Health Insurance Premium: Regression and Resampling Analysis

Data Visualization and Modelling
Master's Degree in Modelling for Science and Engineering
Universitat Autonoma de Bracelona

Anna Bicelli

February 2025

Abstract

This report analyzes factors influencing medical costs and health insurance premiums in the United States based on a dataset containing demographic data and health-related information from a sample of individuals. First, a linear regression model was built to predict the cost of insurance based on age, body mass index (BMI), number of child, whether an individual is a smoker or not, and region of origin. Then two resampling techniques were applied in order to asses the stability and reliability of the model's coefficient estimates: Non-parametric Bootstrap and Jackknife. The results show that...(results confirm the importance of smoking, age, BMI, geographic region, and frequency of exercise on the insurance cost paid by Americans included in the dataset.)

1 Introduction and description of the dataset

The health insurance landscape in the United States is complex and characterized by much controversy, so it is important to correctly predict the price of the insurance premium that each individual, based on his or her habits and characteristics, must pay.
This paper is based on the analysis of the dataset 'Insurance Data for Machine Learning' [1] with the goal to investigate the relationships between various demographic and health-related factors with insurance charges and to develop an accurate predictive model. To obtain rigorous results, a linear regression model was constructed and later, using statistical modeling and two resampling techniques, such as Non-parametric Bootstrap and Jackknife, this report aims to provide a robust assessment of the key drivers of insurance costs.
The dataset contains 10 numerical and categorical variables, with the latter being converted into factors in order to build the regression model. These variables can be categorized into demographic and health-related factors:

- **age**: numerical variable that indicates the age of the insured, expressed in years(, allowing the assessment of aging impact on insurance costs).
- **gender**: categorical variable that indicates the gender of the insured(, useful for identifying possible gender differences in insurance costs).
- **BMI**: numerical variable representing the body mass index, providing a measure of body fat.
- **children**: numerical variable indicating the number of children covered under the insurance plan.
- **smoker**: categorical variable indicating whether an individual is a smoker (*yes* or *no*).
- **region**: categorical variable indicating the insured's region of residence in the United States (*northeast*, *northwest*, *southeast*, *southwest*).
- **medical history**: categorical variable indicating the presence of pre-existing medical conditions in the insured.
- **family_medical_history**: categorical variable indicating the presence of medical conditions in the insured's family.
- **exercise_frequency**: categorical variable indicating how often the insured performed physical activity.
- **occupation**: categorical variable indicating the insured's occupation.
- **coverage_level**: categorical variable representing the level of insurance coverage.
- **charges**: numerical variable that indicates the insured's annual healthcare expenditure, expressed in dollars. This is the dependent variable.

Notably, the dataset does not contain missing values in any variable and therefore does not require specific pre-processing techniques.

1.1 Summary Statistics

Table 1 presents summary statistics for each numerical variable in the dataset, providing an initial overview of it.

2 Problem statement, objectives and hypotheses

Insurance companies must determine fair and competitive premiums for their customers. The insurance cost (charges) can be influenced by various demographic and behavioral factors, such as age, BMI (body mass index), the number of dependent children and smoking habits. This study aims to identify the

Statistic	Age	BMI	Children	Charges
Min	18.0	18.00	0.0	3445
1st Qu.	29.0	26.02	1.0	13600
Median	41.0	34.00	2.0	16622
Mean	41.5	34.00	2.5	16735
3rd Qu.	53.0	41.99	4.0	19781
Max	65.0	50.00	5.0	32562

Table 1: Summary Statistics of the Insurance Dataset.

factors that most influence the insurance cost and build a reliable predictive model.

2.1 Objectives

This report examines the following objectives:

1. Explore the relationships between numerical variables and insurance premium costs through visualizations and correlation analysis.
2. Identify the most influential variables affecting insurance premium costs using linear regression.
3. Analyze the impact of smoking on insurance costs by comparing the distribution of premiums between smokers and non-smokers.
4. Evaluate the stability of the estimated coefficients by repeatedly drawing random samples with replacement using the Non-Parametric Bootstrap.
5. Assess each data point's influence on model parameters by systematically removing one observation at a time using the Jackknife method.

2.2 Hypothesis

The study tests the following null and alternative hypotheses:

1. **Impact of numerical variables on the cost of the insurance premium**
   - H_0: There is no significant relationship between the independent variables (age, BMI, children, smoking, region) and the insurance cost (charges).
   - H_1: There is a significant relationship between at least one of the independent variables and the insurance cost.
2. **Cost for smokers vs non-smokers**
   - H_0: There is no significant difference in insurance costs between smokers and non-smokers.

- H_1: Insurance costs for smokers are significantly different than for non-smokers.

3. **Coefficient stability**

   - H_0: The coefficient estimates obtained from the regression model are stable and do not show significant changes with resampling.
   - H_1: Coefficient estimates show significant variability with resampling, indicating potential stability issues.

2.3 Prior Information Available

The dataset used in this study takes into account various considerations that could influence insurance costs in the United States. It is known that smokers usually have significantly higher insurance costs than non-smokers, that age and BMI are positively correlated with insurance costs and finally that the region of residence also influences these costs due to different regional health policies.

3 Statistical Analysis with Resampling Techniques

First, categorical variables were converted into factors to allow their inclusion in the Linear Regression model.

The first scatter plot shows all combinations of pairs among the four numerical variables: `age`, `bmi`, `children`, and `charges`. Below is the visualization:

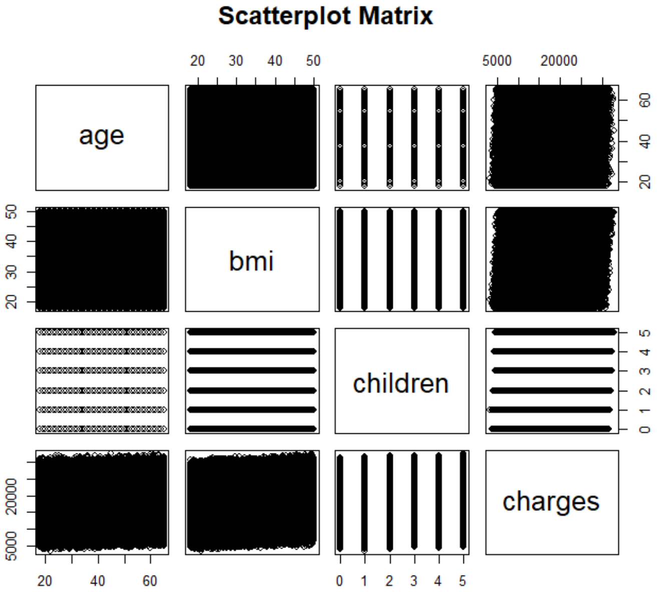


Figure 1: Scatter Plot Matrix of Numerical Variables (`age`, `bmi`, `children`, and `charges`)

It can be observed that there is a slight positive trend between `age` and `charges`, where higher costs tend to be associated with older individuals. However, there is significant dispersion in values, with some younger individuals incurring high costs, possibly due to other factors.

In the relationship between `bmi` and `charges`, a slight tendency for higher `bmi` to be associated with increased insurance costs is observed. Similarly, there is a mild tendency for costs to rise with the number of children, but this variable has a smaller impact compared to `age` and `bmi`.

No significant correlations were found between `age` and `bmi`, nor between `children` and the variables `bmi` and `age`. It can be seen also in this heatmap:

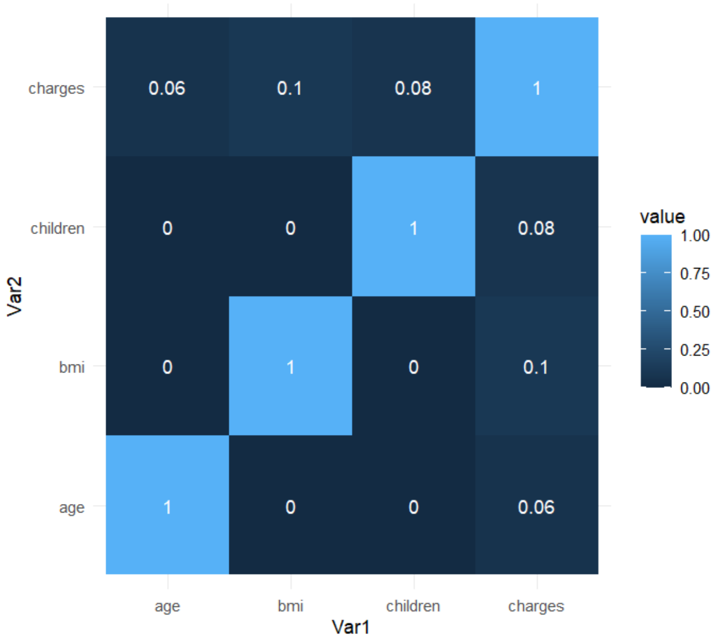


Figure 2: Heatmap correlation between Numerical Variables (`age`, `bmi`, `children`, and `charges`)

To study the relationship between insurance costs and other variables in the dataset, this study used multiple linear regression combined with two resampling techniques for greater robustness in parameter estimation and generation of confidence intervals.

3.1 Multiple Linear Regression

The Multiple Linear Regression was applied twice: first, to model the relationship between insurance costs (dependent variable) and all other independent variables. The second model analyzed the relationship between health insurance premium costs and the variables considered most impactful on the dependent variable, as seen in the section 2.3.

This regression approach allows us to determine how much each factor (such as age or smoking) contributes to explaining the variability in insurance costs and it reduces the complexity of the model and allows for clearer interpretation. The model can be expressed as follows:

$$charges = \beta_0 + \beta_1 \cdot age + \beta_2 \cdot bmi + \beta_3 \cdot children + \beta_4 \cdot smoker + \beta_5 \cdot region_{northwest}$$
$$+ \beta_6 \cdot region_{southeast} + \beta_7 \cdot region_{southwest} + \varepsilon$$

where:

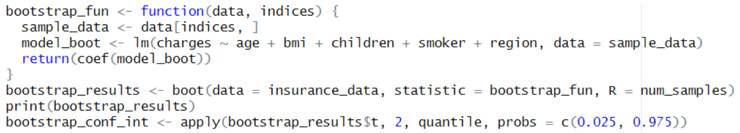
- **charges**: the dependent variable, representing the insurance cost.
- β_0: the intercept of the model.
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$: the coefficients of the most important variables: `age`, `bmi`, `children`, `smoker`, and `region`, respectively.
- ε: the error term, representing the unexplained variance in `charges`.

3.2 Non-Parametric Bootstrap

The Non-Parametric Bootstrap was used in this study to estimate the distribution of the regression coefficients without making assumptions about the distribution of the original data. It studied the variability of the coefficients of the regression model which includes only the most important variables and was divided into the following steps:

1. Resampling: 1000 samples were created by randomly taking data from the original dataset, ensuring a stable estimate of the confidence intervals of the regression coefficients.
2. Model: for each sample a regression model is estimated with the most important variables on the cost of the healthcare insurance premium.This model ensures robust estimates of regression coefficients by reducing the risk of including non-significant variables that would unnecessarily increase the variance of the coefficients and allows for greater interpretability by focusing only on factors that have an important and realistic impact on insurance costs.
3. Estimation of regression coefficients.
4. Estimate of the standard error of the coefficients
5. Built the confidence intervals of the coefficients using the percentile method.

The code used is the following:



```
bootstrap_fun <- function(data, indices) {
  sample_data <- data[indices, ]
  model_boot <- lm(charges ~ age + bmi + children + smoker + region, data = sample_data)
  return(coef(model_boot))
}
bootstrap_results <- boot(data = insurance_data, statistic = bootstrap_fun, R = num_samples)
print(bootstrap_results)
bootstrap_conf_int <- apply(bootstrap_results$t, 2, quantile, probs = c(0.025, 0.975))
```

Figure 3: Non-parametric Bootstrap code

3.3 Jackknife

The Jackkinfe Resampling technique was used to estimate the coefficients of the linear regression model including the independent variables most influential on insurance costs. This approach was also useful for assessing their coefficient stability and identifying any outliers that could significantly influence the estimates. SHOW HOW IN THE RESULTS

The steps performed were:

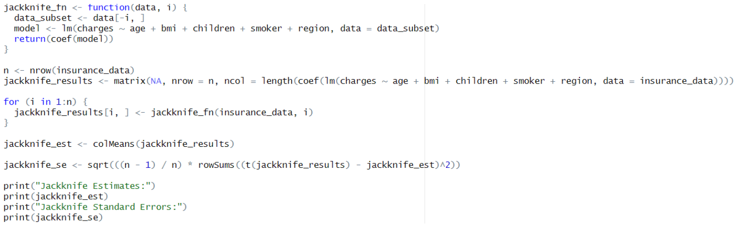
1. Resampling Leave-One-Out: For each observation in the dataset, a sample is created excluding that observation.

2. Fitting the linear regression model on each Jackknife sample.

3. The estimated coefficients are stored in a matrix for subsequent analysis.

4. The standard errors of the coefficients are calculated based on the variability of the Jackknife estimates.

The code used is the following:

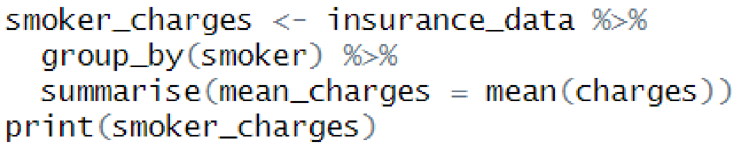


```
jackknife_fn <- function(data, i) {
  data_subset <- data[-i, ]
  model <- lm(charges ~ age + bmi + children + smoker + region, data = data_subset)
  return(coef(model))
}

n <- nrow(insurance_data)
jackknife_results <- matrix(NA, nrow = n, ncol = length(coef(lm(charges ~ age + bmi + children + smoker + region, data = insurance_data))))

for (i in 1:n) {
  jackknife_results[i, ] <- jackknife_fn(insurance_data, i)
}

jackknife_est <- colMeans(jackknife_results)

jackknife_se <- sqrt(((n - 1) / n) * rowSums((t(jackknife_results) - jackknife_est)^2))

print("Jackknife Estimates:")
print(jackknife_est)
print("Jackknife Standard Errors:")
print(jackknife_se)
```

Figure 4: Jackkinfe code

3.4 Charges for smokers and not smokers

è interessante studiare come varia il costo assicurativo in base alla variabile 'smoker'. Per fare ciò è stato utilizzato il seguente codice:



```
smoker_charges <- insurance_data %>%
  group_by(smoker) %>%
  summarise(mean_charges = mean(charges))
print(smoker_charges)
```

Figure 5: Smoke analysis code

4 Results

The statistical analysis of the dataset for Predicting Health Insurance Premiums in the US yielded important results in the relationship between the cost of the insurance premium and demographic and health-related factors.

Table 1: Regression Coefficients, Standard Errors, t-Values, and p-Values

Variable	**Coefficient**	**Standard Error**	**t-Value**	**p-Value**
Intercept (β_0)	11709.56	1.93	5441.1	$< 2e - 16$
Age (β_1)	20.05	0.02	959.2	$< 2e - 16$
BMI (β_2)	50.03	0.03	1598.0	$< 2e - 16$
Children (β_3)	199.12	0.17	1183.6	$< 2e - 16$
Smoker (β_4)	5000.76	0.58	8657.9	$< 2e - 16$
Region: Northwest (β_5)	-712.14	0.82	-857.2	$< 2e - 16$
Region: Southeast (β_6)	-508.39	0.82	-610.8	$< 2e - 16$
Region: Southwest (β_7)	-805.03	0.82	-978.8	$< 2e - 16$

4.1 Multiple Linear Regression Results

After fitting the reduced model to the data, the following results were obtained:

From this table, we can derive the following conclusions:

- $\beta_0 = 11709.56$: This value represents the average baseline insurance cost when all other variables are zero.
- $\beta_1 = 20.05$: For each additional year of age, insurance costs increase by approximately \$20.
- $\beta_2 = 50.03$: Each additional point in Body Mass Index (BMI) is associated with an average increase in insurance costs of \$50.
- $\beta_3 = 199.12$: Having an additional child increases insurance costs by about \$199.
- $\beta_4 = 5000.76$: Being a smoker increases insurance costs by approximately \$5000 compared to non-smokers.
- β_5, β_6, β_7: The negative coefficients for regions indicate that insurance costs are lower compared to the reference region (Northeast), with significant variations among different geographic areas.

All coefficients are highly significant ($p < 0.001$), implying that each variable has a statistically relevant impact on insurance costs.

The reduced model has an R^2 of 0.3464, implying that about 34.64% of the variability in insurance costs is explained by these variables.

The complete model has an R^2 of 0.9957, indicating that it explains nearly all the variability, thanks to the inclusion of numerous significant variables such as medical history, exercise frequency, and coverage level. However, it likely overfits the data. Moreover, the overall F-statistic is 75,7 and its corresponding p-value is ¡ 2.2e-16, so it is possible to conclude that the model is significant and it is possible to refuse the null hypotesis.

The multiple linear regression diagnostic graph has been implemented:

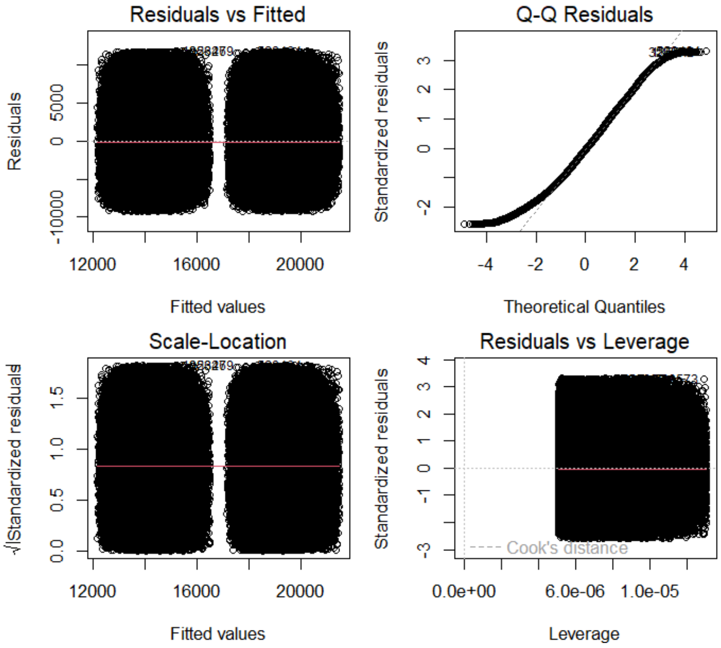


Figure 6: Diagnosis Plot

where:

- Residual vs Fitted plot shows the standardized residuals as a function of the predicted values. The residuals appear grouped in horizontal bands, suggesting the presence of problems of heteroscedasticity or non-constant variability.

- Q-Q Residuals tests the normality of the residuals by comparing them to a theoretical normal distribution. The residuals do not follow the theoretical line perfectly, especially at the extremes, suggesting that the distribution of the residuals is not normal.

- Scale-Location Plot displays the square root of the standardized residuals with respect to the fitted values. The expected flat curve is not visible; there is a certain variation in the dispersion of the residuals along the horizontal axis, indicating the presence of a possible non-constancy of the variance of the residuals.

- Residuals vs Leverage shows the standardized residuals versus the statistical leverage of each point. There are no obvious outliers that significantly affect the model.

4.2 Non-Parametric Bootstrap Results

The bootstrap results include:

- **Original Estimate**: the coefficients of the linear model estimated on the complete dataset.
- **Bias**: the difference between the mean of the coefficients from the resampling technique and the coefficients estimated in the original model.
- **Standard Error**: the standard deviation of the bootstrap coefficient estimates, useful for evaluating the variability of the estimates.

Table 2: Original Estimate, Bias, and Standard Error for Each Coefficient

Coefficient	**Original Estimate**	**Bias**	**Standard Error**
Intercept (β_0)	11709.56	-0.1549	19.29
Age (β_1)	20.05	-0.0032	0.25
BMI (β_2)	50.03	0.0113	0.40
Children (β_3)	199.12	-0.0560	2.20
Smoker (β_4)	5000.76	-0.2092	7.13
Region: Northwest (β_5)	-712.14	-0.1667	9.76
Region: Southeast (β_6)	-508.39	0.3114	10.01
Region: Southwest (β_7)	-805.03	0.0651	10.21

From Table 2, it can be observed that all calculated biases are very close to zero, indicating that the coefficients estimated in the original model are robust and consistent with their distribution obtained through the non-parametric bootstrap. The linear regression model effectively describes the relationship between the independent variables and the 'charges' variable. Furthermore, the standard errors confirm the precision of the coefficient estimates for the multiple linear regression model. For instance, the 'smoker' variable (whether an individual is a smoker or not) has a low margin of error compared to its estimated value of 5000.76 dollars for insurance premium costs. Smoking has the highest impact on insurance costs, increasing medical insurance expenses by approximately 5000 dollars compared to non-smokers. Age and BMI also have a significant impact on insurance costs and are both reliable and consistent predictors. This suggests that insurance policies could benefit from incentives for healthy lifestyles, particularly to reduce the number of smokers, as smoking is the primary cost-increasing factor. Although statistically significant, regional differences have a modest impact compared to smoking.

The use of this resampling technique confirmed the stability of the model and the significance of the coefficients. This reinforces confidence in the model's interpretation and highlights the importance of including smoking, age, and BMI as key drivers of insurance costs.

4.3 Jackknife Results

The standard errors of the regression coefficients estimated using the jackknife resampling technique are presented in the following table:

Table 3: Jacckinfe Estimates and Standard Errors for Regression Coefficients

Variable	**Estimate**	**Standard Error**
(Intercept)	-9541.96	1842.52
Age	271.59	29.19
BMI	353.98	28.97
Children	617.95	108.39
Smoker (yes)	23380.07	972.51
Region: Northwest	-91.13	553.85
Region: Southeast	-656.61	542.67
Region: Southwes	-385.72	543.91

4.4 Charges for smokers or not Results

Table 4: Average Insurance Costs for Smokers and Non-Smokers

Smoker Status	**Average Cost**
No	14234.19
Yes	19234.76

The average insurance cost for smokers is significantly higher compared to non-smokers, confirming the importance of smoking as a key determinant of insurance premiums.

Furthermore, analyzing the histogram in Figure 7, which shows the distribution of insurance costs for smokers and non-smokers, it is possible to conclude that:

- The distribution for non-smokers (in pink) is more concentrated around lower insurance costs.
- The distribution for smokers (in blue) shifts towards higher costs.
- Although there is some overlap, it is evident that the tail for smokers extends further toward higher values, highlighting the influence of smoking on insurance costs.

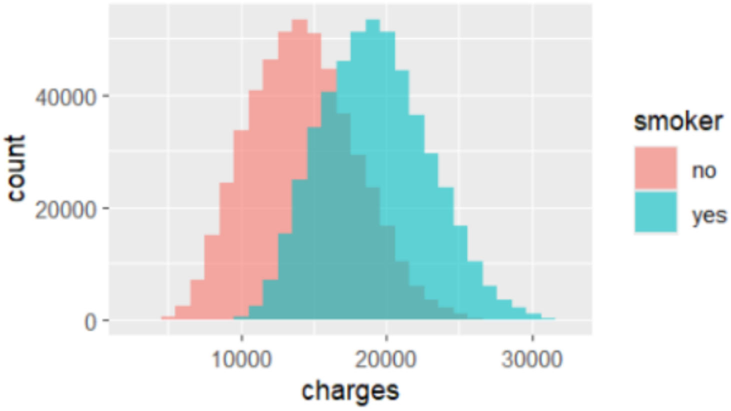


Figure 7: Histogram charges for smokers and not smokers

5 Conclusions

This report demonstrated the application of statistical models and resampling techniques to analyze the factors influencing the cost of health insurance premiums in the United States. The linear regression model revealed that the variables with the greatest impact on the insurance premium cost are age, BMI, smoking status, and region of residence. This finding was further confirmed by the non-parametric bootstrap, particularly highlighting smoking, age, and BMI as key drivers.

The non-parametric bootstrap and jackknife resampling techniques provide more robust estimates of the standard errors of the regression coefficients compared to the standard linear regression model.

6 Discussion

This study highlights the variables that require greater attention to avoid excessively increasing health insurance premium costs. Since the dataset is synthetic, it does not fully reflect the complexity of the real world and should be considered merely as an example for studying the real-world healthcare system. To be effectively utilized, it requires the application of Machine Learning techniques and the inclusion of additional variables to improve the accuracy of the predictions.

A R Code

```r
library(tidyverse)
library(ggplot2)
library(caret)
library(boot)

# Upload the dataset
insurance_data <- read.csv("insurance_dataset.csv")
summary(insurance_data)

# Check null values
missing_values <- colSums(is.na(insurance_data))
print(missing_values)

# Convert categorical variables into factors
insurance_data <- insurance_data %>%
  mutate(
    gender = as.factor(gender),
    smoker = as.factor(smoker),
    region = as.factor(region),
    medical_history = as.factor(medical_history),
    family_medical_history = as.factor(family_medical_history),
    exercise_frequency = as.factor(exercise_frequency),
    occupation = as.factor(occupation),
    coverage_level = as.factor(coverage_level)
  )

# Check the dataset
str(insurance_data)

# Relationship between variables
pairs(~ age + bmi + children + charges, data = insurance_data, main
     =’Scatterplot Matrix’)

# correlation matrix between numeric variables
correlation_matrix <- cor(insurance_data %>%
                            select_if(is.numeric))
print(correlation_matrix)

# Heatmap
library(reshape2)
library(ggcorrplot)
cor_melted <- melt(correlation_matrix)
ggplot(cor_melted, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2)), color = "white") +
  theme_minimal()

# Linear regression
full_model <- lm(charges ~ ., data = insurance_data)
summary(full_model)

# Selection of the most important variables
step_model <- step(full_model, direction = "both")
summary(step_model)

model <- lm(charges ~ age + bmi + children + smoker + region, data
    = insurance_data)
summary(model)
par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))
plot(model)
```

```
cat("R^2 full model:", summary(full_model)$r.squared, "\n")
cat("R^2 important variables model:", summary(model)$r.squared, "\n
    ")

# Resampling techniques
set.seed(123)
num_samples <- 1000

# Bootstrap
bootstrap_fun <- function(data, indices) {
  sample_data <- data[indices, ]
  model_boot <- lm(charges ~ age + bmi + children + smoker + region
      , data = sample_data)
  return(coef(model_boot))
}
bootstrap_results <- boot(data = insurance_data, statistic =
    bootstrap_fun, R = num_samples)
print(bootstrap_results)
cat("Confidence Intervals (Percentile Method):\n")
bootstrap_conf_int <- apply(bootstrap_results$t, 2, quantile, probs
     = c(0.025, 0.975))

# Jackknife
jackknife_fn <- function(data, i) {
  data_subset <- data[-i, ]
  model <- lm(charges ~ age + bmi + children + smoker + region,
      data = data_subset)
  return(coef(model))
}

n <- nrow(insurance_data)
jackknife_results <- matrix(NA, nrow = n, ncol = length(coef(lm(
    charges ~ age + bmi + children + smoker + region, data =
    insurance_data))))

for (i in 1:n) {
  jackknife_results[i, ] <- jackknife_fn(insurance_data, i)
}

jackknife_estimate <- colMeans(jackknife_results)
jackknife_se <- sqrt(((n - 1) / n) * rowSums((t(jackknife_results)
    - jackknife_est)^2))

print("Jackknife Estimates:")
print(jackknife_estimate)
print("Jackknife Standard Errors:")
print(jackknife_se)

smoker_charges <- insurance_data %>%
  group_by(smoker) %>%
  summarise(mean_charges = mean(charges))
print(smoker_charges)

ggplot(insurance_data, aes(x = charges, fill = smoker)) +
  geom_histogram(bins = 30, alpha = 0.6, position = "identity")
```

References

[1] Sridhar Streaks, *Insurance Data for Machine Learning*. Available at: `https://www.kaggle.com/datasets/sridharstreaks/insurance-data-for-machine-learning/data`.

[2] Sun, Jun Jun, *Identification and Prediction of Factors Impact America Health Insurance Premium*, PhD Thesis, National College of Ireland, 2020.

[3] Ward, Zachary J., Bleich, Sara N., Long, Michael W., and Gortmaker, Steven L., *Association of body mass index with health care expenditures in the United States by age and sex*, PloS one, Vol. 16, No. 3, pp. e0247307, 2021.

[4] Xu, Xin, Bishop, Ellen E., Kennedy, Sara M., Simpson, Sean A., and Pechacek, Terry F., *Annual healthcare spending attributable to cigarette smoking: an update*, American Journal of Preventive Medicine, Vol. 48, No. 3, pp. 326–333, 2015.

[5] Waples, Josef, *Jackknife Regression in R/RStudio using Mtcars*, Medium, 2020.