Linear regression model for predicting medical expenses based on insurance data

I. Introduction

Medical expenses is one of the important expenditures in people lives. Factors that are often regarded as contributing to higher medical cost are age, gender, obesity(high body-mass index), smoking... In this study, we will examine which attributes affect a person's medical cost the most.

Firstly, we find the correlation of medial expense with each attribute. Afterwards, we use regression analysis to build models that predict personal medical cost from subsets of attributes. The models are compared using ANOVA, and the best-fit model will be indicated.

The source codes, plots, and images used in this report are stored in the repository:

https://github.com/HoangNV2001/AnalysisLinearRegressionMedicalExpenses

II. Dataset & exploratory analysis

A. Data

We use the medical cost personal dataset from Kaggle[1] which consists of US peoples information and their medical costs billed by health insurance. There are a total of 1338 samples. The table below lists the attributes and their descriptions.

Column	**Description**
Age	Age of primary beneficiary
Sex	Insurance contractor gender, female, male
BMI	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
Children	Number of children covered by insurance
Smoker	If the person smokes
Region	The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
Charges	Individual medical costs billed by health insurance

The variable age, BMI, children are numerical variables, the variables sex, smoker and region are categorical variables. In the summary statistics below, we can see that there are no missing values in the dataset.

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

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

Fig. 1. Summary of Data

B. Data exploration

This section is about exploring the data by scatter plots, box plots, and correlations to find out relations between attributes. The findings shall be used in the later phase for building models.

1. Visualising data:

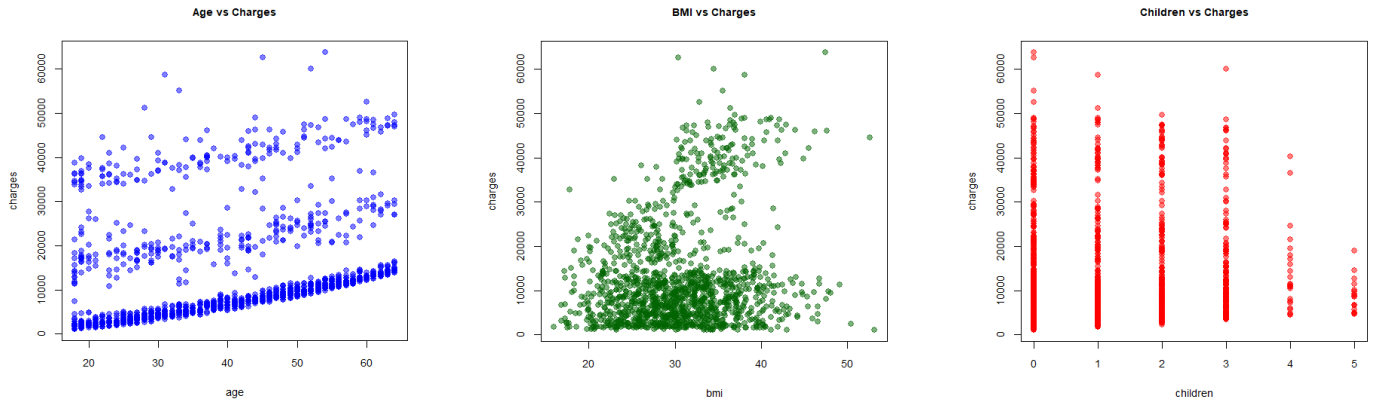


Fig. 2. Charge vs Age, BMI, Children

Regarding the scatter plots, we can notice that age and charges share an increasing trend.

While in BMI vs charge, the trend is less visible, but as bmi increase, there is a higher number of occurence of samples with more expensive charge. The correlation between the number of children and charge is not remarkable.

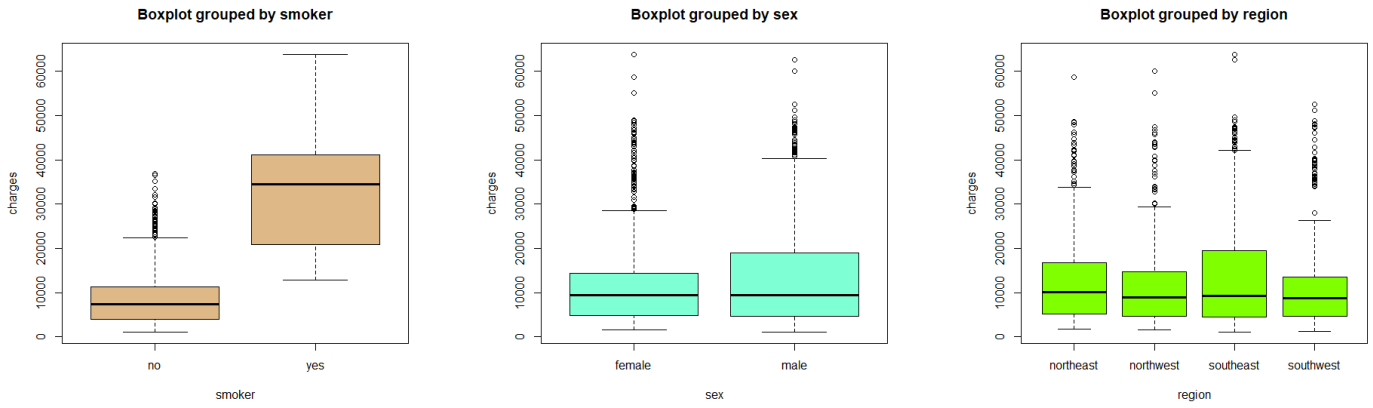


Fig. 3. Charge vs Smoker, Sex, Region

As can be clearly seen from the box diagrams, smokers have much higher medical expenses than non-smokers.This should be regarded as an important factor. On the contrary, gender and region have no significant influences on the charges

2. Correlation:

The heatmap below represents the correlation matrix between the attributes.

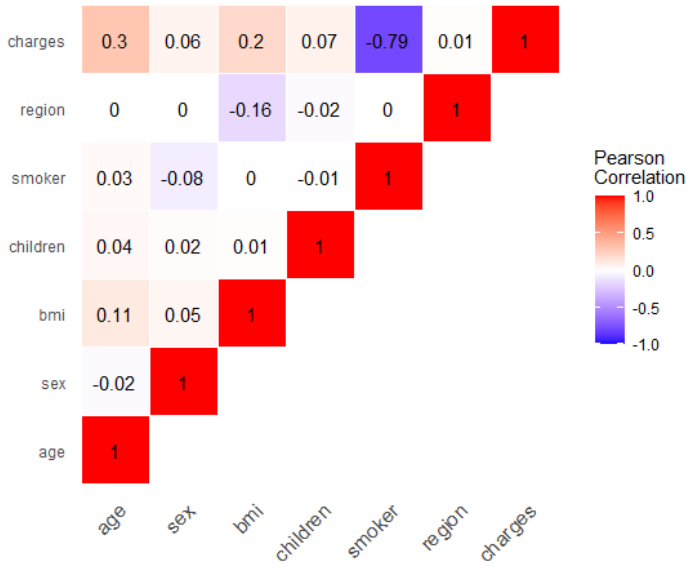


Fig. 4. Correlation heatmap of attributes

As can be seen from the heatmap, smoker have high degree of correlation to charges. Also, age have moderate correlation, followed by BMI. Other attributes have low to almost no correlation to Charges.

III. Model & hypotheses

In this section, we will build linear regression models to predict the personal medical charges. Also, the models will be compared and evaluate so that we can choose the best-fit one.

As inferred from previous section, age, BMI, smoking have the highest correlation to charges among all attributes. Therefore, the initial multiple regression model would be:

charges ~ age + bmi + smoker

From the above formula, we build a regression model using R's lm() function. The snapshot below is the model summary, and we will explain and interpret some of the coefficients.

```
Call:
lm(formula = charges ~ age + bmi + smoker, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-12415.4  -2970.9   -980.5   1480.0  28971.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11676.83     937.57  -12.45   <2e-16 ***
age            259.55      11.93   21.75   <2e-16 ***
bmi            322.62      27.49   11.74   <2e-16 ***
smokeryes    23823.68     412.87   57.70   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6092 on 1334 degrees of freedom
Multiple R-squared:  0.7475,    Adjusted R-squared:  0.7469
F-statistic:  1316 on 3 and 1334 DF,  p-value: < 2.2e-16
```

Fig. 5. Summary of Model (smoking+age+bmi)

- **Call**: This is an R feature that shows what function and parameters were used to create the model.
- **Residuals**: Difference between what the model predicted and the actual value of y.

$$\boldsymbol{r = x - x_0} \quad ; \boldsymbol{with}: \begin{cases} \boldsymbol{r = residual} \\ \boldsymbol{x = measure\ variable} \\ \boldsymbol{x_0 = approximate\ variable} \end{cases}$$

- **Coefficients:**
  - **Estimate:** These are the weights that minimize the sum of the square of the errors. The detail about calculating steps shall be omitted in this report.

    The coefficient for age is 259.55, BMI is 322.62 and smoker is 238223.68. The intercept is

    -11676.83. Thus the model for predicting charges using smoking, age, BMI is:

    charges = -11676.83 + (259.55 * age) + (322.62 * BMI) + (238223.68 * smoker) + error.

  - **Std. Error:** Standard error of our estimate , which allows us to construct marginal confidence intervals for the estimate of that particular feature parameter $\hat{\beta}_{\text{i.}}$

    e.g. If s.e($\hat{\beta}_{\text{i}}$) is the estimated coefficient for feature i, then a 95% confidence interval of $\hat{\beta}$i is given by : $\hat{\beta}$i ± 1.96 × s.e($\hat{\beta}$i)

  - **t-value:** The t-statistic is

$$\frac{\hat{\beta}\text{i}}{s.e(\hat{\beta}\text{i})}$$

    which tells us about how far our estimated parameter is from a hypothesized 0 value, scaled by the standard deviation of the estimate.

  - **Pr(>|t|) :**

    This is the p-value for the individual coefficient. Under the t distribution with n – p – 1 degrees of freedom, this tells us the probability of observing a value at least as extreme as our $\hat{\beta}$i . If this probability is sufficiently low, we can reject the null hypothesis that this coefficient is 0.

    From the model summary, we are can support the claim that age, bmi and smoking have significant influence on charges (reject the null hypothesises that their coefficients are 0)

- **Multiple R-squared**: Also called the coefficient of determination, this is often used as measurement of how well the model fits to the data.

  Formula of R-squared:

$$R^2 = 1 - \frac{RSS}{TSS} \quad ; with \begin{cases} RSS : residual\ sum\ of\ squares \\ TSS : total\ sum\ of\ squares \end{cases}$$

  In the summary of this model, Multiple R-squared is 0.7475. This value represents the square R between age and charges and indicates that 74.75% of the variation in the outcome variable charges can be explained by the predictors.

- **Adjusted R-square**: Adjusted R-Squared normalizes Multiple R-Squared by taking into account how many samples we have and how many variables we're using. *"The use of an adjusted R^2 is an attempt to account for the phenomenon of the R^2 automatically and spuriously increasing when extra explanatory variables are added to the model."*[2] The formula of adjusted R^2 is:

[2] https://en.wikipedia.org/wiki/Coefficient_of_determination#In_a_multiple_linear_model

$$\bar{R}^2 = 1 - \frac{RSS/df_e}{TSS/df_t}$$

$$with \begin{cases} df_t: degrees\ of\ freedom\ n - 1\ of\ the\ estimate\ of \\ the\ population\ variance\ of\ the\ dependent\ variable \\ df_e: the\ degrees\ of\ freedom\ n - p - 1\ of\ the\ estimate \\ of\ the\ underlying\ population\ error\ variance. \end{cases}$$

For this model, the adjusted R-square (0.7494) is almost equal to the multiple R-square. This is due to the fact that this is a relatively simple model with only three explanatory variables.

- **F-statistic**: The F value in regression is the result of a test where the null hypothesis is that all of the regression coefficients are equal to zero. In other words, the model has no predictive capability.

  Basically, the f-test compares the model with zero predictor variables (the intercept only model), and decides whether the added coefficients improved the model.

  From the Fstatistic result (p-value << 0.001), we can conclude that the model is good and that the predictor variables are all significant.

Next, we try to improve this model by including the predictor variables sex, children and region.

```
Call:
lm(formula = charges ~ age + bmi + children + region + sex +
    smoker, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5       137.8   3.451 0.000577 ***
regionnorthwest   -353.0       476.3  -0.741 0.458769
regionsoutheast  -1035.0       478.7  -2.162 0.030782 *
regionsouthwest   -960.0       477.9  -2.009 0.044765 *
sexmale           -131.3       332.9  -0.394 0.693348
smokeryes        23848.5       413.1  57.723  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Fig. 6. Summary of Model (smoker + age + bmi + sex + children + region)

In this updated model the multiple R square is 0.7509, meaning that the new model explains 75.09 of the variation in charges.

Looking at the coefficients, we find that sex attribute is not significant (p-value is 0.69). Therefore, we remove the attribute sex from the model and reevaluate it.

```
Call:
lm(formula = charges ~ age + bmi + children + region + smoker,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-11367.2  -2835.4   -979.7   1361.9  29935.5

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11990.27     978.76 -12.250  < 2e-16 ***
age                256.97      11.89  21.610  < 2e-16 ***
bmi                338.66      28.56  11.858  < 2e-16 ***
children           474.57     137.74   3.445 0.000588 ***
regionnorthwest   -352.18     476.12  -0.740 0.459618
regionsoutheast  -1034.36     478.54  -2.162 0.030834 *
regionsouthwest   -959.37     477.78  -2.008 0.044846 *
smokeryes        23836.30     411.86  57.875  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6060 on 1330 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7496
F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

Fig. 7. Summery of Model (smoker + age + bmi + children + region)

After removing attribute sex from the model, we can observe that the model still explains 75.09 per cent of the variation in charges. Next, we will compare the initial model and the current one using ANOVA.

- ANOVA:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1334	49513219514	*NA*	*NA*	*NA*	*NA*
2	1330	48845249273	4	667970241	4.547015	0.001191318

Fig. 8. Summary of ANOVA (smoker + age + bmi) vs

(smoker + age + bmi + children + region)

The p-value is significant (0.001 < 0.01), so we conclude that our third model provides a better fit than our initial model.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1329	48839532844	*NA*	*NA*	*NA*	*NA*
2	1330	48845249273	-1	-5716429	0.155553	0.6933475

Fig. 9. Summary of ANOVA (smoker + age + bmi + children + region + sex) vs

(smoker + age + bmi + children + region)

Comparing the second and the third model, the p-value is non-significant (0.69 >> 0.01). Therefore, we conclude that there is no significant difference between two models’ performances.

Since the third model has one less variable than the second model, we decide that the former is the most prominent.

- Standardized residual:

One crucial part is outliers, these points are especially important because they can have a strong influence on the least squares line. We take regard to outliers by examine standardized residual. The standardized residual is the residual (the differences between the observed and predicted values) divided by its standard deviation.

$$Standardized\ Residual\ i = \frac{Residual\ i}{Standard\ Deviation\ of\ Residual\ i}$$

Outliers would have high standardized residual value. Upon calculation, the proprotion of samples with an absolute value of standardized residual greater than 2 is 5.01 percent, and greater than 3 is 2.1 percent. This rate is of outliers is acceptable.

```
> standard_res <- rstandard(r3)
> sum(standard_res < -2 | standard_res >2)/ length(standard_res)
[1] 0.05007474
> sum(standard_res < -3 | standard_res >3)/ length(standard_res)
[1] 0.02092676
```

Fig. 10. Calculating proportion of outliers

- Variance Inflation Factor

A variance inflation factor (VIF) detects multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect the regression results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

VIFs are calculated by taking a predictor variable, and regressing it against every other predictor in the model. This gives us the R-squared values (formula above), which can then be plugged into the VIF formula. "i" is the predictor variable we are examine (e.g. x1 or x2)

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

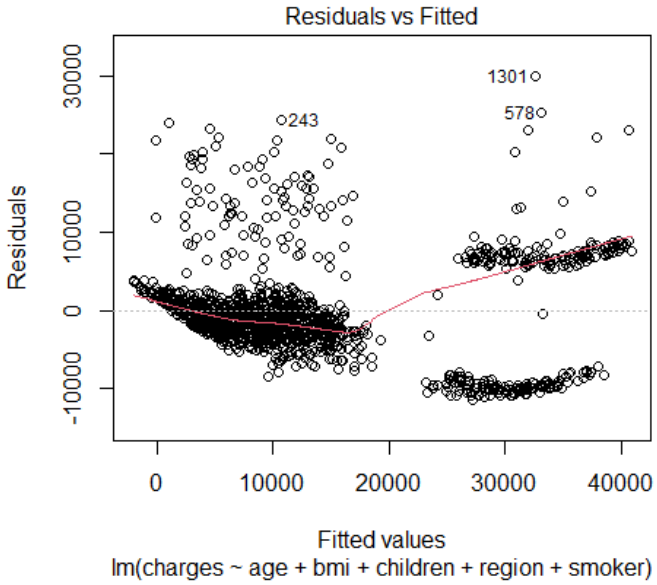
The table below shows the VIF test result. A VIF higher than 5 indicates high correlation. Since none of the VIF values is higher than 5, we can claim that there is no multicolinearity.

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

Fig. 11. Summary of VIF Test

- Residual plots

We will investigate further into residuals through residual plot. No patterns should be present if the model fits well. Also, the plot should look like random dots evenly distributed around the horizontal zero line.



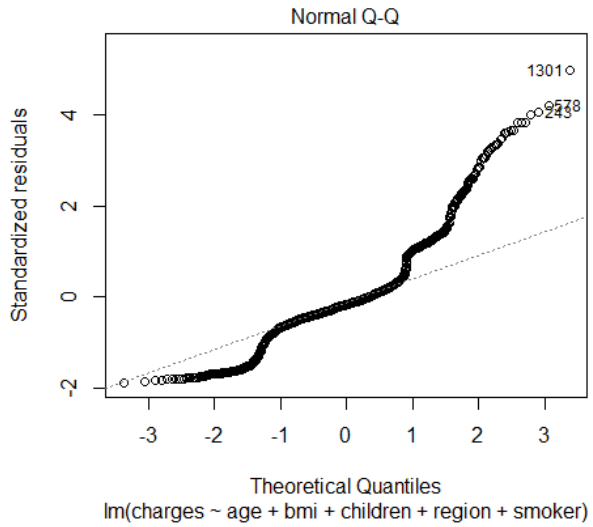
In the chart, that dots seem not to be randomly placed around the horizontal zero line. There are there distinguishable groups. Our conclusion is that there might be other factors that are not taken into account of, which means there is an additional varible (perhaps a categorical one with 3 categories) that is not included in the dataset.

There might also be a non-linear relationship between the charge and the predictor variables. This will be inspected in the following plot.

- QQ Plot

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

The above mentioned non-linear relationship is also reflected by the second plot (Q-Q plot) since the dots deviate from the dotted line.



IV. Results & Conclusion

From the research, we find that age, BMI, number of children, and smoking are the most influential factors to a person's medical charge. Among this smoking seems to affect the most on the medical expenses. On the contrary, region and gender do not contribute much to the change in medical charges.

Our final linear regression model explains 75% of the data, which is a good result. However, there seems to be non-linear relationships between healthcare cost and predictors. Also, there might be hidden patterns that are not explained by the data due to the lack of variables.