

Predicting Medical Insurance Charges from Age, BMI and Smoking Status

DA Group 6

1 Introduction

Health insurance is an insurance that covers a varying portion of the medical expenses a person may incur. Risk is spread across all persons covered by the insurer. Insured individuals enroll in a health insurance policy and pay a recurring monthly or yearly cost (known as premiums) to a health insurance company. In return the company promises to reimburse a proportion of the medical costs of the insured. The health insurance company will set premiums based on the calculated risk of an individual.

The premium an individual pays, depends on a number of factors. In particular, the higher the risk of having health problems, the higher the premium will be. For instance, overweight or obese persons are more likely to develop heart disease when compared with people of normal weight. Moreover, smoking is linked with an increased risk of lung cancer when compared to individuals who do not smoke. Apart from these, there is a tendency for health insurance rates to escalate with increasing age, since older people are more prone to health problems. In these situations, one might expect these individuals to have a higher risk of large medical expenses.

The main goal of this report is to predict the total yearly medical costs (**charges**) a health insurance company will charge an individual based on the age (**age**) (in years), the corresponding body mass index (**bmi**) and from whether the insured smokes tobacco or not (**smoker**). The **bmi** is a measure of body fat, which is defined as the body weight (in kg) divided by the square of the body height (in m²). The analysed data consists of a sample of 364 individuals living in the South East region of the United States from a simulated dataset based on demographic statistics obtained from the US Census Bureau, approximating real world data. The insurance dataset contains 1338 observations, which contain hypothetical data on medical expenses for patients in the United States.

Section 2 consists of an exploratory data analysis to gain a better understanding of the distribution of the features used under this study. In Section 3 a backward elimination stepwise regression is used to select the best regression model to predict the insurance charges after fitting a number of regression models to the data. Moreover, the model assumptions are also checked here. Finally, Section 4 summarises the results obtained from the analysis and provides ideas for model improvement.

2 Exploratory Data Analysis

To get an idea of the distribution of the data, the following summary statistics were obtained for the categorical variable **smoker** (Table 1) and for the continuous variables **charges**, **age** and **bmi** (Table 2).

Table 1: Numbers of smokers and non-smokers.

Smoker	n
no	273
yes	91

Table 1 shows that out of the 364 individuals, only 25% of the individuals in the study sample smoke (273 non smokers vs 91 smokers).

Table 2: Summary statistics on insurance charges, age and bmi.

Variable	n	Mean	SD	Minimum	1st quartile	Median	3rd quartile	Maximum
age	364	38.94	14.16	18	26.75	39	51	64
bmi	364	33.36	6.48	19.8	28.57	33.33	37.81	53.13
charges	364	14735.41	13971.1	1121.87	4440.89	9294.13	19526.29	63770.43

Looking at Table 2, the mean age of the individuals is 38.94 years, with a standard deviation (SD) of 14.16 years and the middle 50% are between 26.75 and 51 years of age. Next, the middle 50% of the bmi lies between 28.57 and 37.81 kg/m^2 , with an average bmi value of 33.36 kg/m^2 and standard deviation 6.48 kg/m^2 . Finally, the middle 50% of the data for the medical costs (charges) lies between \$4440.89 and \$19526.29, with an average of \$14735.41. The variation in the mean total charges appear to be quite high, with a value of \$13971.1.

In order to measure the degree of association between the continuous variables in the study, the pairs scatterplot is plotted in Figure 1. The plot shows that there is no strong relationship between any two continuous variables. To confirm this, the correlation coefficients for each pair of variables was calculated. The correlation between the response variable **charges** and the explanatory variables **age** and **bmi** were found to be 0.311 and 0.143, respectively. Moreover, there does not seem to be any linear association between the two continuous explanatory variables **age** and **bmi** (0.02), implying that there is no evidence of multicollinearity in the data.

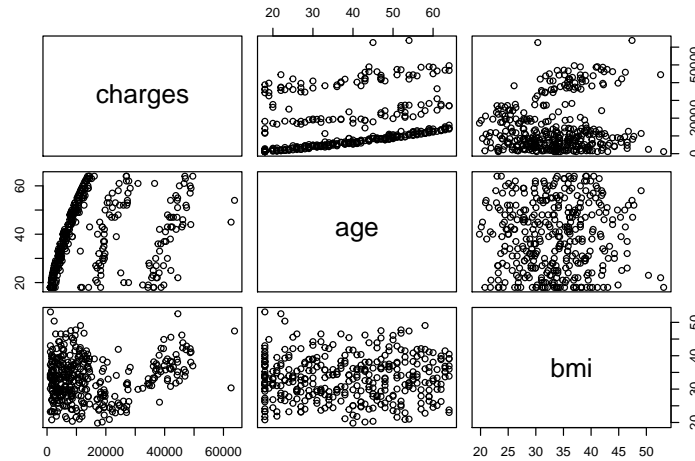


Figure 1: Pairs plot between charges, age and bmi

Figure 2 on the next page, shows two scatterplots of the insurance charges against each of the explanatory variables by smoking status of the individuals. From the left plot, showing the relationship between charges and age with smoking status accounted for, it is evident that older the individual, the higher their health insurance premium. The plot indicates that on average, people who do not smoke, pay less than \$40000 whereas smokers pay up to \$60000 on their health insurance. However, from the parallel slopes, the effect of age on insurance charges appears to be the same between smokers and non smokers. On the other hand, the slopes of the lines of the plot on the right are distinctly different, indicating the effect of bmi on insurance charges changes with smoking status. From the plot, as the bmi of a smoker increases, the corresponding insurance charges increase, whereas in nonsmokers, an increase in bmi does not seem to increase the insurance charges. From both plots it appears that smoking people pay more on their health insurance when compared with those who do not smoke.

To predict the medical insurance charges, a number of linear regression models will be fitted with age, bmi and smoking status as potential predictors.

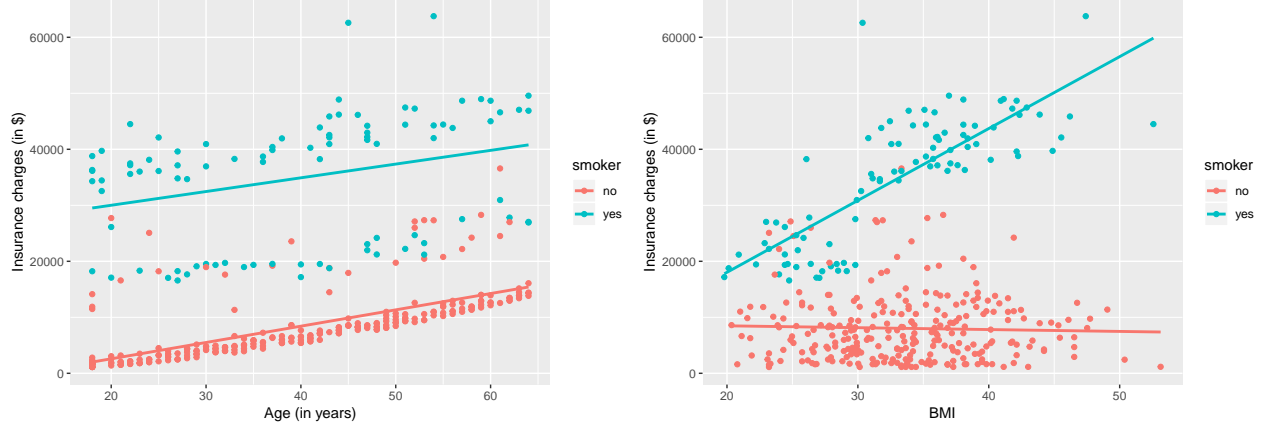


Figure 2: Scatterplots of insurance charges by age and smoking status (left) and insurance charges by bmi and smoking status (right)

3 Formal Data Analysis

Model selection is done via stepwise regression with the full interaction model initially fitted to the data as follows:

$$y_i = \alpha + \beta_{bmi}x_i + \beta_{age}x_i + \beta_{smoker} \cdot \mathbb{I}_{smoker}(i) + \beta_{bmi,smoker} \cdot \mathbb{I}_{bmi,smoker}(i) + \beta_{age,smoker} \cdot \mathbb{I}_{age,smoker}(i) + \epsilon_i$$

$$y_i = \alpha + \beta_{bmi} \cdot bmi_i + \beta_{age} \cdot age_i + \beta_{Smoker} \cdot \mathbb{I}_{Smoker}(i) + \beta_{bmi, smoker} \cdot \mathbb{I}_{bmi, smoker}(i) + \beta_{age, smoker} \cdot \mathbb{I}_{age, smoker}(i) + \epsilon_i$$

where

- α is the intercept of the regression line for non-smokers;
- β_{bmi} is the slope of the regression line for both smokers and non-smokers;
- bmi_i is the bmi of the i^{th} observation
- β_{age} is the slope of the regression line for both smokers and non-smokers;
- age_i is the age of the i^{th} observation
- β_{smoker} is the additional term added to α to get the intercept of the regression line for smokers; and
- $\mathbb{I}_{smoker}(i)$ is an indicator function such that

$$\mathbb{I}_{smoker}(i) = \begin{cases} 1 & \text{if the } i\text{th observation smokes,} \\ 0 & \text{Otherwise.} \end{cases}$$

Also, $\beta_{bmi, smoker} \cdot \mathbb{I}_{bmi, smoker}(i)$ and $\beta_{age, smoker} \cdot \mathbb{I}_{age, smoker}(i)$ correspond to the interaction terms.

The parameter estimates obtained after fitting the above full model are summarised in Table 3.

Table 3: Parameter estimates obtained from the full model $\text{charges} = \text{age} + \text{bmi} + \text{smoker} + \text{age.smoker} + \text{bmi.smoker}$

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-1056.473	1735.094	-0.609	0.543	-4468.730	2355.784
age	292.779	20.600	14.213	0.000	252.268	333.291
bmi	-66.809	46.480	-1.437	0.151	-158.216	24.598
smokeryes	-19119.137	3395.765	-5.630	0.000	-25797.291	-12440.984
age:smokeryes	-6.951	41.646	-0.167	0.868	-88.852	74.951
bmi:smokeryes	1386.060	86.177	16.084	0.000	1216.584	1555.536

The normality assumption for the full interaction model appear valid as the fitted values are randomly scatterd around mean zero and have constant variance for the explanatory variable age. However, for the explanatory variable bmi, the assumption of constant variability is dubious but the assumption of normally distributed residuals seems to hold.

From the results obtained in Table 3, we can conclude that there is insufficient evidence that $\beta_{\text{age, smoker}}$ differs from zero, as the corresponding confidence interval contains zero (-88.852, 74.951). Consequently, this term is eliminated and the following model with only one interaction term **bmi.smoker** is fitted to the data and the results obtained are shown in Table 4:

Table 4: Parameter estimates obtained from the model $\text{charges} = \text{age} + \text{bmi} + \text{smoker} + \text{bmi.smoker}$

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-997.201	1696.062	-0.588	0.557	-4332.666	2338.264
age	291.078	17.879	16.281	0.000	255.918	326.239
bmi	-66.614	46.402	-1.436	0.152	-157.868	24.639
smokeryes	-19408.629	2915.429	-6.657	0.000	-25142.093	-13675.165
bmi:smokeryes	1386.518	86.016	16.119	0.000	1217.359	1555.677

Table 4 shows that the confidence interval for the bmi (-157.868, 24.639) contains zero, implying that bmi is not statistically significantly related to insurance charges, when keeping all the other variables constant. However, the interaction term between **bmi** and **smoker** is found to be significant, since the confidence interval (1217.359, 1555.677) does not contain zero. The interaction between **bmi** and **smoker** can be seen in the right plot in Figure 2 where the regression lines for smokers and nonsmokers have different slopes . This suggests that the best fitted model to predict the insurance charges is the fitted model in Table 4. In order to verify this, the following stepwise model selection approach was also conducted below in Table 5:

Table 5: Model selection

Model	adj.r.squared	AIC	BIC
Full Model	0.8808755	7215.071	7242.351
Model 1	0.8811981	7213.099	7236.482
Model 2	0.7957825	7409.301	7428.787
Model 3	0.7717162	7448.862	7464.450
Model 4	0.7172804	7526.709	7542.298

Full Model corresponds to the full interaction model, in Table 3, **Model 1** consists of only one interaction term between **bmi** and **smoker** shown in Table 4, while **Model 2** is the parallel lines model. Finally, **Model 3**

and **Model 4** corresponds to the models with **age** and **smoker**, and **bmi** and **smoker**, respectively. Table 5 summarizes the Adjusted R^2 , AIC and BIC values obtained for each fitted model. The best model will be the one with that maximizes Adjusted R^2 and minimizes the AIC and BIC values, which in this case corresponds to **Model 1**. The same model found to be the best model when model selection was conducted via confidence intervals.

The equation of the best fitting model with parameter estimates for non smokers is as follows:

$$\widehat{\text{Charges}} = -997.201 + 291.078 \cdot \text{age} - 66.614 \cdot \text{bmi}$$

while the regression line for smokers is given by:

$$\begin{aligned}\widehat{\text{Charges}} &= -997.201 + 291.078 \cdot \text{age} - 66.614 \cdot \text{bmi} - 19408.629 + 1386.518 \cdot \text{bmi} \\ &= -20405.83 + 291.078 \cdot \text{age} + 1319.904 \cdot \text{bmi}\end{aligned}$$

When bmi remains constant, for every one year increase in the age of the insured, the corresponding insurance charges will increase by 291.078, whether the person smokes or not. Holding age constant, a one unit increase in bmi insurance charges for nonsmokers decrease by -66.614 but increase by 1319.904 for nonsmokers. Health insurance charges are significantly larger for smokers than non smokers.

In order for this analysis to be valid, the model assumptions for linear regression corresponding should be satisfied. The assumptions of zero mean and constant variance of the residuals for the explanatory variables and the fitted values are checked in Figures 3 and 4, respectively.

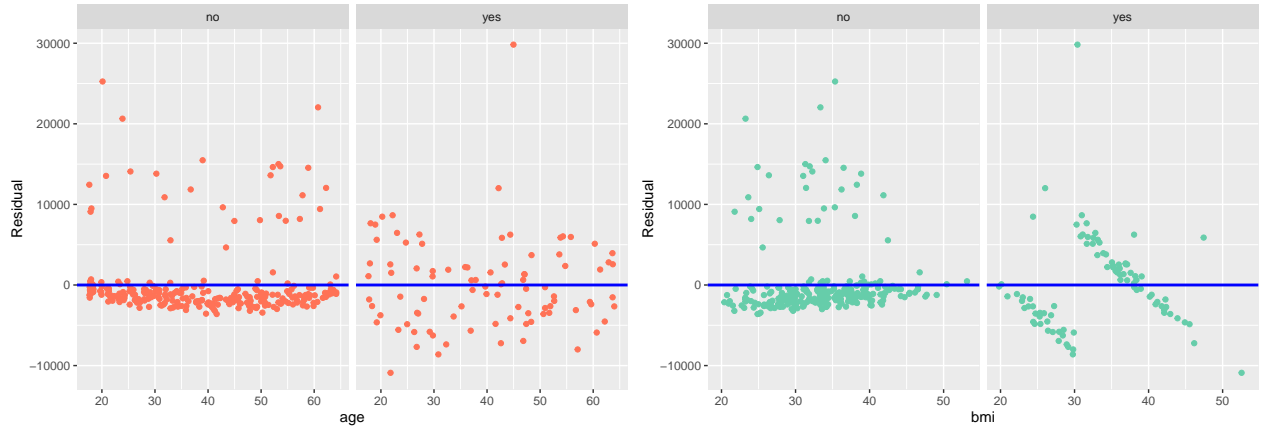


Figure 3: Residuals vs the explanatory variable age and bmi by Smoker.

The residuals in Figure 3 seem to be approximately evenly spread above and below the zero line for smokers and non smokers across all levels of the explanatory variables. Hence, the assumption that the residuals have mean zero appears reasonable. Moreover, the observations seem to be randomly scattered, implying constant variance, with the notable outlier in smokers. The latter could be due to the fact that there are relatively few smokers in the sample (91 smokers from Table 1). Figure 4 shows scatterplots of the residuals against fitted values by smoking status.

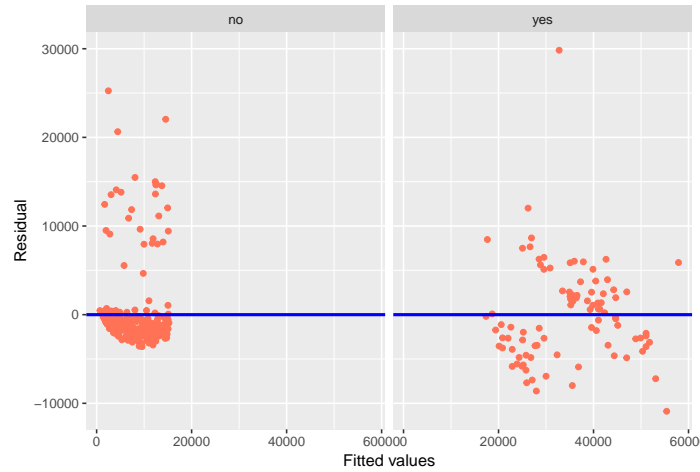


Figure 4: Residuals vs the fitted values by Smoker.

The points in Figure 4 have zero mean and constant variance, but there are a number residuals with large variance above the . The linear model assumptions are somewhat dubious, and further investigation could illuminate possible underlying covariate patterns.

Finally, to assess whether the residuals are normally distributed, a histogram is plotted in Figure 5. The residuals for smokers seem to be bell-shaped and centered at zero, with the exception of one outlier, having a large positive residual. On the other hand, the histogram for people who do not smoke is centered around zero skewed to the right. Thus, the assumption of normally distributed random errors might be slightly dubious. However, overall, both histograms appear to be relatively symmetrical and bell-shaped, and hence, the assumption of normally distributed random errors seems plausible for the best fitted model.

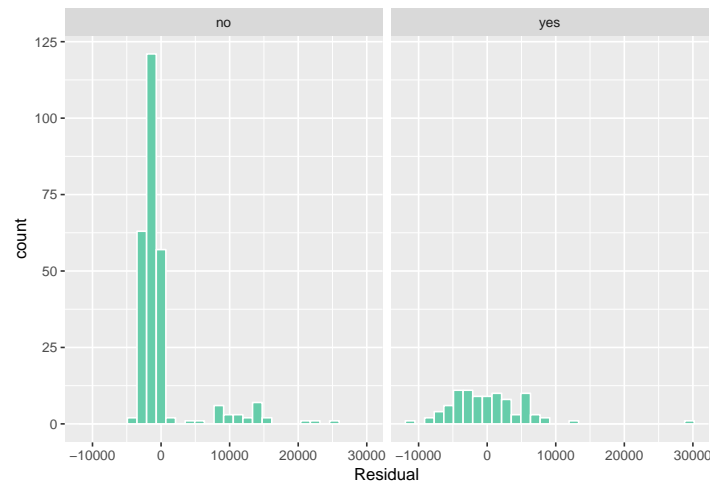


Figure 5: Histogram of residuals.

4 Conclusions

- Limitations?

How to improve:

- A larger dataset would be better/preffered to judge the assumptions.

- Also, can consider transformations on our variables.
- More variables \rightarrow there seems to be some confounding factors