1 Introduction

1.1 Background

Health insurance is a critical component of personal financial planning and risk management, and its importance has grown in recent years. In 2016, the world spent 7.5 trillion US dollar on health (representing close to 10 percent of global GDP), and the average per capita health expenditure was 1,000 US dollar. These increasing patterns and trends identified in the report are confirmed by the 2016 data published in WHO's Global Health Expenditure Database (World Health Organization, 2018). As healthcare costs continue to rise, having adequate health insurance coverage can provide individuals and families with financial protection against unexpected medical expenses. In addition to providing financial security, health insurance can also improve access to healthcare services and promote preventative care, which can help individuals stay healthy and avoid more serious health issues down the line (Zhang et al., 2017).

Meanwhile, when determining the insurance charge, some factors might be important to consider, such as age, gender, health condition, and smoking status. Therefore, understanding the relevant attributes of insurance pricing can help individuals make informed decisions about their healthcare coverage and financial well-being.

1.2 Objective

In this project, our main goal would be to learn more about insurance and **we would focus on investigating the association between clients' demographic factors, lifestyle choices and their insurance premiums by using linear models.**

The resulting multiple linear regression model could potentially be a powerful tool for both insurance purchasers and insurance companies. For insurance purchasers, the model can provide valuable insights into their future insurance charges based on their individual characteristics (Yang et al., 2018). This information can help them save ahead for future high-cost insurance purchases and choose an insurance plan that is right for them. For insurance companies, the estimated model can be used to set their insurance charges based on the information of their customers. By understanding the relationships between these variables, insurance companies can tailor their insurance offerings to different segments of the market, and improve their profitability and customer satisfaction.

1.3 About the Dataset

Our data was found from Kaggle (Prediction of Insurance Charges), which contains detailed information about medical insurance, mainly the characteristics about the insurance clients and the insurance charges for them. The dataset was collected by Bob Wakefield from multiple sources and uploaded to Kaggle in 2020, and it contains 1338 sampled insurance data. However, it does not contain additional information about the specific source and time.

Now, we would provide a complete summary of the variables contained in this dataset that are used to describe the insurance clients:

Variable Name	Description
age	The age of the clients (in years)
sex	The sex of the clients (female or male)
bmi	The body max index: a person's weight in kilograms divided by the square of height in meters (in kg/m2)
children	The number of children the client has (eg. 0, 1, 2, 3...)
smoker	The smoking status of the client (yes for smokers, and no for non-smokers)
region	The region the client lives in the city (including southwest, southeast, northwest and northeast)
charges	The insurance charges for the customer (in dollars)

Based on our objective, we plan to use the characteristics of clients to model their insurance charges. Hence, the charges variable would be our response variable. The first six variables (age, sex, bmi, children, smoker, region) are possible explanatory variables.

2 Analysis

2.1 Preliminary Analysis

Before we fit the model, it is important to first explore the relationship between possible covariates and the response variable (charges) since this could provide us with some ideas about what variables to include in the model. Thus, in this section, we would analyze how the characteristics of insurance clients are related to their insurance charges based on visualizations.

2.1.1 Box Plot

To start with, we would first look at the influence of categorical variables with side by side box plot.

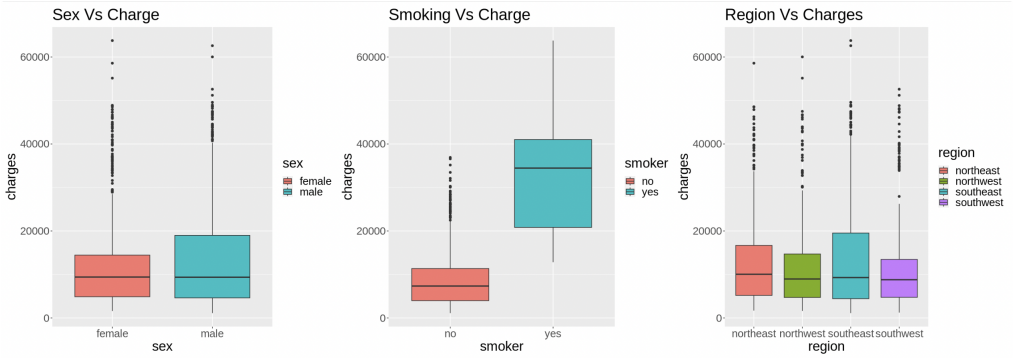


Figure 1: Side by Side Boxplot for Sex, Smoker and Region

Here, we examine the three categorical variables separately. By looking at the boxplots above, we can see that:

For different sex, the median insurance charges for males and females in our sample are close to each other. While the spread of males' insurance charges is a bit larger, there is no clear evidence from the plot that male clients would receive higher insurance charges than female clients.

For smoking status, it is clear from the plot that there is a difference between median insurance charges for non-smokers and smokers. Also, we can see that the variability of data for smokers is larger than the one for non-smokers. Overall, based on this plot, we expect smoking status to be a significant factor in building our model.

For regions, the boxplot for four different regions looks very similar. (i.e. the median value of insurance charges for four regions are close and the spread also looks close to each other) Hence, we would not expect the region to be an important factor here. This also makes intuitive sense since it is unlikely for clients to get different insurance charges just because they live in different regions of the city.

2.1.2 Scatter Plot

After looking at the three categorical variables, we would now explore the linear relationship between our continuous covariates and the response variable, and also the interaction between explanatory variables by using scatter plots.

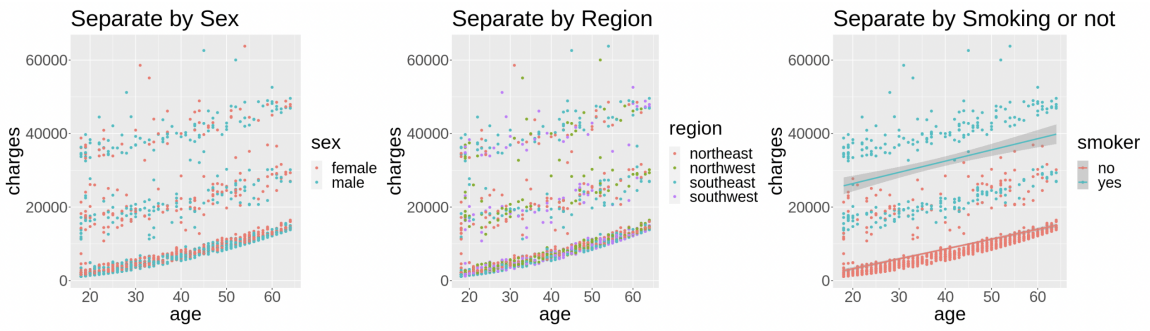


Figure 2: Scatter Plots for Age and Charges Index by Three Categories

By looking at Figure 2 above, we can see that there is a general positive linear relationship between age and charges. Also, sex and region do not seem to separate the data into different relationships, which matches what we observed from the boxplot.

While for smoking status, for both smokers and non-smokers, there is a positive linear relationship between age and charges but with a different intercept. The slopes are similar, suggesting there might be no strong interaction between smoking status and age. However, it seems that smoking status is not the best categorical data that separate the data here. From the plot, we can see three different linear relationships with the same slopes but different intercepts between age and charges. This hints that there might be another categorical variable with three levels that could better describe this pattern. We suspect that one possible choice is smoking status with levels: Never, Occasionally, and Frequently.

After investigating the relationship between age and charges, we would now look at how children and bmi influence the charges. Again, we create scatter plots indexed by sex, location and smoking status, separately. Similar to before, sex and location do not seem to have any influence on the relationship. Hence, in the following part, we would focus on presenting the relationship between children, bmi and charges, indexed by smoking status, and we attached other plots related to sex and location in Appendix for reference.

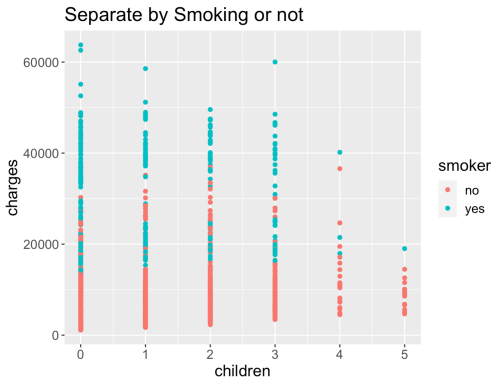


Figure 3: Scatter Plot for Children and Charges Indexed by Smoking Status

From Figure 3 above, since there is quite a large variability in charges at each positive x value (number of children), it is hard to visualize a clear linear relationship between children and charges. Meanwhile, there does not seem to have an interaction between children and smokers.

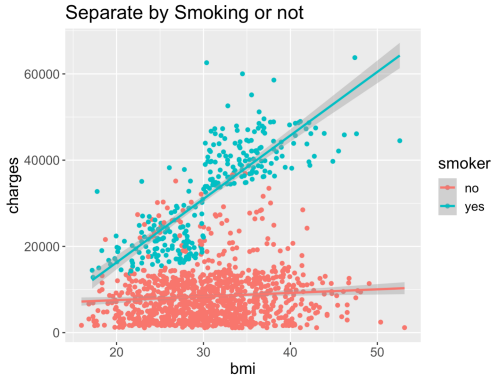


Figure 4: Scatter Plot for Bmi and Charges Indexed by Smoking Status

From Figure 4, it is clear that there are two different relationships between bmi and charges for different smoking statuses. For non-smokers, it seems that change in bmi does not influence the charges a lot; while for smokers, there is a positive linear relationship, suggesting a higher bmi means a higher charge. Hence, we think there is a strong interaction between bmi and smokers based on this figure.

2.2 Model Fitting

In this section, we would fit several different models and choose the best one based on some model selection criteria. We focus on using adjusted R squared (a model with a good prediction power will have an adjusted R squared close to 1), AIC (a model with a low AIC value is favourable), and residual plots (a good linear fit is indicated by a plot with a random pattern around zero) to evaluate our models. But before we start, since there are some categorical variables in our data, we would like to first describe the dummy variables that would appear in the model summary in R output:

1. smokeryes = 1 if the client smokes, 0 if the client does not
2. sexmale = 1 if the client is male, 0 if the client is female
3. regionnorthwest = 1 if the client lives in the northwest of the city, 0 otherwise
4. regionsoutheast = 1 if the client lives in the southeast of the city, 0 otherwise
5. regionsouthwest = 1 if the client lives in the southwest of the city, 0 otherwise

2.2.1 Model 1 (Full Model)

To start with, the first model that we fit includes all variables provided in the dataset. We only include an interaction term between bmi and smoker variables based on the visualizations from preliminary results, and we consider this as the full model.

From the model summary, we see that the adjusted R squared is 0.8398 and the AIC is 26517.57. For the coefficients, we noticed that the coefficient for sex and regionnorthwest have p-values greater than 0.05. Additionally, the two other dummy variables for the region are moderately significant. For age, smoker status, children and the interaction between bmi and smoker status, they are all highly significant. Although the bmi does not have a low p-value, it makes sense as the interaction term is significant. Compared with our preliminary analysis, most of the p-values from this model match with what we discovered in visualizations, but it is a bit surprising that children is a very important factor here.

Variable	age	bmi	smokeryes	sexmale	children
P-Value	$<$ 2e-16	0.35814	$<$ 2e-16	0.06079	3.06e-06
	regionnorthwest	regionsoutheast	regionsouthwest	bmi:smokeryes	
	0.12447	0.00160	0.00131	$<$ 2e-16	

Table 1: P-value for Model 1

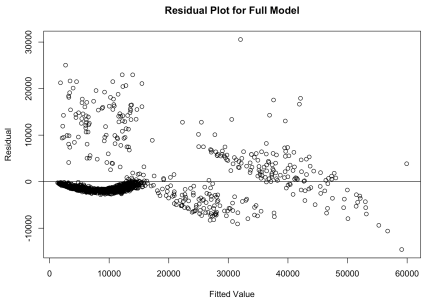


Figure 5: Residual Plot for Full Model

However, the residual plot for this model seems to have a very strange pattern as shown in Figure 5, which violates the linear model assumption. We can see that many observations are clustered in the region of residuals equal to zero. Ideally, we would expect a residual plot to show no pattern for a good linear fit.

2.2.2 Model 2 (Reduced Model)

In summary, some terms in Model 1 are not significant. Additionally, the residual plot for Model 1 shows a strange pattern, which does not indicate a good linear fit. In an attempt to improve the model and residual, we created a second reduced model by removing terms that do not contribute significantly.

For our reduced model, we incorporated all terms from Model 1 except the sex and region variables. Although regions are somewhat significant based on p-value, we still choose to remove it since it is not highly significant in the model and the visualization also indicated that the charges do not differ a lot for all four regions.

This model had an adjusted R-squared of 0.8382, which is lower than the previous full model but not too different. The AIC is 26526.9, which is in fact higher than the previous model. Additionally, all terms are highly significant except the bmi, which is becoming significant through the interaction term (as also seen in Model 1).

Variable	age	bmi	smokeryes	children	bmi:smokeryes
P-Value	$<$ 2e-16	0.82014	$<$ 2e-16	4.61e-06	$<$ 2e-16

Table 2: P-value for Model 2

By looking at the residual plot (Figure 6), we see that it is not too different from the one for the full model, hinting that we should try other methods like transformation to alleviate the problem.

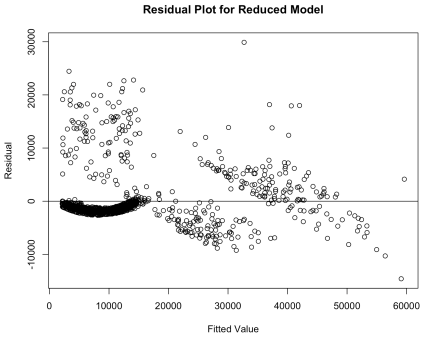


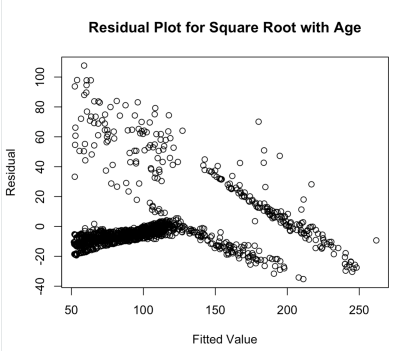
Figure 6: Residual Plot for Reduced Model

2.2.3 Models with Transformation

Although model 2 seems sensible based on the low p-values, its residual plot shows a clear pattern, meaning the basic assumptions for linear regression are violated. Additionally, the AIC value is slightly

higher than the previous model, which suggests that this model is not a great improvement from before. Hence, in this subsection, we aim to achieve a model with a better residual plot.

(a) Figure 7 (a)



(b) Figure 7 (b)

Figure 7: Residual Plots for Log and Square Root Transformations

We first try to add log and square root transformation towards charges in model 2. Both of these methods do not fix the problem in the residual plot as the residual plots show clear patterns in both cases (in Figure 7 above). We suspect that the problem is brought by the missing categorical variable that divides the age variable into three categories, as discussed in Figure 2. Hence, we try to investigate the effect of removing the age variable and we call this Model 3.

Model 3

Model 3 has a much better residual plot than all previous models as shown in Figure 8. Although the residual plot still shows some pattern for large fitted values, it is overall much more random. As we suspect that this pattern here is brought by the missing categorical variable, it can be hardly fixed purely without having more information about other explanatory variables. This model has children and the interaction between smoking status and bmi being very significant. The bmi and smokeryes are less significant, but again it is not a problem. Meanwhile, it has an adjusted R squared of 0.4936 which seems less satisfactory than before. However, since our aim is to explore association and not make predictions, the low adjusted R squared is not a big problem as the residua plot improves a lot. Additionally, the AIC of this model is 2669.087, which is very less compared to Models 1 and 2. This also indicates that this is a better model.

Variable	bmi	smokeryes	children	bmi:smokeryes
P-Value	0.00118	0.15327	2.41e-15	4.92e-08

Table 3: P-value for Model 3

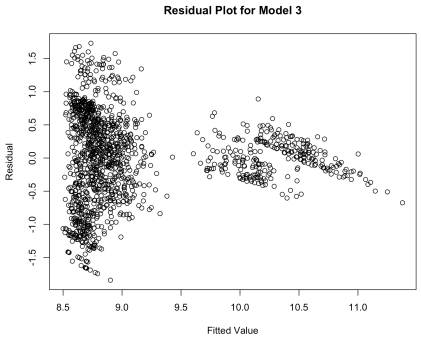


Figure 8: Residual Plot for Model 3

2.2.4 Exploring PCA as a Possibility

Based on the different model metrics that we used in this project, Model 3 yields the best results. However, we wanted to explore whether Principal Component Analysis (PCA) could help here.

We apply principal component analysis to the sample correlation matrix of the three continuous variables (age, children and bmi). We then take a look at the scree plot (shown in Figure 9), which shows us the proportion of variation in the data each principal component is able to explain. The scree plot shows us that all three principal components can explain about the same amount of variation in the data. Thus, it does not make sense to choose only the first or first two principal components to do the regression. Hence, we think that it is not very sensible to use principal component regression here. As if we incorporate all three principal components, the model would be similar to before.

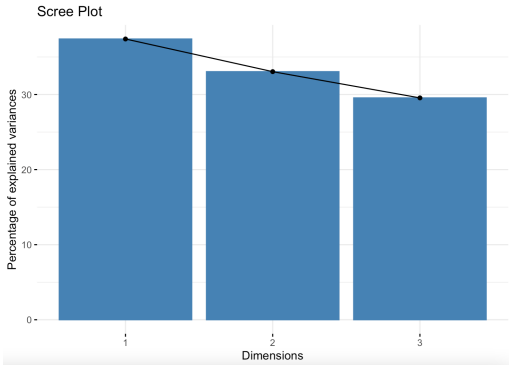


Figure 9: Scree Plot for PCA

2.3 Interpreting the Final Model (Model 3)

Based on all the models we investigate in Section 2.2, we decide to use Model 3 as our final model, and we will interpret the results from that model in this section. We first provide the equation of the fitted final model based on the R output.

$$\hat{y} = 8.3275 + 0.0108x_1 + 0.3175x_2 + 0.1190x_3 + 0.0389x_1x_2$$

where

$$y = \text{log of the insurance charges;}$$

$$x_1 = \text{the value for the bmi;}$$

$$x_2 = \begin{cases} 1 & \text{if the client smokes} \\ 0 & \text{otherwise} \end{cases};$$

$$x_3 = \text{ the number of children}$$

From the model above, we can see that bmi is positively associated with log charges since our model estimate that a unit increase in bmi would lead the estimated mean log charges to increase by 0.0108 assuming other factors are fixed. Besides, our model shows that the number of children is positively associated with log charges.

A very interesting thing to note from this model is the difference between the smoker and non-smoker categories. Compared with the non-smoker, the intercept of the fitted model is 0.3175 units higher if the client is a smoker. For the slope of bmi, it is 0.0389 units larger for smokers than non-smokers. Hence, these two coefficients reflect that a smoker tends to be charged a higher fee when purchasing insurance than non-smokers when they have the same bmi and children.

To summarize, our model provides evidence that both bmi and children are positively associated with the charges (or log(charges)). It also shows that smoking status interacts with bmi.

3 Conclusion & Discussion

From our final model, we can conclude that **there is a positive association between individual's bmi, smoking status, the number of children they have and the log of the insurance charges.** The adjusted R squared for the final model is 0.4936, which means that around 49% of the variation in the response variable (insurance charges) can be explained by the explanatory variables in the final model. Although the adjusted R squared is relatively lower than model1 and model2, our full model is the most sensible one since we care more about using a proper model to explain the association instead of getting a model that clearly violates the model assumptions but might possibly yield good prediction results.

In our final model, we did not consider the age of the insureds. However, we still think age should be important to the insurance charge based on the significant p-value in models 1 and 2 as well as the pattern from Figure 2. Age is also integral to the insurance industry, thus future studies could focus on a more thorough exploration of the missing categorical data that can be used to differentiate age brackets. In this case, the age variable could then be incorporated into a linear model that does not yield a pattern for the residual plot.

Meanwhile, there might also be other variables that are associated with insurance charge that is not included in our model. Hence, information about clients' history of other ailments (eg: heart disease, liver disease, etc.) and also other forms of dependencies (like elderly parents, etc.) could be collected to form a better model. Additionally, various types of medical insurance exist, including Exclusive Provider Organization (EPO), Health Maintenance Organization (HMO), Point of Service (POS), and Preferred Provider Organization (PPO). Having additional knowledge of which type the data refers to will result in building models more suitable for their intended purpose.

Based on all the points mentioned above, our final model can explain some of the variables that are associated with insurance charges, but to a limited degree. If future studies could achieve the points we mentioned in the previous paragraphs, we believe the model would yield high adjusted R Square and have residual plots that are more random.

4 Appendix

4.1 References

Wakefield, B. (2018, October 15). *Prediction of insurance charges*. Kaggle. Retrieved April 5, 2023, from https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender

World Health Organization, 2018. *Public spending on health: a closer look at global trends (No. WHO/HIS/HGF/HFWorkingPaper/18.3)*. World Health Organization. https://www.who.int/publications/i/item/WHO-HIS-HGF-HFWorkingPaper-18.3

Yang, C., Delcher, C., Shenkman, E. et al. Machine learning approaches for predicting high cost high need patient expenditures in health care. *BioMed Eng OnLine 17 (Suppl 1)*, 131 (2018). https://doi.org/10.1186/s12938-018-0568-3

Zhang, A., Nikoloski, Z., & Mossialos, E. (2017). Does health insurance reduce out-of-pocket expenditure? heterogeneity among China's middle-aged and elderly. *Social Science & Medicine*, 190, 11–19. https://doi.org/10.1016/j.socscimed.2017.08.005

4.2 Plots

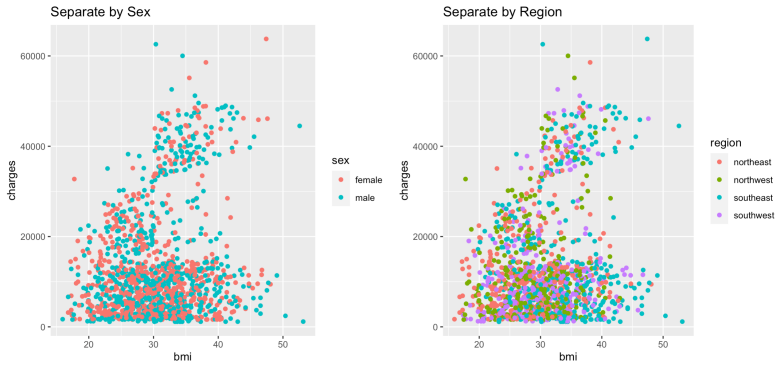


Figure 10: Scatter Plots for bmi and charges indexed by Sex and Region

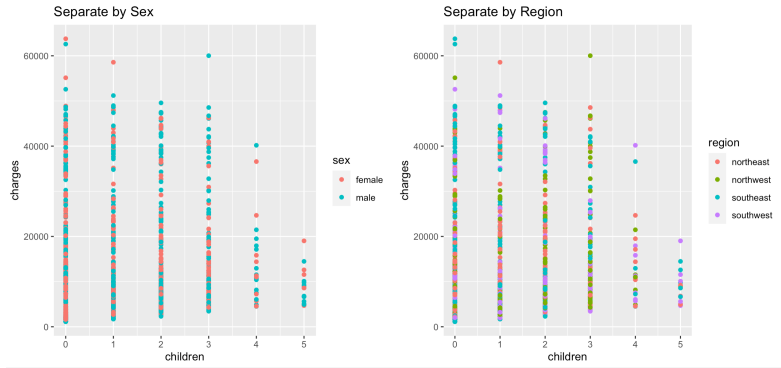


Figure 11: Scatter Plots for children and charges indexed by Sex and Region