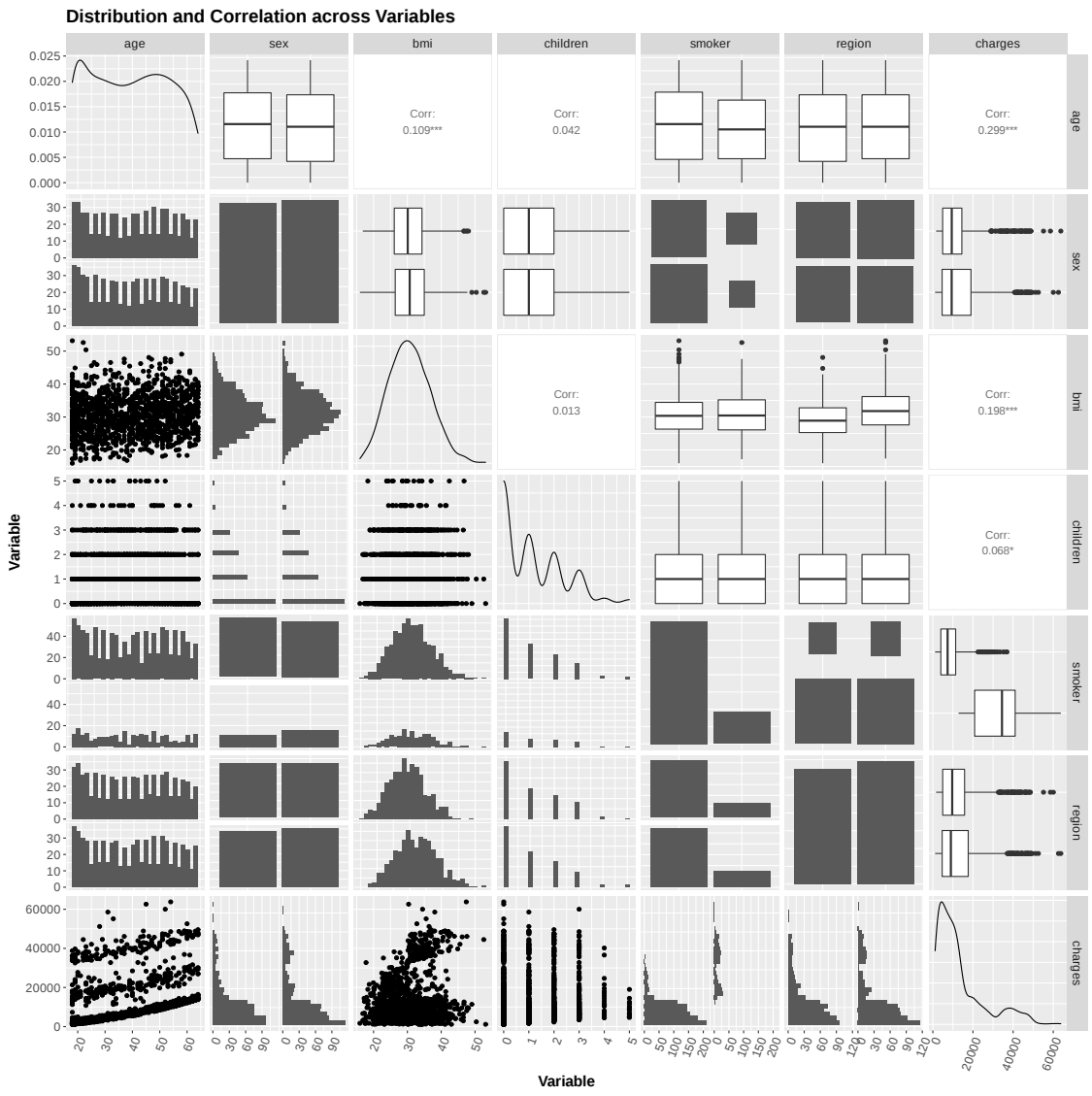

Figure 1.2: Distribution and Correlations Across Variables in the Dataset.

1.2.5 Splitting Data into Training and Testing Sets

[8]:
```
#Splitting the data by labelling the data by ID
set.seed(123456)
insurance$ID <- rownames(insurance)
insurance_training <- sample_n(insurance, size = nrow(insurance) * 0.70,␣
 ↪replace = FALSE)
insurance_testing <- anti_join(insurance, insurance_training, by = "ID")

insurance_training <- select(insurance_training, -ID)
insurance_testing <- select(insurance_testing, -ID)
```

```
head(insurance_training)
head(insurance_testing)
```

A tibble: 6 × 7

age <dbl>	sex <fct>	bmi <dbl>	children <dbl>	smoker <fct>	region <fct>	charges <dbl>
32	male	31.50	1	no	south	4076.497
59	male	27.50	1	no	south	12333.828
42	female	25.30	1	no	south	7045.499
42	male	34.10	0	no	south	5979.731
18	male	35.20	1	no	south	1727.540
18	female	26.73	0	no	south	1615.767

A tibble: 6 × 7

age <dbl>	sex <fct>	bmi <dbl>	children <dbl>	smoker <fct>	region <fct>	charges <dbl>
19	female	27.900	0	yes	south	16884.924
32	male	28.880	0	no	north	3866.855
37	female	27.740	3	no	north	7281.506
60	female	25.840	0	no	north	28923.137
23	male	23.845	0	no	north	2395.172
63	female	23.085	0	no	north	14451.835

Table 1.3 (top): Preview of the training set of the dataset

Table 1.4 (bottom): Preview of the testing set of the dataset

[9]:
```
#Printing number of rows within split dataset
cat("Number of instances in the training set: ", nrow(insurance_training), "\n")
cat("Number of instances in the testing set: ", nrow(insurance_testing), "\n")
```

```
Number of instances in the training set:  936
Number of instances in the testing set:  402
```

We split the dataset into a training and testing set as our project proposes to produce a predictive model which predicts charges on the new observations of patients.

1.3 Methods: Plan

With our proposal to build a predictive model, we have split the data into a training and testing set as seen by Table 1.3 and Table 1.4. Using the same set of data for model selection and for evaluating will create a biased result with possibly inflated type 1 errors thus we eliminate the problem by splitting the dataset. The training set will be used for model selection and finding a subset of input variables that are relevant to the response, and the testing set will be used to compute out-of-sample predictions and to evaluate the goodness of the models.

The dataset is reliable as it produced or simulated for the purposes of prediction, and it is helpful as it contains a variety of variables of different types and ranges, as deduced from the distribution of variables against charges (Figure 1.2). The plot also suggests that there is no correlation across input variables and response variables; however, this would be further confirmed through computing the General Variance Inflation Factors in the project.

The proposed project plans to build an additive predictive multi-linear regression model, and we comparing a full and reduced model. We will first build a full multilinear regression model, where all input variables in the dataset are used to predict the response. Through forward feature selection, we will find a subset of variables in a model where the Mallow's Cp is minimized. To select the model which best fits and predicts the data, the root mean squared error will be compared between models.

We expect to build a multilinear regression model which accurately predicts a patient's medical cost given some inputs of a patient's personal information.

Hopefully our project impacts the greater society as our results could help medical organizations or particular government departments decide on their financial assignment to provide help to those in need. Moreover, can assist an individual to predicting their own healthcare costs to prepare financial plans.

1.4 Results

1.4.1 Estimating Multilinear Regression Model

[10]:
```
#Estimate additive MLR model using all variables in the dataset
insurance_full_MLR <- lm(charges ~ ., data = insurance_training)
tidy(insurance_full_MLR)
```

A tibble: 7 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-10653.6330	1151.96202	-9.2482502	1.524723e-19
age	241.0604	14.44958	16.6828650	7.448097e-55
sexmale	-313.8132	402.86313	-0.7789573	4.362032e-01
bmi	317.5317	33.42351	9.5002491	1.719942e-20
children	513.3964	168.26007	3.0512073	2.343990e-03
smokeryes	23962.2000	504.52291	47.4947711	9.059586e-251
regionsouth	-1004.6347	412.39565	-2.4360943	1.503383e-02

Table 1.5: Components of Full Model

[11]:
```
#Obtaining out-of-sample predictions with the full model
insurance_test_pred_full_MLR <- predict(insurance_full_MLR, newdata =␣
↪insurance_testing)
head(insurance_test_pred_full_MLR)
```

1 25743.2140450543 **2** 5916.80188595368 **3** 8614.12006448715 **4** 12015.0101810138 **5** 2148.48619451487 **6** 11863.3916231639

[12]:
```
#Storing model evaluations
insurance_RMSE_models <- tibble(
    Model = "Full Additive Regression",
    R_MSE = rmse(
        preds = insurance_test_pred_full_MLR,
        actuals = insurance_testing$charges)
    )
```

```
insurance_RMSE_models
```

A tibble: 1 × 2

Model	R_MSE
<chr>	<dbl>
Full Additive Regression	5888.656

Table 1.6 : The Model and it's Evaluation on the Testing Set

1.4.2 Mutlicollinearity

[13]:
```
#Check if there is multicollinearity within the dataset using General Variance Inflation Factors
vif_insurance <- gvif(insurance_full_MLR)
round(vif_insurance, 3)
```

```
           GVIF df GVIF^(1/(2*df))
age      1.0143  1          1.0071
sex      1.0062  1          1.0031
bmi      1.0625  1          1.0308
children 1.0018  1          1.0009
smoker   1.0053  1          1.0026
region   1.0511  1          1.0252
```

A matrix: 6 × 3 of type dbl

	GVIF	df	GVIF^(1/(2*df))
age	1.014	1	1.007
sex	1.006	1	1.003
bmi	1.062	1	1.031
children	1.002	1	1.001
smoker	1.005	1	1.003
region	1.051	1	1.025

Table 1.7: General Variance Inflation Factors for the Input Variables in the Full Model

In combination of the general variance inflation factors computed and the visualization of distribution and correlation across variables in Figure 1.2, we conclude that there is no multicollinearity across the datqaset.

1.4.3 A Smaller Model using Forward Selection

[14]:
```
#Using the forward selection algorithm to find a reduced LR model
insurance_forward_sel <- regsubsets(
    x = charges ~ ., nvmax = 6,
    data = insurance_training,
    method = "forward",
    )

insurance_forward_summary <- summary(insurance_forward_sel)
insurance_forward_summary
```

```
Subset selection object
Call: regsubsets.formula(x = charges ~ ., nvmax = 6, data = insurance_training,
    method = "forward", )
6 Variables  (and intercept)
            Forced in Forced out
age             FALSE      FALSE
sexmale         FALSE      FALSE
bmi             FALSE      FALSE
children        FALSE      FALSE
smokeryes       FALSE      FALSE
regionsouth     FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: forward
         age sexmale bmi children smokeryes regionsouth
1  ( 1 ) " " " "     " " " " "      "*"       " "
2  ( 1 ) "*" " "     " " " " "      "*"       " "
3  ( 1 ) "*" " "     "*" " " "      "*"       " "
4  ( 1 ) "*" " "     "*" "*" "      "*"       " "
5  ( 1 ) "*" " "     "*" "*" "      "*"       "*"
6  ( 1 ) "*" "*"     "*" "*" "      "*"       "*"
```

[15]:
```r
#Summary table of measures for all forward selection models
insurance_forward_summary_df <- tibble(
    n_input_variables = 1:6,
    RSS = insurance_forward_summary$rss,
    ADJ.R2 = insurance_forward_summary$adjr2,
    BIC = insurance_forward_summary$bic,
    Cp = insurance_forward_summary$cp
    )
insurance_forward_summary_df
```

A tibble: 6 × 5

n_input_variables <int>	RSS <dbl>	ADJ.R2 <dbl>	BIC <dbl>	Cp <dbl>
1	50967889939	0.6233669	-901.3072	418.680882
2	38792983690	0.7130275	-1149.9528	98.038270
3	35659260671	0.7359263	-1221.9509	16.992655
4	35301815892	0.7382926	-1224.5390	9.520146
5	35078675062	0.7396672	-1223.6326	5.606774
6	35055778451	0.7395570	-1217.4021	7.000000

Table 1.8: Summary of Measures across Models Computed in the Forward Selection Algorithm

[16]:
```r
#Illustrates Cp of forward selection models
options(repr.plot.width = 8, repr.plot.height = 8)
plot(summary(insurance_forward_sel)$cp,
  main = "Cp for forward selection",
  xlab = "Number of Input Variables", ylab = "Cp", type = "b", pch = 19,
  col = "red"
```

```
)
```

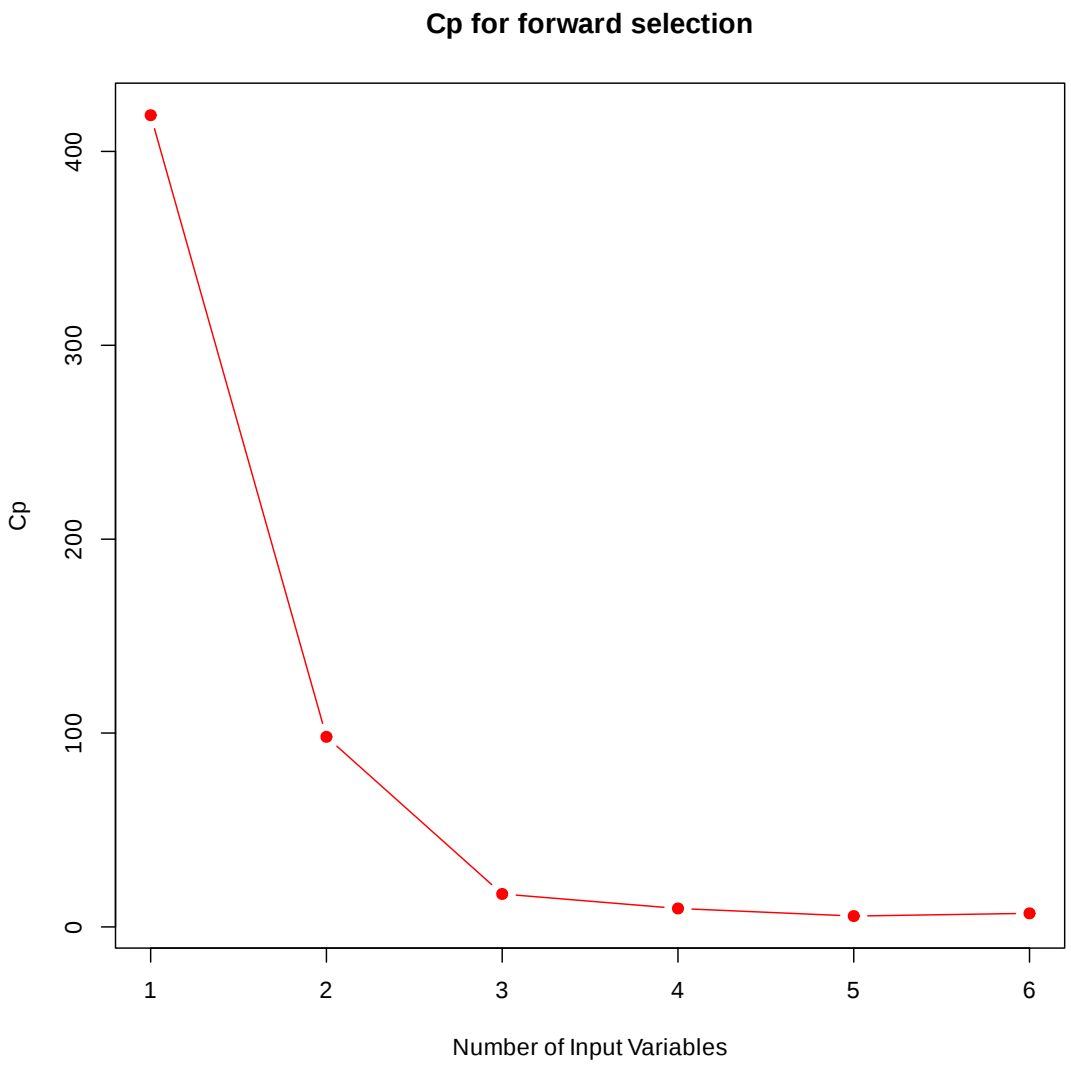

Figure 1.3: Plot of Mallow's Cp across Forward Selection's Models

Based of the Figure 1.3 and Table 1.8, we select the model with 5 input variables as it minimizes the Cp which is an estimate for the test mean squared error. The selected model from forward selection includes the input variables age, bmi, children, smoker, and region.

[17]:
```
#Estimation with selected model from forward selection
insurance_reduced_MLR <- lm(charges ~ age + bmi + children + smoker + region,␣
 ↪data = insurance_training
)
```

```
#Out-of-sample predictions using the reduced model
insurance_test_pred_reduced_MLR <- predict(insurance_reduced_MLR, newdata =␣
↪insurance_testing)
head(insurance_test_pred_reduced_MLR)
```

1 25560.9156966106 **2** 6075.9940649472 **3** 8459.58829171677 **4** 11871.3874310342 **5** 2309.21870039048 **6** 11722.9980702746

[18]:
```
#Combining reduced model's results
insurance_RMSE_models <- rbind(
    insurance_RMSE_models,
    tibble(
    Model = "Reduced Forward Selection Additive Regression",
    R_MSE = rmse(
        preds = insurance_test_pred_reduced_MLR,
        actuals = insurance_testing$charges)
    )
    )
insurance_RMSE_models
```

A tibble: 2 × 2

Model <chr>	R_MSE <dbl>
Full Additive Regression	5888.656
Reduced Forward Selection Additive Regression	5882.354

[19]:
```
#Obtaining coefficents of the reduced model
tidy(summary(insurance_reduced_MLR))
```

A tibble: 6 × 5

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-10793.5098	1137.63963	-9.487635	1.916535e-20
age	241.3615	14.44135	16.713221	4.956678e-55
bmi	316.6875	33.39887	9.481981	2.013649e-20
children	512.6034	168.22142	3.047195	2.375121e-03
smokeryes	23935.7995	503.27681	47.559909	2.824810e-251
regionsouth	-1002.8241	412.30190	-2.432257	1.519283e-02

Table 1.9: Estimated Coefficients of the Reduced Model

1.5 Methods: Results

Based off Figure 1.1, individuals in the United States on average have a medical charge of 150,000 USD and with the response variable in our data set having a left-skew distribution we note that majority of individuals have a medical costs of less than 150,000 USD.

To build a predictive linear regression model for our project, where some subset of the dataset's input variables (age, sex, bmi, children, smoker, region) predict the response variable, charges, we did a 7:3 splitting of the dataset into training and testing sets. The training set was used to build predictive models and for feature selection using the forward selection algorithm and the testing set was used to evaluating the out-of-sample prediction performance of the models.

We first built and evaluated a full model that uses all input variables to predict the response. With the full model and general variance inflation factors, we checked if the input variables were correlated and if there was a violation across the dataset. General variance inflation factors were used over variance inflation factors as there are categorical variables in our dataset. Table 1.7, which had relatively low GVIF values, in combination with Figure 1.2, led us to conclude that there is no multicollinearity present in the dataset.

For feature selection, we used the forward selection algorithm to find a subset of variables to predict the response for our reduced linear regression model. We chose our model from the forward selection based off Mallow's Cp as it is an appropriate estimate for the test mean squared error. As we want our models to accurately predict the medical costs of an individual, we noticed that the model with the lowest Cp is when 5 input variables were selected, specifically the input variables age, bmi, children, smoker, and region. Thus the reduced model excludes the variable sex in respect to the full model, which can be seen intuitive as the medical costs of an individual should not differ based off the individual's sex.

Through evaluating the root mean squared error of both the full and reduced model, we see that the full model has an RMSE of \$5888.656 USD squared and the reduced model has an RMSE of \$5882.354 squared. They are not largely different but we can still conclude that the model we get from the forward selection has a better performance than the full model. Moreover, this small different can be predicted as Table 1.8 and Figure 1.3 showcase that there is little difference in Mallows Cp from the model with 5 input variables and 6 input variables.

1.6 Discussion

In our analysis, we found that our selected reduced model had a lower RMSE than the full model, thus the reduced model is our most preferred model. Thus implies that the age, bmi, number of children, the participation in smoking, and region of an individual impact and accurately predict their medical costs. Based off the results of Table 1.9, we can deduced an estimate or expectation of how these attributes affect the medical costs of an individual. For example, holding all other variables in the reduced model constant, we expect an increase of \$513 USD in an individual's medical costs for every 1% increase in an individual's child or dependent. From Table 1.9, we see that there is a strong and large difference in expected medical costs between being a smoker and being a non-smoker. This could imply those who smoke are at higher risk for various diseases thus increasing their medical costs.

The healthcare system is an important part of a working society and one's health is important to the people within the society. We hope that our analysis and model can help shape the law to provide better services to the poeple in need and help prepare people for their financial burden that come with medical costs. However, our analysis is imperfect, and there is still room for improvement as our study is based on a limited number of factors that could have an impact on medical costs. With more data and personal attributes about an individual, we may find other features that may exist and influence our response variable, charges. We can also improve the model by seeing if interactions are involved or by building a logistic regression model with shrinkage. In addition, we may miss some confounding factors beyond the variables of the data set that might also affect the results.

In the future, researchers can collect data including more possible variables and try to locate whether multicollinearity exists. Some other model selection methods such as the Akaike information criterion(AIC), and Bayesian information criterion(BIC) can also be used. We expect sincerely

that a more accurate predictive model will be constructed in order to better manage an individual's spending in healthcare.

Our project could lead to the future and further discoveries of attributes that impact one's medical cost and reason why the medical costs in the United States continue to increase drastically. Moreover, we hope that this study can also be used to find outliers in the healthcare system who are in fraudulent insurance plans, where their medical costs are severely greater than others for no meaningful reason.

1.7 References

Huang, J. Z. (2014). An introduction to statistical learning: With applications in R by gareth james, trevor hastie, robert tibshirani, daniela witten. Journal of Agricultural, Biological, and Environmental Statistics, 19(4), 419. https://doi.org/10.1007/s13253-014-0179-9

Lantz, B., & O'Reilly for Higher Education. (2015). Machine learning with R: Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R. Packt Publishing.

JAMA: Journal of the American Medical Association. "Factors Associated with Increases in US Health Care Spending, 1996–2013."

Word Count: 1494