UNIVERSITI KEBANGSAAN MALAYSIA

The National University of Malaysia

Medical Expenses Analysis Report

STQD6214

An Hongyu P120996   Liu Baichuan P120378   He Yawen P124878

Catalogue

1. Introduction

1.1 Background

1.1.1 Health Insurance in the US

Health insurance is a means for financing a person’s or persons' health care expenses. In the US the majority of people have private health insurance, usually obtained through a current employer, and the minority are covered through government sponsored programs.

The insurer calculates premiums for their insurance policies relying on two primary factors - the cost the insurer predicts to pay under their policies, and the cost of operating particular policies or plans. The cost of medical expenses are calculated in many ways - the policyholder's health status, region of residence, employment status, and wages can all be included in the estimate.

1.1.2 Regression analysis

In terms of statistical methods, regression analysis is often used to estimate healthcare costs and calculate insurance premiums. A policyholder's medical expenses can be influenced by a myriad of factors, such as their habits, chronic illnesses, age, economic factors, occupational hazards, place of residence, and so on. Regression analysis can be used to identify the factors that are significant in their influence on medical costs. Regression analysis can also be used to predict the true cost of an insurance policy, allowing for insurance companies to set competitive prices. Setting the same price for all policyholders is not a competitive strategy as those with low expenses would overpay and possibly leave the service, and those with high expenses would remain using the service and make a loss of the insurance company. Regression models are tools that can be used to establish proper classification systems that offer a fair price to customers and maximise the company's profits.

1.2 Dataset

The dataset for this report comes from the book Machine Learning with R by Brett Lantz and is in the public domain. The dataset includes information about the insurance policy holder, their dependents, and their medical expenses throughout a year.

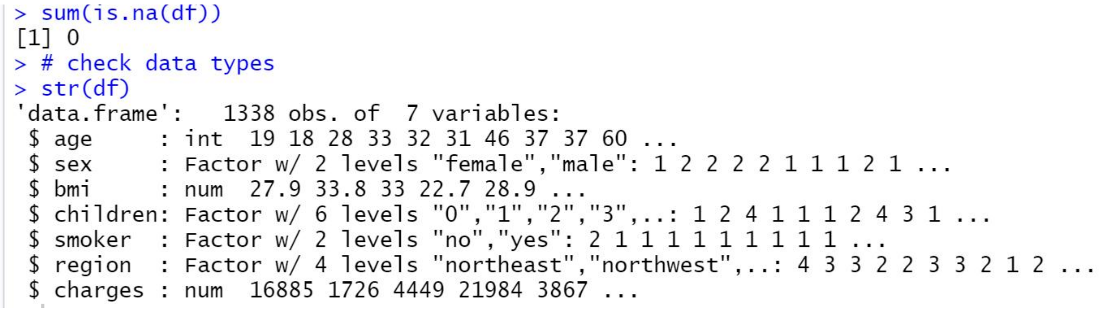
- **Age**: Age of primary policyholder.
- **Sex**: Sex of the policy policyholder.
- **BMI**: Body Mass Index of policyholder, defined as the body mass divided by the square of the body height (kg/m2).
- **Smoker status**: Whether the policyholder is a smoker or a non-smoker.
- **Children**: Number of children/dependents covered in the policy.
- **Region of residence**: Residential areas of the policy holder (in the US) - North East, South East, South West, North West.
- **Charges**: Yearly medical expenses billed by the medical insurance provider ($).

1.3 Objectives

- To determine if there is a relationship between attributes and medical costs.
- To determine if there a significant difference in medical costs between different groups.
- To fit a multiple linear regression to predict costs.

2. Import and Pre-processing

2.1 Check the Dataset



```
> sum(is.na(df))
[1] 0
> # check data types
> str(df)
'data.frame':	1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: Factor w/ 6 levels "0","1","2","3",..: 1 2 4 1 1 1 2 4 3 1 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

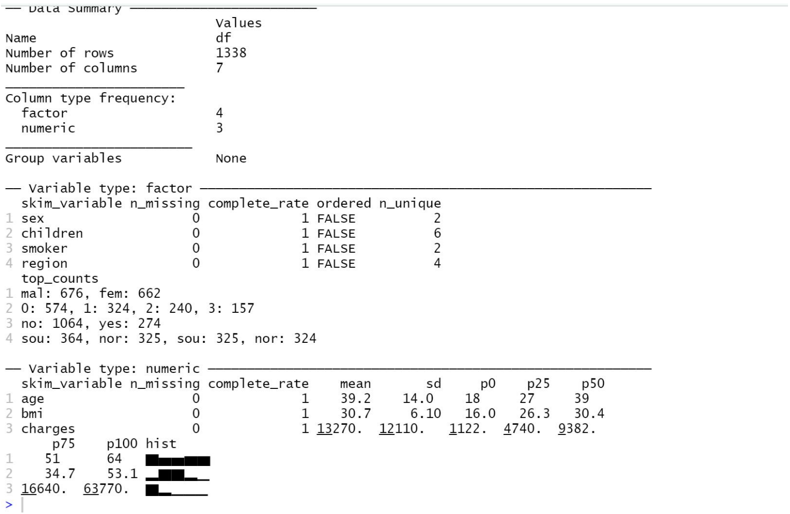
Figure 2-1

Summary

- There are no missing values.
- All columns are assigned correct data types.

3. Exploratory Data Analysis

3.1 Overview



```
── Data Summary ────────────────────────
                           Values
Name                       df
Number of rows             1338
Number of columns          7
_______________________
Column type frequency:
  factor                   4
  numeric                  3
________________________
Group variables            None

── Variable type: factor ──────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique
1 sex                   0             1 FALSE          2
2 children              0             1 FALSE          6
3 smoker                0             1 FALSE          2
4 region                0             1 FALSE          4
  top_counts
1 mal: 676, fem: 662
2 0: 574, 1: 324, 2: 240, 3: 157
3 no: 1064, yes: 274
4 sou: 364, nor: 325, sou: 325, nor: 324

── Variable type: numeric ─────────────────────────────────────────
  skim_variable n_missing complete_rate    mean       sd     p0    p25    p50
1 age                   0             1    39.2    14.0    18     27     39
2 bmi                   0             1    30.7     6.10   16.0   26.3   30.4
3 charges               0             1 13270.  12110.   1122.  4740.  9382.
      p75    p100 hist
1    51      64   ▇▅▅▅▆▇
2    34.7    53.1 ▁▅▇▅▁▁
3 16640.  63770.  ▇▂▁▁▁▁
> |
```

Figure 3-1

Findings:

- There are four numerical variables:
  - Continuous: Age, BMI, Charges
  - Discrete: Dependents
- There are three categorical variables:
  - Sex
  - Smoker
  - Region
- No missing values

3.2 Distributions of variables

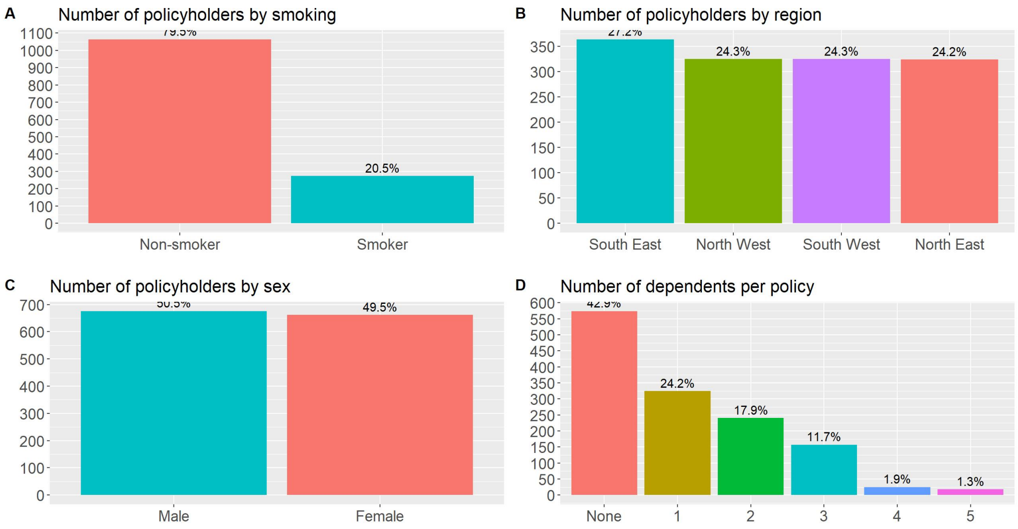


Figure 3-2

Summary

- Smoking status: There are many more non-smokers (80%) than smokers (20%).
- Region of residence: Policyholders are evenly distributed across regions with South East being the most populous one (27%) with the rest of regions containing around 24% of policyholders each.
- Sex: There are slightly more men (51%) than there are women (49%) in the sample.
- Dependents: Most policyholders (43%) do not have dependents covered in their policy. For those who do have dependents covered in their policy, most have one dependent (24%). Maximum number of dependents covered is five (1%).

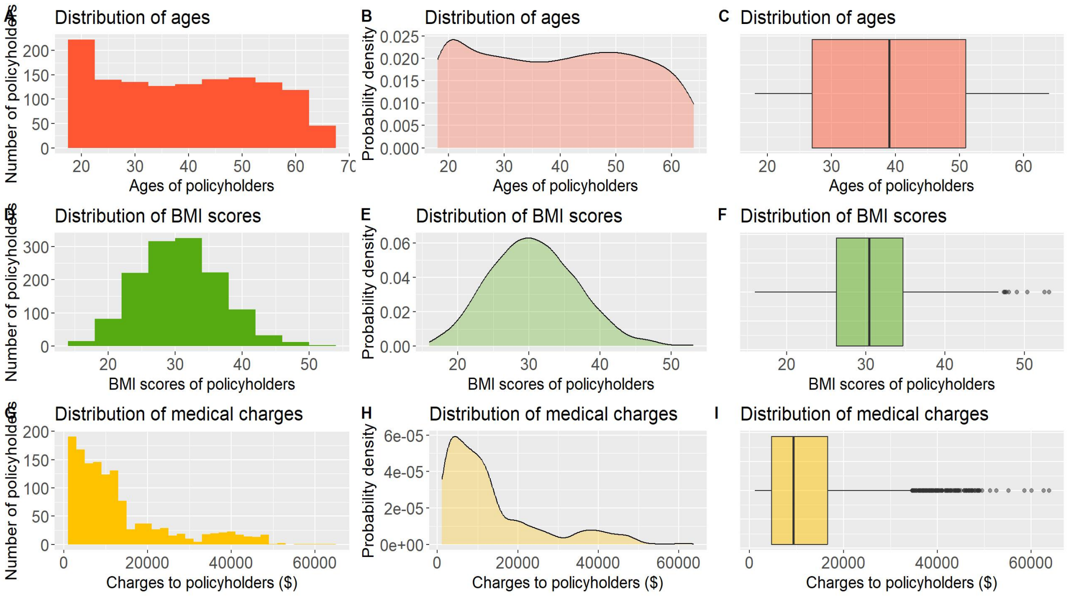


Figure 3-3

Summary

- Age: Youngest policyholder is 18 and the eldest is 64. All ages in the range are represented fairly equally apart from the youngest and eldest policyholders. 18-23-year-olds are the most populous group (among all 5-year segments) and 60-64-year-olds are the least represented 5-year age group. There are no outliers.
- BMI: BMI is normally distributed with the smallest and the largest values being the least common and median and mean being almost identical. There are a few outliers on the larger side. Minimum recorded BMI score is 16 and maximum is 53.1.
- Charges: Charges are heavily right-skewed with many outliers on the larger side. This means most charges are fairly low with a few particularly high charges. Smallest charge is $1,122 and largest charge is $63,770.

3.3. Charges vs all other factors

Since the expenses are not normally distributed, it is not helpful to compare the means. Medians are compared below in box plot (A-D) and violin plot graphs (E-H).

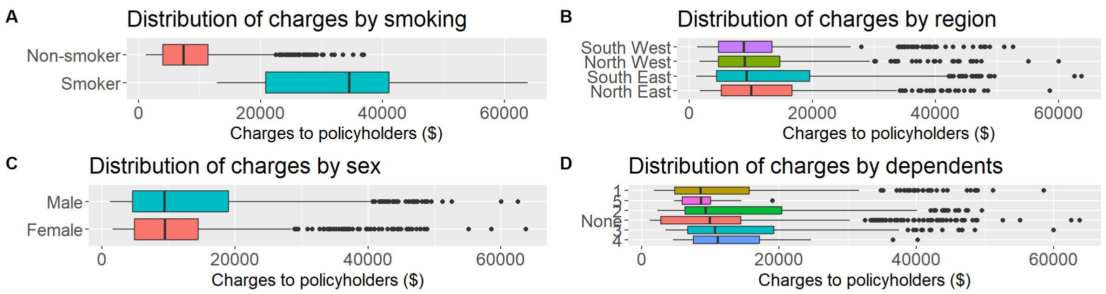


Figure 3-4

Summary

- **Smoking**: There is a big difference in medians between smokers ($34,456) and non-smokers ($7,345). Non-smokers show many outliers on the larger side, while the vast majority of charges are on the smaller side. Smokers show bimodal distribution and no outliers.
    - Larger charges for smokers are to be expected as smoking is a known serious health risk.
- **Region of residence**: There are slight differences in medians between all groups. All groups have outliers. The spread of values is fairly similar for all groups apart from South East which has a larger interquartile range (IQR).
- **Sex**: Males have a marginally larger median ($9,413) than females ($9,370), a difference of just $43. Both groups show outliers on the larger side. The spread of values is fairly similar.
- **Dependents**: There are some differences in medians between the groups, but they are not drastic.

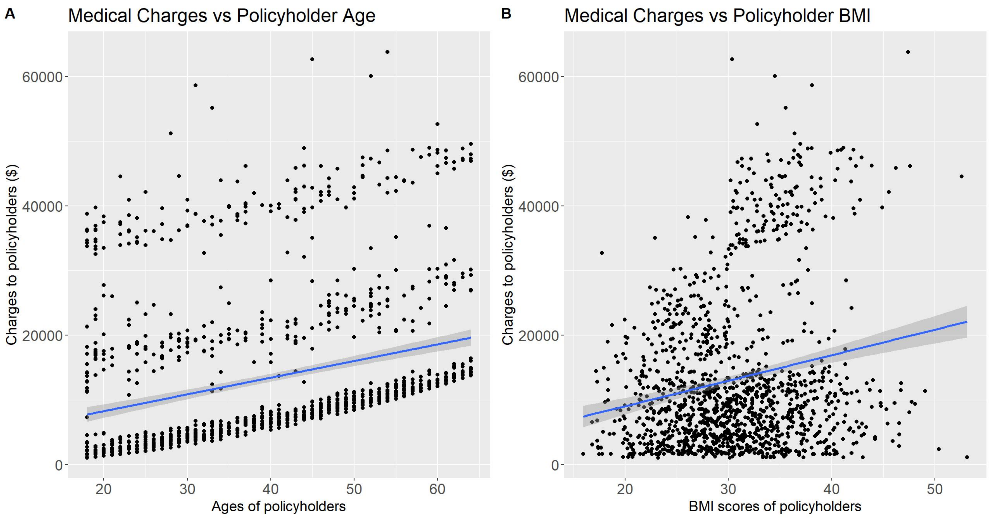
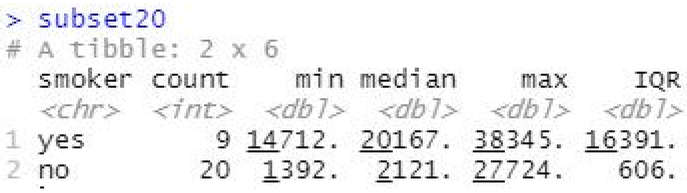


Figure 3-5

Summary

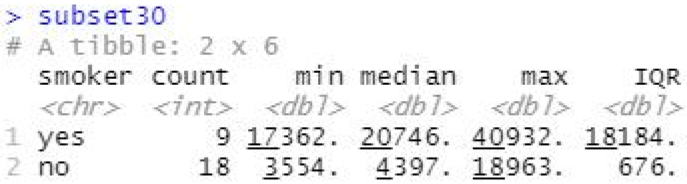
- **Medical Charges vs Policyholder Age**: There are interesting patterns here, showing three groups. The lower band shows a very strong relationship between medical charges and age, with two other bands showing a less strong relationship. The general trend is a positive correlation, meaning as age increases, so do medical expenses.
- **Medical charges vs Policyholder BMI**: There is a positive relationship between BMI and medical expenses, meaning people with higher BMI scores have higher medical bills. The relationship is not, however, very strong. There are possibly two groups to the scatter plot, judging by the spread of points above and below the regression line.

We can find something interesting inside plot A. It shows there are three groups, and we can infer that smoke caused a much higher different medical costs of person with the same age, we can check the medical costs of person by age of 20,30,40,50 and 60.



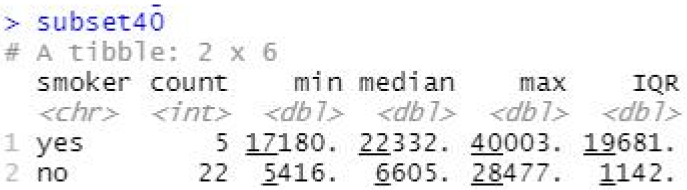
```
> subset20
# A tibble: 2 x 6
  smoker count    min median    max    IQR
  <chr>  <int>  <dbl>  <dbl>  <dbl>  <dbl>
1 yes        9 14712. 20167. 38345. 16391.
2 no        20  1392.  2121. 27724.   606.
```

Figure 3-6



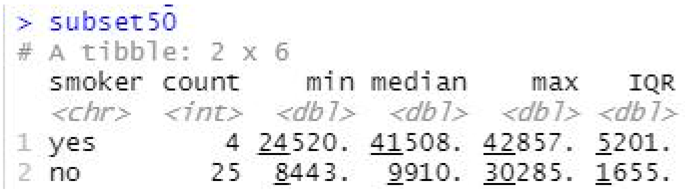
```
> subset30
# A tibble: 2 x 6
  smoker count    min median    max    IQR
  <chr>  <int>  <dbl>  <dbl>  <dbl>  <dbl>
1 yes        9 17362. 20746. 40932. 18184.
2 no        18  3554.  4397. 18963.   676.
```

Figure 3-7



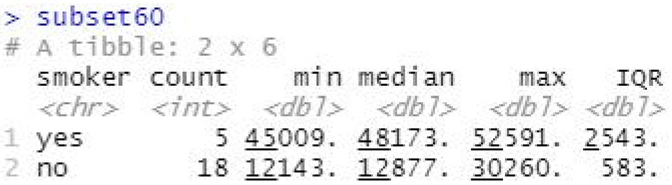
```
> subset40
# A tibble: 2 x 6
  smoker count    min median    max    IQR
  <chr>  <int>  <dbl>  <dbl>  <dbl>  <dbl>
1 yes        5 17180. 22332. 40003. 19681.
2 no        22  5416.  6605. 28477.  1142.
```

Figure 3-8



```
> subset50
# A tibble: 2 x 6
  smoker count    min median    max   IQR
  <chr>  <int>  <dbl>  <dbl>  <dbl> <dbl>
1 yes        4 24520. 41508. 42857. 5201.
2 no        25  8443.  9910. 30285. 1655.
```

Figure 3-9



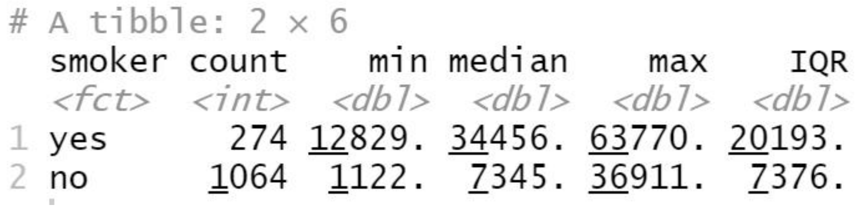
```
> subset60
# A tibble: 2 x 6
  smoker count    min median    max   IQR
  <chr>  <int>  <dbl>  <dbl>  <dbl> <dbl>
1 yes        5 45009. 48173. 52591. 2543.
2 no        18 12143. 12877. 30260.  583.
```

Figure 3-10

- As some minimum medical cost of smoker is catching up maximum medical cost of non-smoker in some age groups, and the huge difference in median medical cost between non-smokers and smokers in the same age explains the reason of three bands showing in plot A.

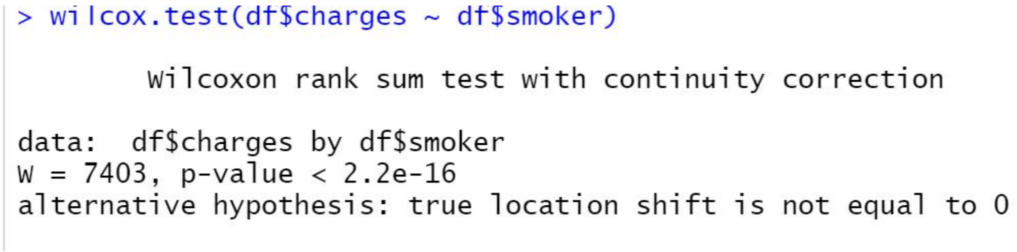
4. Hypothesis Testing

4.1. Smoking



```
# A tibble: 2 × 6
  smoker count    min median    max    IQR
  <fct>  <int>  <dbl>  <dbl>  <dbl>  <dbl>
1 yes      274 12829. 34456. 63770. 20193.
2 no      1064  1122.  7345. 36911.  7376.
```

Figure 4-1



```
> wilcox.test(df$charges ~ df$smoker)

        Wilcoxon rank sum test with continuity correction

data:  df$charges by df$smoker
W = 7403, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Figure 4-2

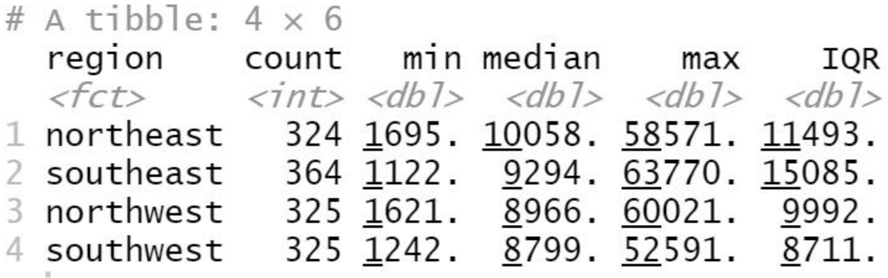
Exploratory data analysis has indicated that smoking has an effect on charges. Smokers' (N: 274) median medical expenses are $34,456 (range: \$12,829-$63,770) and medical expenses for non-smokers (N: 1064) are $7,345 (range: $1,122-\$36,911). The is also considerable difference in the distribution of the observations between groups. An independent 2-group Mann-Whitney U Test was performed to determine if this difference is statistically significant. Assumptions of the test are as follows - dependent variable is continuous, two independent categorical variables are present, there is no relationship between the observations in each group of the independent variables or between the groups themselves, and the shape of distributions of the independent variables must be known. Since the distributions of charges grouped by sex are different, the test is used to determine whether there are differences in the distributions of the two groups. All of these assumptions are met.

- H0: There is no difference in the distribution scores.
- HA: There is a difference in the distribution scores.

The test indicated that there is a significant difference between the groups, W = 7403, *p* < 0.001. The null hypothesis is rejected.

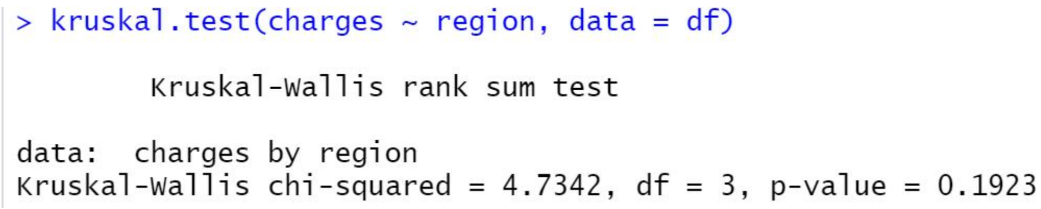
4.2. Regions

There was also some difference in medical charges between regions.



```
# A tibble: 4 × 6
  region    count   min median   max   IQR
  <fct>     <int> <dbl>  <dbl> <dbl> <dbl>
1 northeast   324 1695. 10058. 58571. 11493.
2 southeast   364 1122.  9294. 63770. 15085.
3 northwest   325 1621.  8966. 60021.  9992.
4 southwest   325 1242.  8799. 52591.  8711.
```

Figure 4-3



```
> kruskal.test(charges ~ region, data = df)

	Kruskal-Wallis rank sum test

data:  charges by region
Kruskal-Wallis chi-squared = 4.7342, df = 3, p-value = 0.1923
```

Figure 4-4

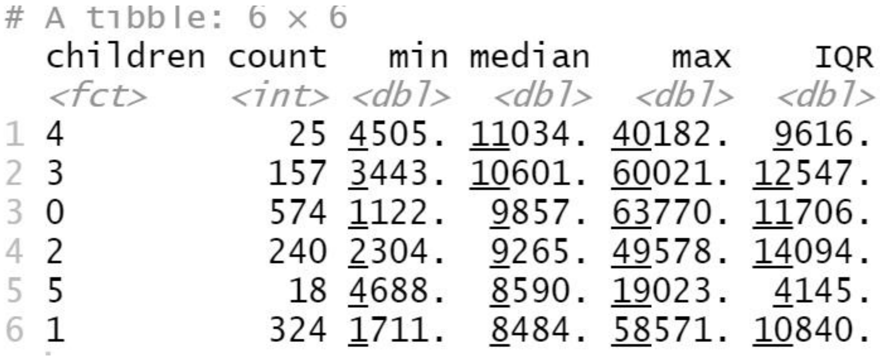
North East sees the largest charges (*Mdn* = $10,058), followed by South East (*Mdn* = \$9,294), then by North West (*Mdn* = $8,966), with South West (*Mdn* = \$8,799) sees the smallest charges. Kruskal-Wallis test was performed to determine if these differences are significant. Assumptions of the test are as follows - dependent variable is continuous, two independent categorical variables are present, there is no relationship between the observations in each group of the independent variables or between the groups themselves. All assumptions are met.

- H0: There is no difference between the medians.
- HA: There is a difference between the medians.

The test showed that the difference between the median medical charges in different regions is not significant, *H*(3) = 4.73, *p* = 0.19. A significant level of 0.19

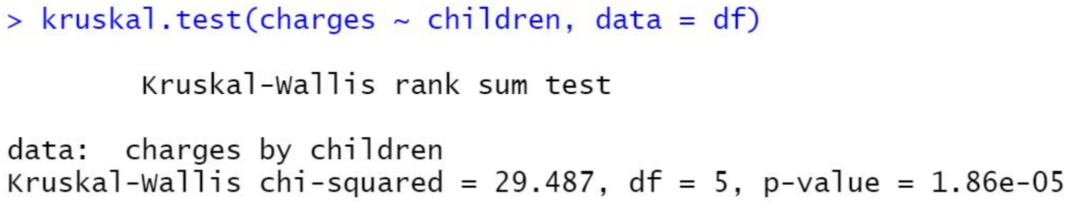
indicates a 19% risk of concluding that a difference exists when there is no actual difference. The null hypothesis is accepted.

4.3. Children



```
# A tibble: 6 × 6
  children count   min median    max    IQR
  <fct>    <int> <dbl>  <dbl>  <dbl>  <dbl>
1 4           25 4505. 11034. 40182.  9616.
2 3          157 3443. 10601. 60021. 12547.
3 0          574 1122.  9857. 63770. 11706.
4 2          240 2304.  9265. 49578. 14094.
5 5           18 4688.  8590. 19023.  4145.
6 1          324 1711.  8484. 58571. 10840.
```

Figure 4-5



```
> kruskal.test(charges ~ children, data = df)

	Kruskal-Wallis rank sum test

data:  charges by children
Kruskal-Wallis chi-squared = 29.487, df = 5, p-value = 1.86e-05
```

Figure 4-6

A Kruskal-Wallis test (assumptions met) also showed that the number of dependents covered by the insurance policy significantly affects medical costs billed on that policy by the insurance company, $H(5) = 29.49$, $p < 0.001$. Medical expenses for the rest of the groups can be seen in the table below.

5. Multiple Linear Regression

Multiple linear regression (MLR) models allow for effective summarisation of multivariate datasets. It is an extension of the single linear regression in which instead of one independent variable, multiple independent variables are used to predict the value of the response variable.

Response variable (charges) is to be transformed to reduce skewness and meet the assumption of normality for the MLR model.

The dataset is to be split into a training dataset (80% of all data) and a testing dataset (20% of all data).

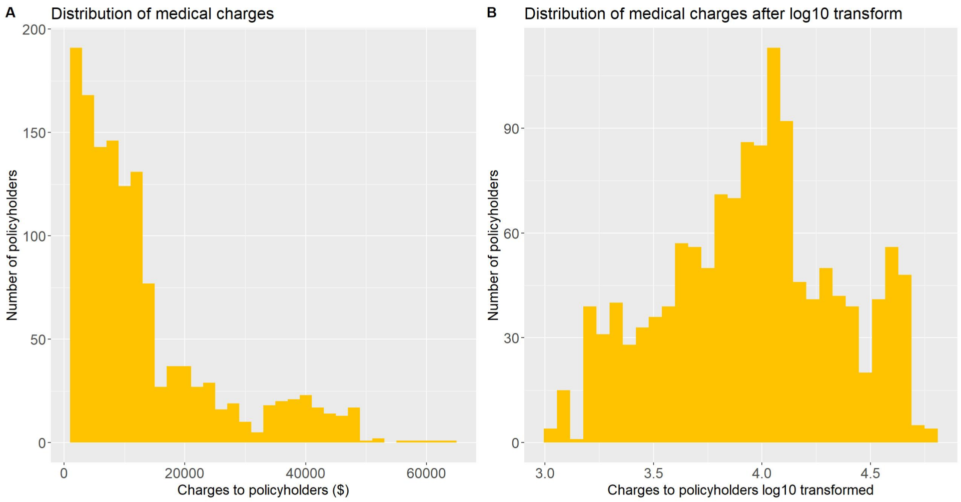
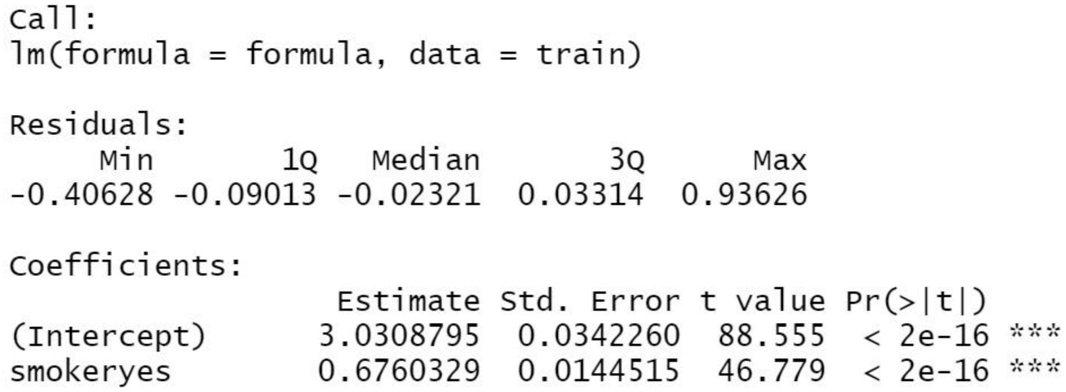


Figure 5-1

The hypotheses for this model are such:

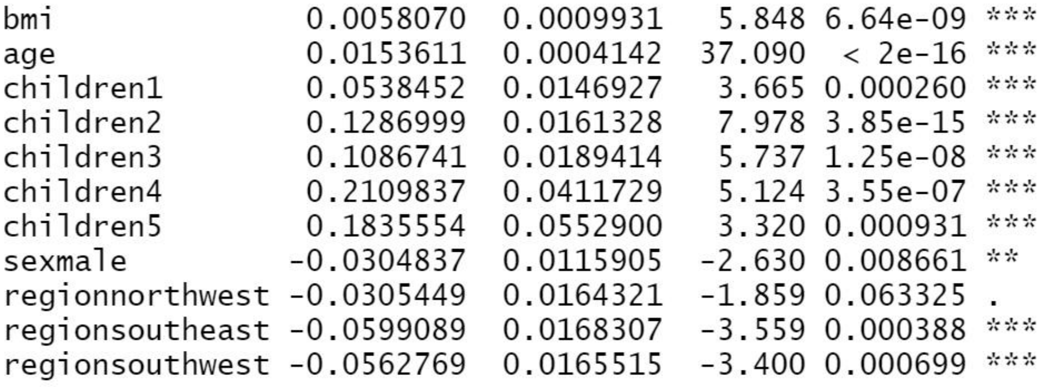
- **Null hypothesis**: there will be no significant prediction of medical expenses by the policyholder's smoking status, BMI score, age, region of residence, sex, and number of dependents covered by the policy.
- **Alternative hypothesis**: there will be significant prediction based on the above mentioned factors.

5.1. Split the dataset and train the model



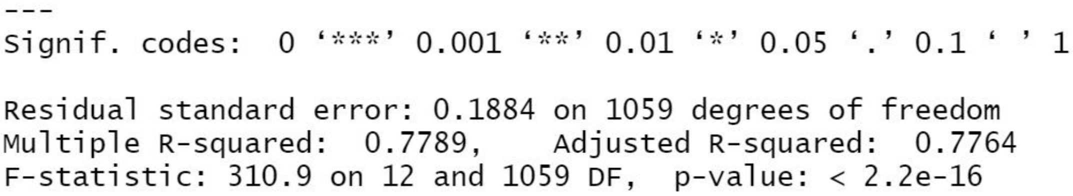
```
Call:
lm(formula = formula, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.40628 -0.09013 -0.02321  0.03314  0.93626 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.0308795  0.0342260  88.555  < 2e-16 ***
smokeryes       0.6760329  0.0144515  46.779  < 2e-16 ***
```

Figure 5-2



```
bmi             0.0058070  0.0009931   5.848 6.64e-09 ***
age             0.0153611  0.0004142  37.090  < 2e-16 ***
children1       0.0538452  0.0146927   3.665 0.000260 ***
children2       0.1286999  0.0161328   7.978 3.85e-15 ***
children3       0.1086741  0.0189414   5.737 1.25e-08 ***
children4       0.2109837  0.0411729   5.124 3.55e-07 ***
children5       0.1835554  0.0552900   3.320 0.000931 ***
sexmale        -0.0304837  0.0115905  -2.630 0.008661 ** 
regionnorthwest -0.0305449  0.0164321  -1.859 0.063325 .  
regionsoutheast -0.0599089  0.0168307  -3.559 0.000388 ***
regionsouthwest -0.0562769  0.0165515  -3.400 0.000699 ***
```

Figure 5-3



```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1884 on 1059 degrees of freedom
Multiple R-squared:  0.7789,	Adjusted R-squared:  0.7764 
F-statistic: 310.9 on 12 and 1059 DF,  p-value: < 2.2e-16
```

Figure 5-4

A significant regression equation was found (F(12,1057) = 303.9, p < 0.001), with an adjusted R-squared of 0.7764. In other words, the model explains 77.6% of total variance in the sample. Null hypothesis is rejected.

6. Discussion and Conclusions

Smoking having the strongest effect on medical expenses is quite expected.

Increases in the BMI score lead to rather small expense increases, however, it is worth pointing out that normal BMI scores are not indicative of ill health. Only people in the underweight (BMI < 18.5), overweight (BMI 25.0 to 29.9), and obese (BMI ≥≥ 30) ranges would be expected to have poorer health outcomes.

Same should be said of the effect of aging - 22-year-olds would be expected to enjoy the same level of health as 18-year-olds despite being 4 years older. However, middle aged and elderly people will most likely see a rapid decline in health year by year.

Medical expenses increasing with increased number of dependents is to be expected. However, having three dependents covered by insurance seems to be cheaper than having two dependents, and five dependents sees a lesser increase in charges than four. This may be explained by the uneven number of observations in each group. For example, no dependents group has 574 observations when five dependents group only has 18.

It is also interesting to note that even though the median difference of medical charges between men and women is only $43, the relationship between sex and medical charges was significant in the multiple linear regression model.

7. References

1. Fulton, B. D. (2017). Health care market concentration trends in the United States: evidence and policy responses. Health Affairs, 36(9), 1530-1538.

2. Ho, K. (2009). Insurer-provider networks in the medical care market. American Economic Review, 99(1), 393-430.

3. Frees, E. W. (2009). Regression modeling with actuarial and financial applications. Cambridge University Press.

4. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300.