Forecasting Health Insurance Cost

2022-12-16

Introduction

Background

Medical and healthcare services are one of the inelastic demand in the modern society. However, personal medical costs have been rising rapidly and becoming a major public health issue. In 2020, health expenditure constitutes of 19.7% of US GDP. Understanding the factors affecting health outcomes and costs as well as predicting future expenditure is essential to many parties involved in the ecosystem of healthcare. These knowledge can aid risk management and Value at Risk (VaR) based pricing strategies for medical insurance companies, drive more informed decision-making process for healthcare policy makers, and educate individual about healthy lifestyle choices. According to various research studies, factors that influence personal medical care costs include obesity (Cawley et al., 2021), ageing (Alemayehu et al., 2004), smoking (Hall et al., 2016), etc. Specifially on smoking habit, in the issue of PLOS Medicine, Lightwood and Glantz have estimated how much, on average, a 1% reduction in smoking prevalence in a US state was associated with reduced health costs in that state a year later (Lightwood, 2016). Moreover, research about health care costs shows wide variations in spending across the U.S., with spending more than three times higher in some regions than others(KFF Family Foundation, 2009). Patients in higher cost areas were not necessarily receiving better care; rather, the cost variations were explained by the availability and volume of services used by similar patients. More interesting factors of influence are found in recent year, such as single women spend nearly twice as much of their income on health insurance as single men (Festa, 2022), suggesting gender, potential marital status, and number of dependents as important considerations when deciding health insurance cost.

In this report, we seek to explore the gathered influencing factors, such as demographics, personal habits, and household situations, and their potential interactions to predict the individual medical cost billed by health insurance annually. Conducting various types of regression fit and model selection, we attempt to interpret the relative influence of each factors based on the model output, controlling for other factors. Using more complex models like generalized additive model (GAM), we can extrapolate new data beyond the scope of the known observations and answer question like "what is the likely range of the medical costs of 60-year-old females living in southwest US who smokes with high BMI?" These results will help insurance companies classify their clients and pricing and policy-makers to attend to and potentially intervene certain high-risk populations.

Data

We used a public medical insurance dataset that was published in the book Machine Learning with R by Brett Lantz in 2013 and later updated on Kaggle in 2017. The dataset consists of 1338 observations and covers 7 variables, including the amount of medical cost billed by health insurance and potentially important factors related to the charges. There are no missing data in any of these columns.

- Age - a numeric variable that records the age of the primary beneficiary. Its range is from 18 years old to 64 years old, consecutively.
- Sex - a binary categorical variable that records the primary beneficiary's gender. Only two values are presented: "female" and "male".

- Body Mass Index (BMI) - a numeric variable that is calculated by the primary beneficiary's weight in kilograms divided by height in meters, providing an understanding of the ratio of body metrics. Note that BMI screens for weight categories that may lead to health problems, but it does not diagnose the body fatness or health of an individual. A healthy range of BMI is from 18.5 (kg/m^2) to 24.9 (kg/m^2). Our variable in the dataset ranges from 15.96 (kg/m^2) to 53.13 (kg/m^2).
- Number of Children - a numeric variable that records the number of children / dependents covered by the primary beneficiary's health insurance. Note that a child is no longer a dependent to their parents at the age of 26, and have the choice to apply for their own health insurance line at the age of 18. Our variable in the dataset ranges from 0 (no child) to 5 children.
- Smoking Status (Smoker) - a binary categorical variable that records if the primary beneficiary's identifies as a smoker or not. Only two values are presented: "yes" (smoker) and "no" (non-smoker).
- Region - a categorical variable that records if the primary beneficiary's residential area in the US. Only four values are presented: "northeast", "southeast", "southwest", "northwest".
- Charges - a numeric variable that records the amount of annual medical cost billed by health insurance (in dollars) to the primary beneficiary. This variable ranges from 1121.9 to 63770.4.

Objective

The main objective of this report is to build regression models with medical costs as the response and conduct model selection for the best predictive power by exploring the distribution and interaction of several factors (geographic information, personal demographic and health information, as well as lifestyle habits) and their relationship with the individual medical costs. Based on the selected models, we also aim to interpret the relative impact of each factor to the predicted insurance charges and help insurance companies classify their clients and pricing with predicted outcome from new beneficiary's information.

The secondary objective is to investigate whether possessing a habit of smoking will lead to distinctly higher insurance charges, controlling all other factors. We will explore both singular effect of smoking and interactive effect of smoking and other factors to generate a more comprehensive analysis accounting for the potential influence of smoking to all aspects of our life. This question is particularly brought up as smoking is a factor that the beneficiaries can personally decide to take on or not, even though they understand the detrimental effect to their health and family.

Exploratory Data Analysis

The response variable medical charges is log-transformed to achieve a normal distribution. Independent variables Sex, Number of Children, Smoking Status, and Region are treated as factors due to their categorical nature; Age and BMI remain numeric.

Inspecting each variable independently in Figure 1, we can see that higher BMI, older age, and possessing a smoking habit potentially contribute to higher charges.

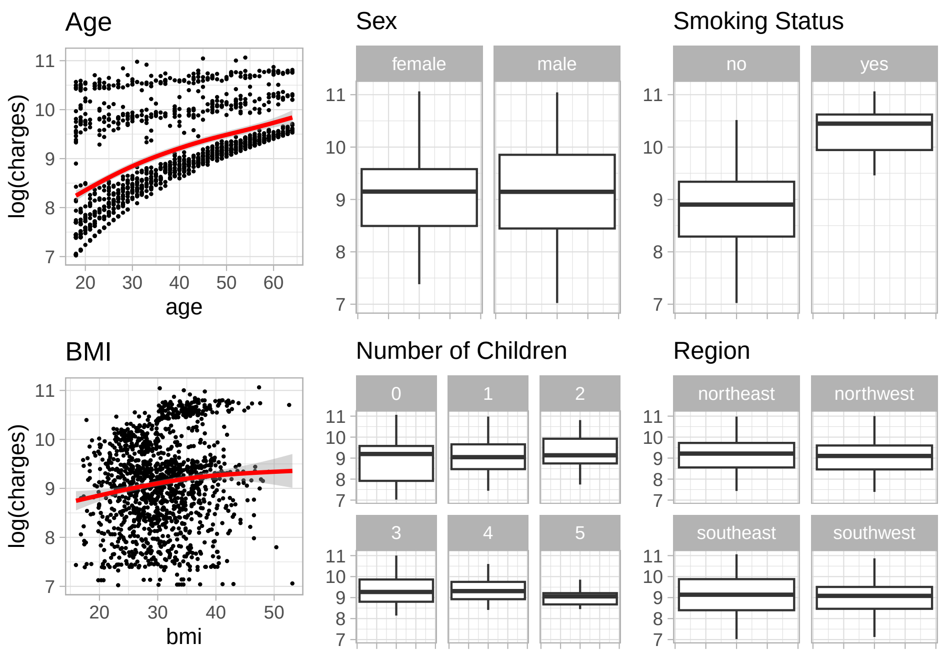

Figure 1: Relationship between medical charges and 6 independent variables

In order to explore the interaction effect between some variables, we decide to choose selected variables in focus and visualize their effect. Since one of our objective is to analyze the impact of possessing a habit of smoking, we attempt to discover all possible interactions between smoking and other five variables. From the distinct patterns between smoker and non-smoker groups in the sample plots in Figure 2, we can see that smoking habit affect the relationship between medical charges and Age, BMI, and Number of children covered by the insurance plan. Exploring more combinations, we discover Region and Age are also involved in potential interaction as to affecting the final charges (See Appendix Figure 8).

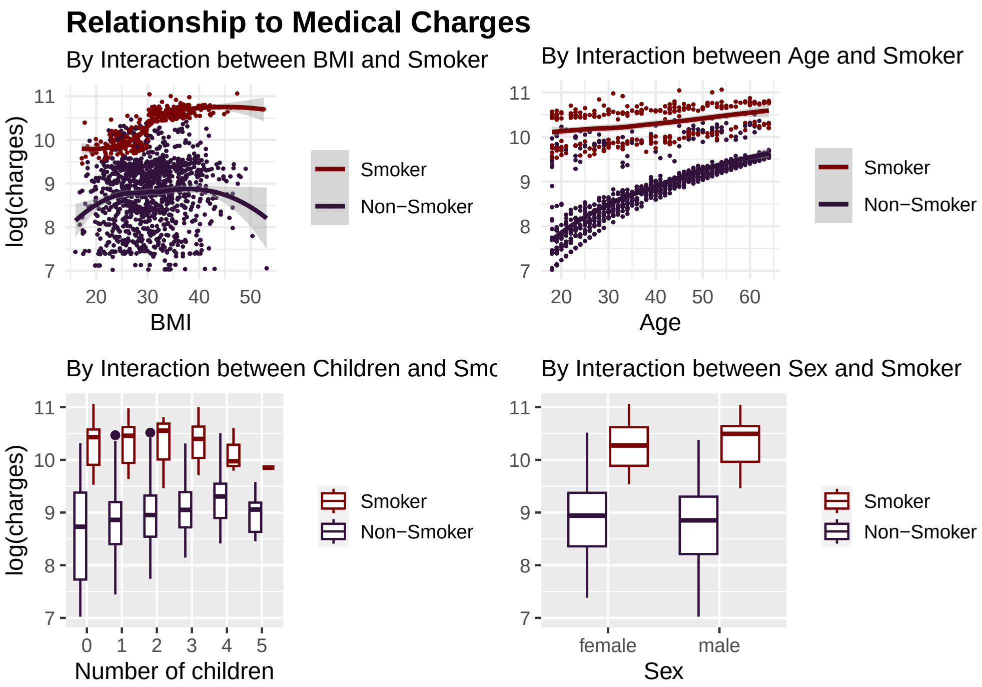

Figure 2: Interaction effects between smoking status and other variables

Methodology

For modeling the relationship between the medical charges and six demographics and personal habits variables, we decide to utilize various types of regression models, including linear least squares regression (with and without shrinkage, such as ridge and Lasso regression), nonlinear polynomial regression, and generalized additive model (GAM). We use AIC and validation set mean squared error (MSE) to evaluate all the models. For AIC, We calculate the metrics based on the entire dataset. For validation set MSE, we conduct a 7:3 validation split, where the models are trained on 70% of the data and tested on 30% of the data to get the testing MSE.

Linear Regression Models:

Least Squares Regression:

We first fit the simplest model, a least squares linear regression model. All the variables are included: Age and BMI are kept as numeric variables. Number of children, Sex, Smoker, and Region are categorical variables. Then, we fit a linear model with all selected interaction terms in the EDA section, namely Smoker & Age, Smoker & Sex, Smoker & BMI, Smoker & Number of Children, and Region & Age. From the output Table 1 below, we can see that the interaction model has a lower AIC and MSE compared to the linear model without interactions, leading to a better fit for the data.

Table 1: Linear Model (without Regularization) Selection

Model	Interactions	Response	Number of Parameters	AIC	MSE
l1	No interactions	Log transformed insurance charges	12	1630.32	66051108
l2	**Smoke or not & Sex, Smoke or not & Age, Smoke or not & BMI, Smoke or not & Number of Children, Region & Age**	**Log transformed insurance charges**	**23**	**1220.38**	**24253712**

The selected least squares model (without regularization) is:

$$\begin{aligned} log(Charges) \quad = \quad & \beta_0 + \beta_1 \times Smoker + \beta_2 \times Age + \beta_3 \times Region + \beta_4 \times bmi + \beta_5 \times Numberofchildren + \beta_6 \times Sex \\ & + \beta_7 \times Smoker * Sex + \beta_8 \times Smoker * Age + \beta_9 \times Smoker * BMI \\ & + \beta_{10} \times Smoker * Numberofchildren + \beta_{11} \times Region * Age \end{aligned}$$

The above simplified model format does not specify each categorical factor's different levels of values. Please refer to Appendix Table 10 for each level's corresponding coefficient.

Ridge and Lasso Shrinkage Models:

Based on the least squares model and interaction model mentioned above, we decide to apply shrinkage method with regularization that helps shrink the coefficients, reduce their variance, and prevent multi-collinearity. Thus, shrinkage method often leads to better model fits. We first use ridge regression and select the best λ by minimizing MSE based on 10-fold cross validation (CV).

Using ridge regression, all coefficients are shrunken towards 0 (Figure 3), but no variable selection is performed. In the least squares model, the coefficient (see Appendix Table 11) with the largest absolute value (1 order of magnitude larger than the rest) is of Smoker (when the beneficiary identifies as a smoker). This is consistent with our EDA on the importance of "smoker or not". In the interaction model(Appendix Table 12), the coefficients with the largest absolute values are of BMI & Smoker (the interaction term between BMI and Smoker), Age, 2-children category, Age & Smoker, and male (descending order of magnitude).

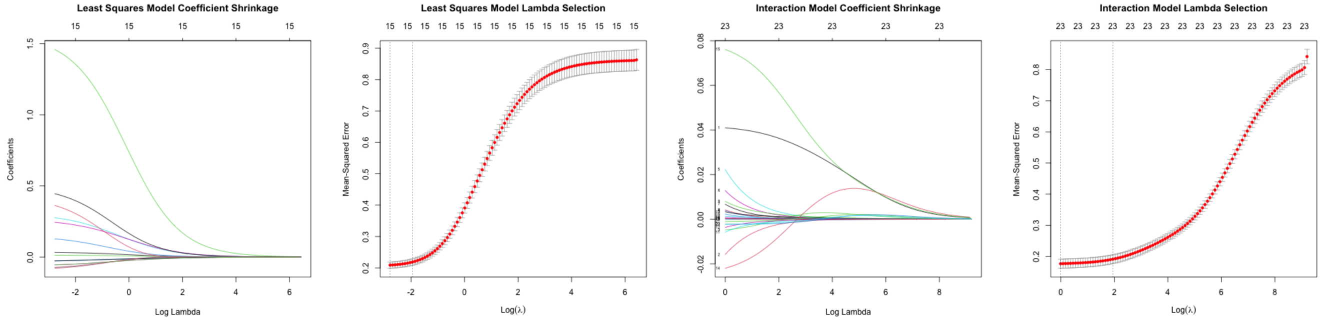

Figure 3: Coefficient Shrinkage for Ridge Regression Models

Ridge regression on the interaction model produces a lower MSE compared to ridge regression on least squares model (Table 2). However, its MSE is higher than that of the interaction model without any regularization, showing that ridge regularization is not leading to better model performance.

Table 2: Ridge Regression Model Selection

Model	Interactions	Response	Lambda	Number of Parameters	MSE
ri1	No interactions	Log transformed insurance charges	0.6793598	12	104077925
ri2	**Smoke or not & Sex, Smoke or not & Age, Smoke or not & BMI, Smoke or not & Number of Children, Region & Age**	**Log transformed insurance charges**	**0.9914619**	**23**	**35328082**

Thus, we proceed to Lasso regression (Figure 4). Comparing to ridge regression, Lasso is able to perform variable selection, shrinking unimportant variables' coefficients to exactly zero without incurring much loss in information. Similarly, we select the best λ by minimizing MSE based on 10-fold cross validation. In the least squares model (see Appendix Table 13), the coefficient with the largest absolute value (1 order of magnitude larger than the rest) is of Smoker. In the interaction model (Appendix Table 14), the coefficients of the interaction term between 4-children category and Smoker, and the interaction term between 5-children category and Smoker are shrunken to 0. The coefficient with the largest absolute value is of Smoker (1 to 2 orders of magnitude larger than the rest), followed by Number of children, Region, the interaction term between Number of children and Smoker, and Male (descending order of magnitude).

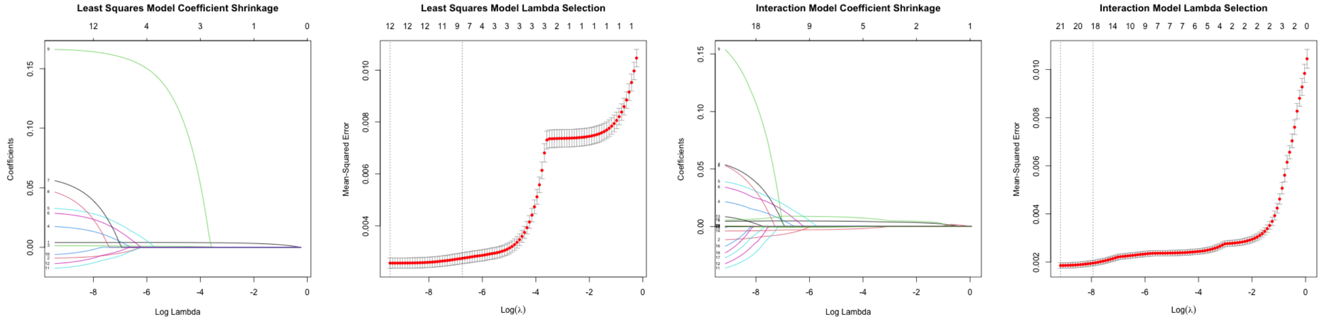

Figure 4: Coefficient Shrinkage for Lasso Regression Models

In Table 3, we can see that Lasso regression on the interaction model produces a lower MSE compared to Lasso regression on least squares model. Moreover, this MSE is lower than that of an interaction model without regularization, leading to a better fit of the data. This also shows that Lasso is able to select significant variables, prevent overfitting, and produce a simpler model that leads to better generalization by involving fewer parameters.

Table 3: Lasso Model Selection

Model	Interactions	Response	Lambda	Number of Parameters	MSE
la1	No interactions	Log transformed insurance charges	0.0000788	13	69823570
la2	**Smoke or not & Sex, Smoke or not & Age, Smoke or not & BMI, Smoke or not & Number of Children, Region & Age**	**Log transformed insurance charges**	**0.0001056**	**22**	**23813495**

Polynomials Regression Models:

Proceeding from linear model fits, we attempt to include orthogonal polynomial fits on selected variables in our current dataset. Based on variable types and the EDA section, we decide to modify previous linear fits on numeric variables Age and BMI to polynomial fits while keeping Sex, Smoking Status, and Region as factors with different levels. Notice that we also attempt to cast Number of children back to numeric and fit a polynomial regression for comprehensiveness. (However, the results below prove that Number of children should stay as categorical factor.)

The process of polynomial fitting starts from univariate modeling based on the three chosen numeric variables individually. Each variable is subject to both MSE and 10-fold CV Error evaluations with increasing degrees of polynomials. Based on the yielded values, we select the associated degrees with minimum error rates as our chosen degrees of polynomial fit of the corresponding variable. Then, we move on to a generalized multivariate model with polynomial and factor terms, encompassing all predictors with minor variations of the full model. The full models are compared with each other using AIC and MSE metrics. The selected model with the lowest AIC will be kept and added interaction terms to render the final polynomial model.

Univariate Modeling:

In order to first decide the degrees of polynomials, we implement MSE and CV Error for degree selection univariately. For predictor Age, the test MSE plots (Figure 5) over different degrees is presented below, indicating degree-4 as the optimized choice with the lowest error value. (Note that the response variable Medical Charges has been log transformed, so we take the exponent of the output predicted values.) Then, we conduct 10-fold CV on the current training set. However, since CV randomly split the data each time, we get two different optimal degrees from two CV trials with different seeds (see Appendix Figure 9). To maximize the randomization of sampling and retrieve a more generalized optimal degree, we repeat the 10-fold CV for 1000 iterations (with random seeds) and keep track of the degree with the lowest error per iteration. The count plot below demonstrates that degree-3 is more frequent compared to degree-2 in terms of yielding a better test error estimation.

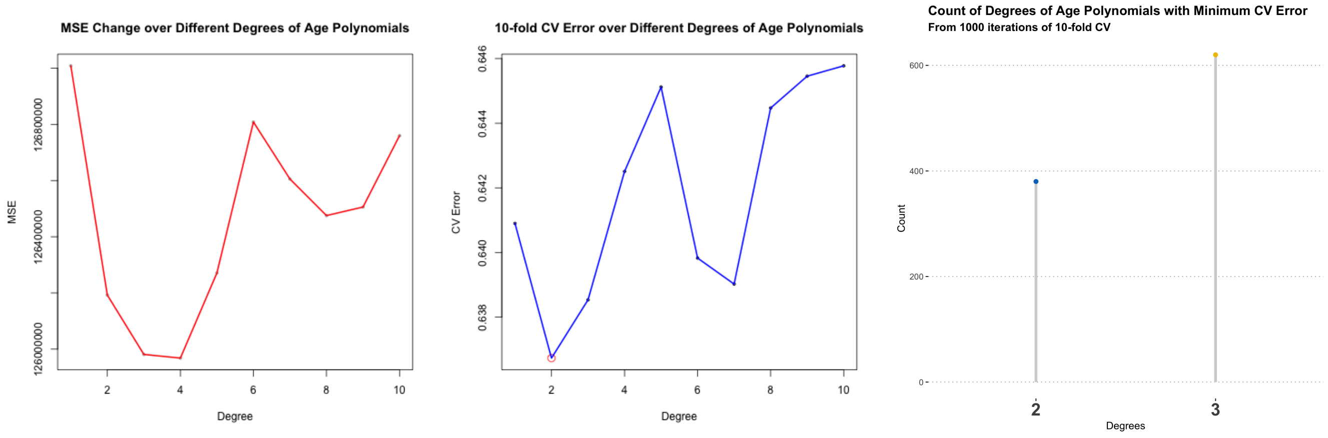

Figure 5: Choosing Optimal Degree for Polynomials of Age

Similarly, we reproduce the process on BMI and Number of children (see plots in Appendix Figure 10). For BMI, results from MSE and repeated CV correspond that a linear fit will generate the lowest error. For Number of children, if we treat it as numeric as mentioned above, MSE suggests a linear fit while CV suggests degree-2 polynomials. Notice that by observing the data, the number of children only ranges from 0 to 5, discretely. Thus, we decide to cast it back to factor abiding by the nature of the variable instead of a linear fit under MSE suggestion. Table 4 below generalizes potential optimal degrees generated from the process:

Table 4: Potential Optimal Degrees Chosen by MSE and CV

Term	MSE Chosen	CV Chosen
Age	4	3
BMI	1	1
Children	As factor	2

Multivariate Model Selection

In order to fit a generalized polynomial model, we incorporate the variables with selected degrees and the remaining variables that have been held unchanged. From previous discoveries (two potential degrees for both Age and Number of children), we first attempt four base models without interaction between predictors. Table 5 shows the four different model parameters and their corresponding AIC and test MSE. Aiming at small AIC and low MSE values, we select the second model with degree-4 polynomials on Age, factorized Number of children, BMI, Sex, Smoking Status, and Region as our final base without interactions.

Table 5: Base Generalized Polynomial Model Selection

Model	Varying Predictors	Same Predictors	Response	Number of Parameters	AIC	MSE
p1	Degree-3 polynomial on age, Number of children (factor)	BMI, Sex, Smoke or not, Region	Log transformed insurance charges	16	1626.90	65625300
p2	**Degree-4 polynomial on age, Number of children (factor)**	**BMI, Sex, Smoke or not, Region**	**Log transformed insurance charges**	**17**	**1618.74**	**63880783**
p3	Degree-3 polynomial on age, Degree-2 polynomial on number of children	BMI, Sex, Smoke or not, Region	Log transformed insurance charges	13	1629.26	66878556
p4	Degree-4 polynomial on age, Degree-2 polynomial on number of children	BMI, Sex, Smoke or not, Region	Log transformed insurance charges	14	1621.88	65502505

Proceeding with the selected base, we add the previously identified interaction terms. The corresponding AICs are listed in Table 6. Comparing the AIC values, though P2 has a slightly lower MSE, we decide to select model P1 with interactions effects that uses a linear function of age instead of polynomial fit. This gives better predictive power and more interpretability than the polynomial interaction fit, which could lead to overfitting on the training set.

Table 6: Generalized Polynomial Model with Interaction Terms Selection

Model	Base Model	Varying Interactions between	Same Interactions between	Number of Parameters	AIC	MSE
P1	**p2**	**Smoker or not & Age, Region & Age**	**Smoker or not & Sex, Smoker or not & BMI, Smoker or not & Number of children (factor)**	**28**	**1209.15**	**22563314**
P2	p2	Smoker or not & Degree-4 polynomial on age, Region & Degree-4 polynomial on age	Smoker or not & Sex, Smoker or not & BMI, Smoker or not & Number of children (factor)	40	1227.86	22539880

The selected polynomial model is:

$$\begin{aligned} log(Charges) &= \beta_0 + (\beta_1 \times Age + \beta_2 \times Age^2 + \beta_3 \times Age^3 + \beta_4 \times Age^4) \\ &+\beta_5 \times BMI + \beta_6 \times Numberofchildren + \beta_7 \times Sex + \beta_8 \times Smoker + \beta_9 \times Region \\ &+(\beta_{10} \times Smoker * Age + \beta_{11} \times Smoker * BMI + \beta_{12} \times Smoker * Numberofchildren \\ &+\beta_{13} \times Smoker * Sex + \beta_{14} \times Region * Age) \end{aligned}$$

where Number of children, Sex, Smoker, and Region are treated as factor (categorical variables that have a fixed and known set of possible values) and the interaction effects related to Age use the first order Age function. The above simplified model format does not specify each categorical factor's different levels of values. Please refer to Appendix Table 15 for each level's corresponding coefficient.

Generalized Additive Models:

To incorporate flexible nonlinearities in several variables, generalized additive models (GAM) were fitted. Indeed, GAMs can fit separate nonlinear functions for each predictor while retaining the additive structure of linear models, allowing us to examine and interpret the nonlinear effects of each predictor individually. Based on the EDA, the relationships between Age and Medical Charges and between BMI and Medical Charges were rather non-linear. We choose natural cubic spline to model these nonlinearities because it can stabilize polynomials fits near boundaries and, more importantly, reduce uncertainties near boundaries. One of the goals of this project is to produce a reliable prediction of medical charges, which allows insurance companies to be more confident in setting their premium prices. Using natural cubic spline can help us achieve this goal by reducing the uncertainties near boundaries at some cost of complexity. The degree of freedoms for Age and BMI were determined using 10-fold cross validation.

Moreover, according to EDA, interactions between a few variables appeared to be significant and worth investigating. Therefore, two GAMs were first fitted: one without interaction effects and one with interaction effects. We then performed a series of ANOVA tests (Table 7) in order to determine which of these models is the best: a GAM without interaction effects M_1, a GAM with interactions effects that uses a linear function of Age and BMI in the interactions M_2, a GAM with interactions effects that uses a non-linear function (natural cubic spline) of Age and BMI in the interactions M_3. We found compelling evidence that a GAM with linear functions of Age and BMI in interaction effects is better than a GAM that does not include any interaction effects at all. However, there was no evidence that non-linear functions of Age and BMI in interactions are needed, looking at the p-values. Moreover, M_2 also has the lowest AIC and MSE compared to the other two models. Based on the results of the ANOVA, AIC, and MSE, we selected M_2 (Table 8).

One thing to note is that including interaction terms in GAM improved the predictive power of our model but at the cost of interpretability. Therefore, we select GAM only for prediction purpose not inference. The full model summary can be found in Appendix.

Table 7: ANOVA for Comparing GAMs

Model	Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)	Significance
M1	1307	253.0715	NA	NA	NA	NA	
M2	1296	181.5496	11	71.52190	46.65556	0.00000	***
M3	1267	176.5711	29	4.97851	1.23185	0.18542	

The selected GAM is:

$$\begin{aligned} log(Charges) &= \beta_0 + \beta_1 \times Smoker + \beta_2 \times Region + \beta_3 \times Numberofchildren + \beta_4 \times Sex \\ &+f_1(Age) + f_2(BMI) \\ &+(\beta_5 \times Smoker * Age + \beta_6 \times Smoker * BMI + \beta_7 \times Smoker * Numberofchildren \\ &+\beta_8 \times Region * Age) \end{aligned}$$

where Number of children, Sex, Smoker, and Region are treated as factor (categorical variables that have a fixed and known set of possible values) and the interaction effects related to age and BMI use the linear function. Here, f_1 is a fitted natural spline on Age with 5 degrees of freedom; f_2 is a fitted natural spline on BMI with 14 degrees of freedom. Natural spline visualizations along with plot diagnosis for all parameters can be found in Results section. The full model summary can be found in Appendix Table 16.

Table 8: Generalized Addition Model Selection

Model	Varying Interactions between	Same Interactions between	Same Predictors	Response	AIC	MSE
M1	No interactions	No interactions	Natural spline on BMI with 14 degrees of freedom, Natural spline on Age with 5 degrees of freedom, Sex, Smoke or not, Region, Number of children (factor)	Log transformed insurance charges	1630.8	66993678
M2	**Smoker or not & BMI**	**Smoker or not & Sex, Smoker or not & Age, Smoker or not & Number of children (factor), Region & Age**	**Natural spline on BMI with 14 degrees of freedom, Natural spline on Age with 5 degrees of freedom, Sex, Smoke or not, Region, Number of children (factor)**	**Log transformed insurance charges**	**1208.7**	**22431011**
M3	Smoker or not & Natural spline on BMI with 14 degrees of freedom	Smoker or not & Sex, Smoker or not & Age, Smoker or not & Number of children (factor), Region & Age	Natural spline on BMI with 14 degrees of freedom, Natural spline on Age with 5 degrees of freedom, Sex, Smoke or not, Region, Number of children (factor)	Log transformed insurance charges	1229.5	18899718

Results

Comparing selected models from linear regression (with shrinkage), polynomial regression, and GAM, we pinpoint GAM as our final chosen model with maximum predictive ability and relative interpretability. Figure 6 below plots the relationship between each feature and the response Medical Charges in the selected GAM model. (Evaluation plots for the selected polynomial model is attached in Appendix Figure 11. Since only BMI-related predictors are different between the selected polynomial and GAM, we generalize both models' explanations for analysis here.) Each plot displays the fitted function and point-wise standard errors. Age and BMI are fitted using natural splines with five and fourteen degrees of freedom, respectively. Step function were fitted to the qualitative variables, Smoker, Region, Number of children, and Sex.

These figures can be easily interpreted. The top right panel indicates that holding all other variable constant, medical charges tend to be higher for people who smoke compared to those who don't. The top right panel indicates that holding all other variable constant, medical charges tend to be highest for intermediate values of age, and lowest for the very young and very old. The middle left panel indicates that holding all other variables constant, medical charges tend to be higher for people who come from northeast followed by northwest than those from south. There was no evident difference between people from southeast and southwest. The middle right panel indicates that holding all other variables constant, medical charges tend to increase with BMI with some fluctuation in the intermediate values of BMI: the higher a person's BMI, the higher their medical charges annually, on average. The bottom left panel indicates that holding all other variables constant, medical charges tend to increase with number of children a person has, while this pattern does not apply to those who have three children. The bottom right panel indicates that holding all other variables constant, medical charges tend to be higher for females than males. [**For the in-depth description of "holding all other predictors constant", see Appendix (Baseline Metrics for Models).**]

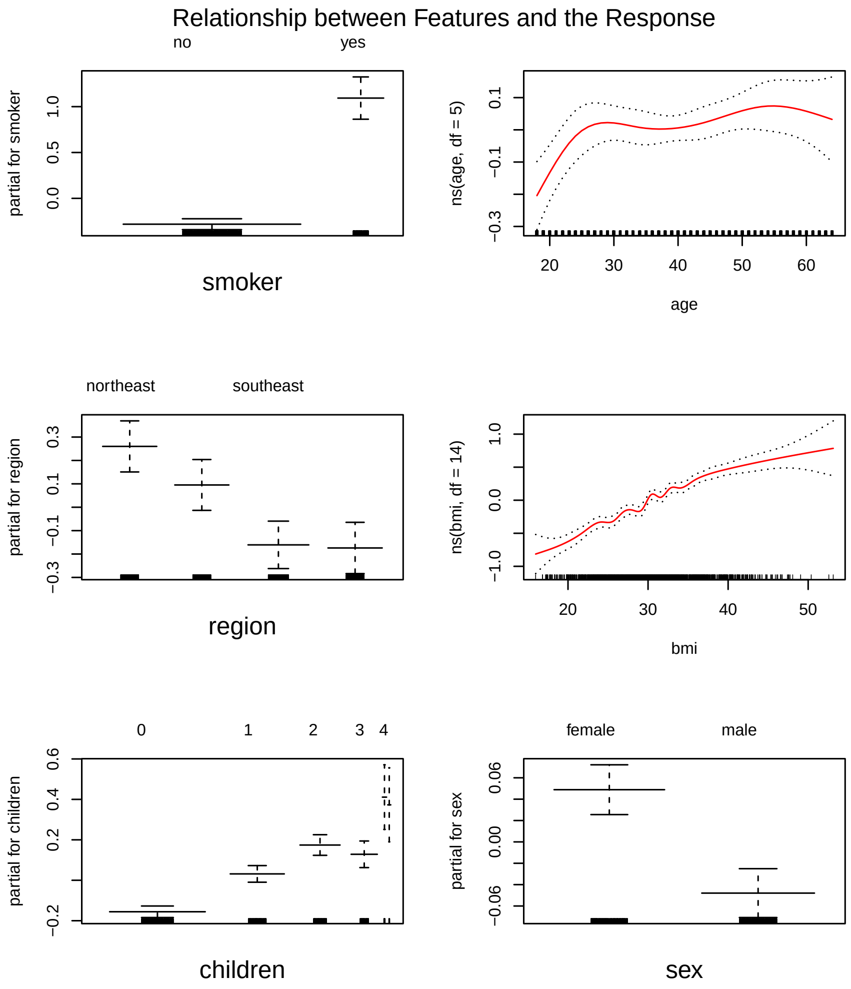

Figure 6: Relationship between Features and the Response

Most of these findings are intuitive. For instance:

- Beneficiary who smoke are more likely to get lung problems such as pneumonia and lung cancer as time passes, so they may be billed more on medical cost and healthcare than those who don't smoke. This answers part of our question in the secondary objective, which proves that smoking, when considered singularly, has a relative impact on medical charges billed by insurance.
- Beneficiary with higehr BMI tend to have higher risk for certain diseases such as heart disease, high blood pressure, type 2 diabetes, gallstones, breathing problems, and certain cancers. All of these could lead to a higher yearly medical cost.
- Northeast of U.S. on average has a higher cost of living than any other regions, which may suggest a higher expenses on healthcare.
- Beneficiary with more children also tend to be older and thus become more prone to medical services.
- Female beneficiary tend to spend more on healthcare than males might be due to the fact that women's health issues primarily revolve around chronic conditions and menopausal symptoms in the population aged 45 to 64 years. With the onset of menopause, the risk of cardiovascular disease (CVD), breast cancer, and osteoporosis increases significantly. However, it is very surprising that middle-aged people tend to have higher medical cost than young and old people.
- Intuitively, very young or old people are more susceptible to different kinds of diseases, resulting in more expensive medical cost than other age groups. However, it is very surprising that middle-aged people (40 to 60 years old)

tend to have relatively higher medical cost than young and old people. Future studies can investigate on why middle-aged people tend to spend more on healthcare.

Overall, GAM retain some interpretabilities compared to other non-linear models, but the interactions effects remain uninterpretable. Therefore, we decided to use GAM for mainly prediction and interpret interactions from the polynomial fit. For example, Figure 7 (left) shows the interaction between the beneficiary's BMI and if they identify as a smoker. Setting non-smoker as the baseline, as BMI metrics increase, medical insurance charges for the smokers surge exponentially compared to the non-smokers. This is due to the fact that a smoker with high BMI indicates an unhealthy state of the body, which will lead to various lung problems / immune system syndrome under the habit of smoking plus the potential threats from obesity. Thus, incorporating both factors, the insurance company is predicted to charge more compared to the non-smoking low-BMI group.

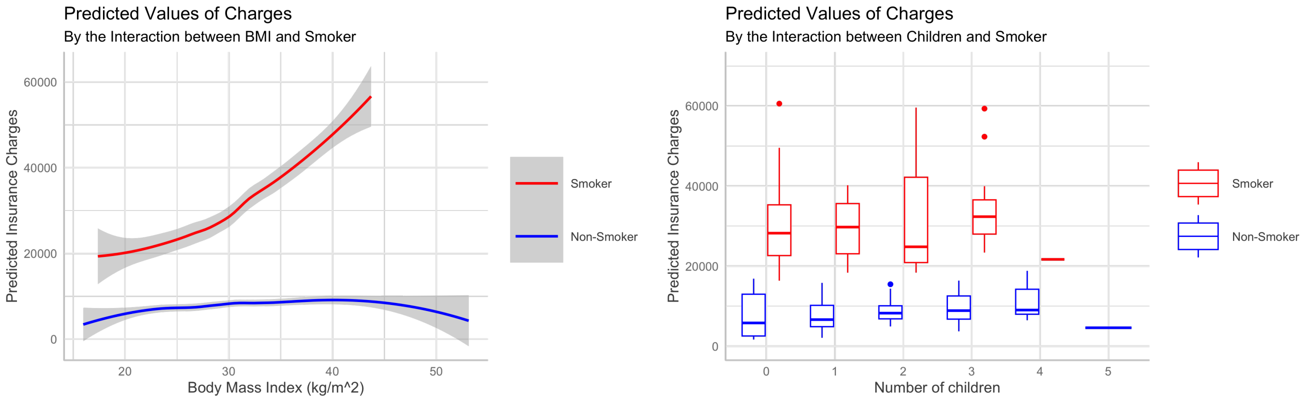

Figure 7: Interaction Terms in Polynomial Model Fit

Moreover, Figure 7 (right) shows that, with the interaction between Smoker and Number of children, beneficiaries who identify as non-smokers have lower and more steady growth of the median predicted charges among different numbers of children / dependents. As the number of children goes up, the non-smoker charges start from a wide confidence interval at no-child beneficiaries, peak at the four-child size, and suddenly drop to the lowest at five-child size. The increasing trend of charges to the growing number of children is intuitive, as the policyholder's health insurance coverage might be used to their dependents' medical cost. However, for the smokers, their predicted insurance charges present more variability. For example, the smoker insurance charges start with no obvious increase from no-child beneficiaries to one-child, experience a sudden median drop (but wider confidence interval) at two-child size, and bounce back to the peak at three-child size. Compared to the non-smoker group, we suspect that beneficiaries who identify as smokers prefer a smaller size of children number as of their negative effect on children health with regard to smoking habit. The wide interval at dependent size 2 validates the assumption. Anywhere after size-2 is shown to present more caution to a household with the potential threat of second-hand smoke. Note that in both groups, there is a sudden drop for large number of children (smokers after size 4 and non-smokers at size 5) due to either data singularity or two or more beneficiaries covering for the children expense within the same household. Final output summary with full interaction terms significance and coefficients can be found in Appendix. These interactions on smoking status lead us to answer the second part of the secondary objective of our report, which is possessing a habit of smoking not only singularly affects the medical charges, but also presents interactive impact on demographic variables, which collaboratively influence the insurance companies to set a higher charge on the smoker group.

Conclusion

Cost forecasting is a crucial component in the US healthcare insurance system with a $2.1 trillion market and about 300 million Americans covered. This report dives deep into a crucial question on health insurance forecasting: how personal demographic, health information, geographic information, and lifestyle habits influence and predict medical costs. From the model results, there are implications regarding the gender gap in healthcare costs. Since females tend to spend more on healthcare at an older age, more intervention can be taken to mitigate the burst of post-menopause diseases. An unexpected finding that middle-aged people spend more on healthcare also deserves attention. Although it's possible that this population undertake the responsibility to support and pay for their family's cost, it might also imply that the onset age of certain diseases become earlier. If that's the case, insurance companies need to reset their assumptions and design new premium strategies. More investigation should be done on this finding in the future.

For the secondary objective, we highlight smoking habit as a significant factor that not only generates more healthcare costs but also affects how age, BMI, sex, and number of children on coverage influence the costs. For example, beneficiaries identified as smokers, in general, bear higher medical charges compared to people who don't have a habit of smoking. Smoking also exacerbates the negative impact of BMI and age on body health with the interaction effect. In fact, the habit of smoking should not be considered simply as a factor of predicting health insurance cost, but more so on its alarming effect on people's health in general. According to Hall, smoking remains a leading cause of preventable death in the United States, where it is responsible for more than 480,000 deaths per year, including nearly 42,000 deaths from secondhand smoke exposure. Thus, our report functions as a call for population-based tobacco control policies as a societal investment by governments, and an alert for the general audience to reconsider the negative effect of smoking and shift to a healthier lifestyle.

One meaningful application for accurately predicting yearly medical costs is to help insurance companies set health premiums. A popular metric they often use is named Value at Risk (VaR), which is also known as quantile risk measure or quantile premium principle. Instead of setting the expected value as the health premiums for a population of interest, VaR takes the quantiles of choice. Table 9 below shows 75%, 95% and 99% quantile of medical costs grouped by age, sex and BMI according to our dataset. Being able to predict medical costs with reliability and certainty, therefore, can provide useful information on how to set health premium in a reasonable range. Understanding what factors can explain medical charges also helps public health professionals identify high-risk populations and hidden trends so that they can establish early alert system for potential healthcare crisis. For example, reproductive specialists can caution prospective parents (if they fall into the non-smoking category) that they are at risk of spending a lot more on healthcare cost for every additional member in the household. This will lead to informed-decision making process and reduces the pain of under-preparing for newborns, potential causing more health problems for infants and financial burden in the family.

Table 9: Value at Risk Pricing Grouped by Age, Sex, and BMI

Age group	Sex	BMI group	75% quantile	95% quantile	99% quantile
18-33	female	Normal	14676.78	22529.72	25530.65
18-33	female	Obesity	8673.84	37271.67	52578.87
18-33	female	Overweight	5312.17	19046.76	21615.75
18-33	female	Underweight	14731.11	27415.90	31670.53
18-33	male	Normal	5346.97	20427.72	22650.68
18-33	male	Obesity	33668.47	38820.62	44554.35
18-33	male	Overweight	16412.63	20348.03	25404.05
18-33	male	Underweight	2775.19	10818.60	12427.28
34-49	female	Normal	19444.27	22771.26	25555.09
34-49	female	Obesity	11863.98	43759.44	46198.36
34-49	female	Overweight	9348.37	23635.19	26712.32
34-49	female	Underweight	15986.94	18416.00	18901.81
34-49	male	Normal	15820.70	22482.68	32299.73
34-49	male	Obesity	31980.28	42660.47	46106.21
34-49	male	Overweight	20352.78	30408.87	37975.28
34-49	male	Underweight	6259.53	6564.34	6625.30
50-64	female	Normal	20257.29	26698.13	27097.97
50-64	female	Obesity	15421.26	47298.18	48793.76
50-64	female	Overweight	17071.89	29388.64	32882.45
50-64	female	Underweight	11597.66	12882.96	13140.02
50-64	male	Normal	21706.97	26720.85	29453.80
50-64	male	Obesity	26734.04	48511.60	51927.93
50-64	male	Overweight	19691.03	30053.31	30274.29
50-64	male	Underweight	11534.87	11534.87	11534.87

Limitations and Future Work

One major limitation of this study is that our data is lack of comprehensiveness as of the complex topic on health insurance prediction. For example, the dataset itself has been cleaned and manipulated when we retrieve it, such as no missing data, which leads to potential prediction bias of our final selected model if applied on real-world observations. Moreover, the dataset is only constituted of six key predictors, which is not enough to account for complex and massive beneficiary information. Even within each single variable, some levels of the categorical values are prone to singularity issue, such as when the number of children is larger than four. Therefore, we seek to collect more realistic and comprehensive data in the future to build a model with better accountability and predictive power.

Another limitation is related to the model itself. The current model selection process follows the standard regression model fitting and evaluation metrics on MSE and AIC. However, if with more beneficiary information, the model complexity will grow in exponential form. Also, the selected GAM lacks adequate interpretability other than predictability, which reminds us to build a more applicable model, such as using Bayesian hierarchical modeling with known priors on a subset of variables (since our key predictors are mostly common demographics input), or conducting variable selection when the pool of predictors grows larger.

References

Alemayehu B, Warner KE. The lifetime distribution of health care costs. Health Serv Res. 2004 Jun;39(3):627-42. doi: 10.1111/j.1475-6773.2004.00248.x. PMID: 15149482; PMCID: PMC1361028.

Cawley J, Biener A, Meyerhoefer C, Ding Y, Zvenyach T, Smolarz BG, Ramasamy A. Direct medical costs of obesity in the United States and the most populous states. J Manag Care Spec Pharm. 2021 Mar;27(3):354-366. doi: 10.18553/jmcp.2021.20410. Epub 2021 Jan 20. PMID: 33470881.

Festa, Liz. "All the Single Ladies … Pay More for Insurance." ValuePenguin, ValuePenguin, 24 Oct. 2022, https://www.valuepenguin.com/insurance-costs-gender-gap-study.

Hall W, Doran C. How Much Can the USA Reduce Health Care Costs by Reducing Smoking? PLoS Med. 2016 May 10;13(5):e1002021. doi: 10.1371/journal.pmed.1002021. Erratum in: PLoS Med. 2017 Apr 12;14 (4):e1002295. PMID: 27164007; PMCID: PMC4862676.

Hardy, Mary R. An Introduction to Risk Measures for Actuarial Applications. CAS Exam Study Note, 2006, https://www.researchgate.net/publication/242469445_An_Introduction_to_Risk_Measures_for_Actuarial_Applications.

KFF. "Explaining Health Care Reform: How Do Health Care Costs Vary by Region?" KFF, THE HENRY J. KAISER FAMILY FOUNDATION, 30 Nov. 2009, https://www.kff.org/health-costs/issue-brief/explaining-health-care-reform-how-do-health/.

Lightwood, James, and Stanton A. Glantz. "Smoking Behavior and Healthcare Expenditure in the United States, 1992–2009: Panel Data Estimates." PLOS Medicine, vol. 13, no. 5, 2016, https://doi.org/10.1371/journal.pmed.1002020.

Appendix

Baseline Metrics for Model

- When the beneficiary is female.
- When the beneficiary's residential region is northeast.
- When the beneficiary's number of children / dependent is zero.
- When the beneficiary does not identify as a smoker.

Additional Plots and Summary Output

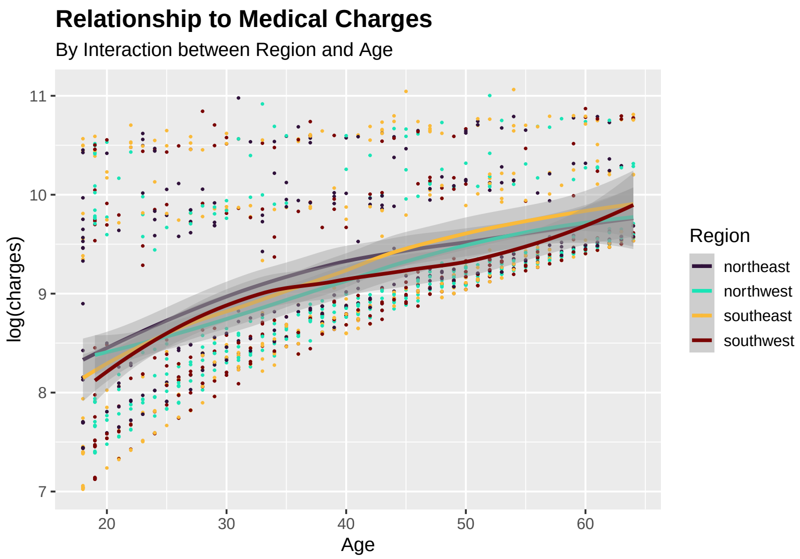

Figure 8: Interaction effects between Region and Age

Table 10: Selected Interaction Model without Regularization

term	estimate	std.error	statistic	p.value	Signif. codes
(Intercept)	7.243	0.086	84.348	0.000	***
smokeryes	1.386	0.146	9.490	0.000	***
age	0.038	0.002	24.443	0.000	***
regionnorthwest	-0.143	0.089	-1.606	0.109	
regionsoutheast	-0.415	0.087	-4.792	0.000	***
regionsouthwest	-0.401	0.089	-4.502	0.000	***
bmi	0.001	0.002	0.560	0.576	
children1	0.204	0.029	6.969	0.000	***
children2	0.357	0.033	10.823	0.000	***
children3	0.318	0.039	8.093	0.000	***
children4	0.585	0.083	7.074	0.000	***
children5	0.533	0.094	5.677	0.000	***
sexmale	-0.101	0.023	-4.361	0.000	***
smokeryes:sexmale	0.101	0.053	1.909	0.056	.
smokeryes:age	-0.033	0.002	-17.589	0.000	***
smokeryes:bmi	0.051	0.004	12.113	0.000	***
smokeryes:children1	-0.248	0.067	-3.693	0.000	***
smokeryes:children2	-0.329	0.071	-4.658	0.000	***
smokeryes:children3	-0.303	0.081	-3.737	0.000	***
smokeryes:children4	-0.579	0.238	-2.436	0.015	*
smokeryes:children5	-0.290	0.395	-0.736	0.462	
age:regionnorthwest	0.002	0.002	0.911	0.362	
age:regionsoutheast	0.007	0.002	3.247	0.001	**
age:regionsouthwest	0.006	0.002	2.968	0.003	**

Table 11: Least Squares Ridge Regression Model

term	s0
(Intercept)	7.3637840599
age	0.0339787037
sexmale	0.0020630982
bmi	0.0111439611
children1	0.0005076234
children2	0.0371867735
children3	0.0255777701
children4	0.0072079036
children5	0.0016192769
smokeryes	0.2782956938
regionnorthwest	-0.0025992782
regionsoutheast	-0.0043908185
regionsouthwest	-0.0184868311

Table 12: Least Squares Ridge Regression Model with Interactions

term	s0
(Intercept)	7.3926914157
age	0.0409597300
sexmale	-0.0158901646
bmi	-0.0048139895
children1	0.0039008092
children2	0.0223747980
children3	0.0128743416
children4	0.0067102689
children5	0.0041980027
smokeryes	0.0080181207
regionnorthwest	0.0023046318
regionsoutheast	-0.0058508049
regionsouthwest	-0.0037131016
sexmale:smokeryes	0.0033063676
age:smokeryes	-0.0220191093
bmi:smokeryes	0.0761090664
children1:smokeryes	-0.0001123569
children2:smokeryes	0.0013117785
children3:smokeryes	0.0007737479
children4:smokeryes	0.0002655750
children5:smokeryes	0.0003445808
age:regionnorthwest	-0.0010512652
age:regionsoutheast	-0.0022448731
age:regionsouthwest	-0.0022793500

Table 13: Least Squares Lasso Regression Model

term	s0
(Intercept)	1.97040
age	0.00401
sexmale	-0.00905
bmi	0.00131
children1	0.01747
children2	0.03271
children3	0.02862
children4	0.05600
children5	0.04645
smokeryes	0.16614
regionnorthwest	-0.00612
regionsoutheast	-0.01761
regionsouthwest	-0.01369

Table 14: Least Squares Lasso Regression Model with Interactions

term	s0
(Intercept)	1.985242
age	0.004622976
sexmale	-0.01153432
bmi	-1.075754e-06
children1	0.02153547
children2	0.03896208
children3	0.03410351
children4	0.05360324
children5	0.05294988
smokeryes	0.156726
regionnorthwest	-0.001282904
regionsoutheast	-0.036962
regionsouthwest	-0.03261877
sexmale:smokeryes	0.008624693
age:smokeryes	-0.003923684
bmi:smokeryes	0.005490503
children1:smokeryes	-0.01724203
children2:smokeryes	-0.02724688
children3:smokeryes	-0.02292508
children4:smokeryes	.
children5:smokeryes	.
age:regionnorthwest	-0.0001184835
age:regionsoutheast	0.0005355378
age:regionsouthwest	0.0004211812

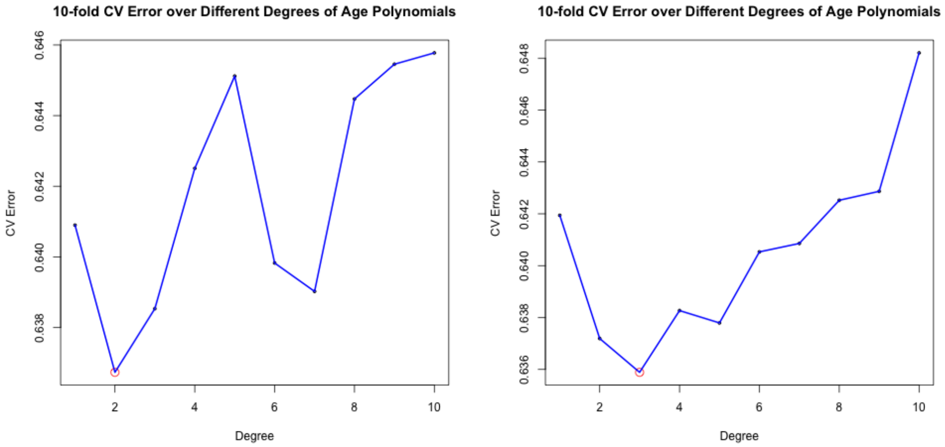

Figure 9: Choosing Optimal Degree for Polynomials of Age with 10-fold CV

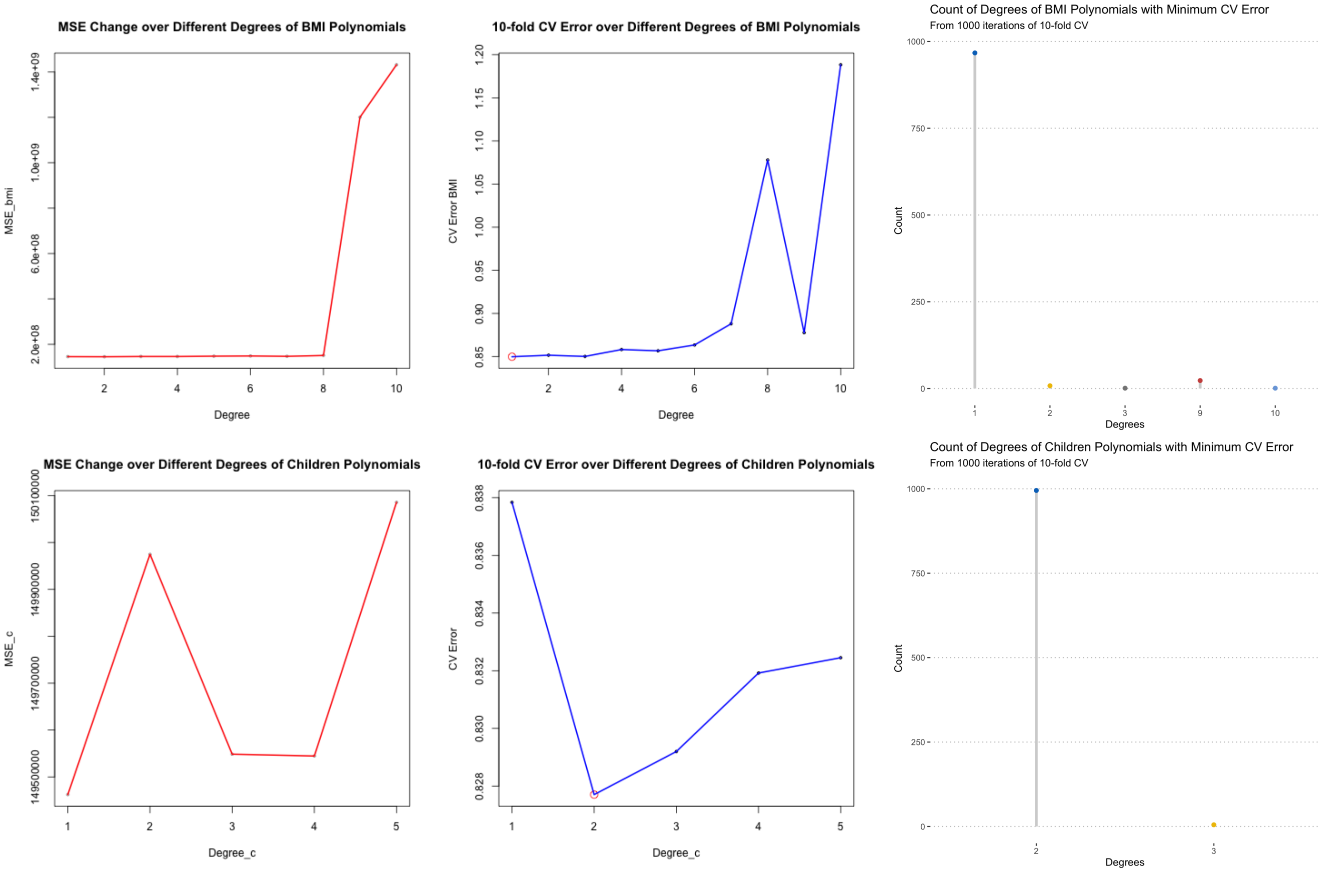

Figure 10: Choosing Optimal Degree for Polynomials of BMI and Number of Children

Table 15: Selected Polynomial

term	estimate	std.error	statistic	p.value	Signif. codes
(Intercept)	7.446	0.095	78.358	0.000	***
poly(age, 4)1	2.489	1.095	2.273	0.023	*
poly(age, 4)2	-1.151	0.406	-2.832	0.005	**
poly(age, 4)3	0.699	0.378	1.850	0.064	.
poly(age, 4)4	-0.902	0.381	-2.367	0.018	*
bmi	0.001	0.002	0.705	0.481	
children1	0.184	0.030	6.071	0.000	***
children2	0.334	0.034	9.771	0.000	***
children3	0.294	0.040	7.374	0.000	***
children4	0.561	0.083	6.784	0.000	***
children5	0.513	0.094	5.466	0.000	***
sexmale	-0.101	0.023	-4.349	0.000	***
smokeryes	1.389	0.145	9.557	0.000	***
regionnorthwest	-0.163	0.089	-1.838	0.066	.
regionsoutheast	-0.414	0.086	-4.812	0.000	***
regionsouthwest	-0.424	0.089	-4.774	0.000	***
sexmale:smokeryes	0.096	0.052	1.837	0.066	.
smokerno:age	0.033	0.002	17.517	0.000	***
bmi:smokeryes	0.050	0.004	12.123	0.000	***
children1:smokeryes	-0.252	0.067	-3.764	0.000	***
children2:smokeryes	-0.326	0.070	-4.638	0.000	***
children3:smokeryes	-0.305	0.081	-3.784	0.000	***
children4:smokeryes	-0.582	0.237	-2.458	0.014	*
children5:smokeryes	-0.297	0.393	-0.756	0.450	
regionnorthwest:age	0.002	0.002	1.132	0.258	
regionsoutheast:age	0.007	0.002	3.269	0.001	**
regionsouthwest:age	0.007	0.002	3.214	0.001	**

*Note: In the selected polynomial model summary table, term ‘smokeryes:age’ has been removed due to NA outputs. This is because of the fact that `age` is fitted linearly with the categorical factor `smoker`, and it functions as the baseline for the interaction when smoker is “yes” with age. This is normal as we are quantifying interactions between categorical and numeric variables, but removed for clarity of the report.

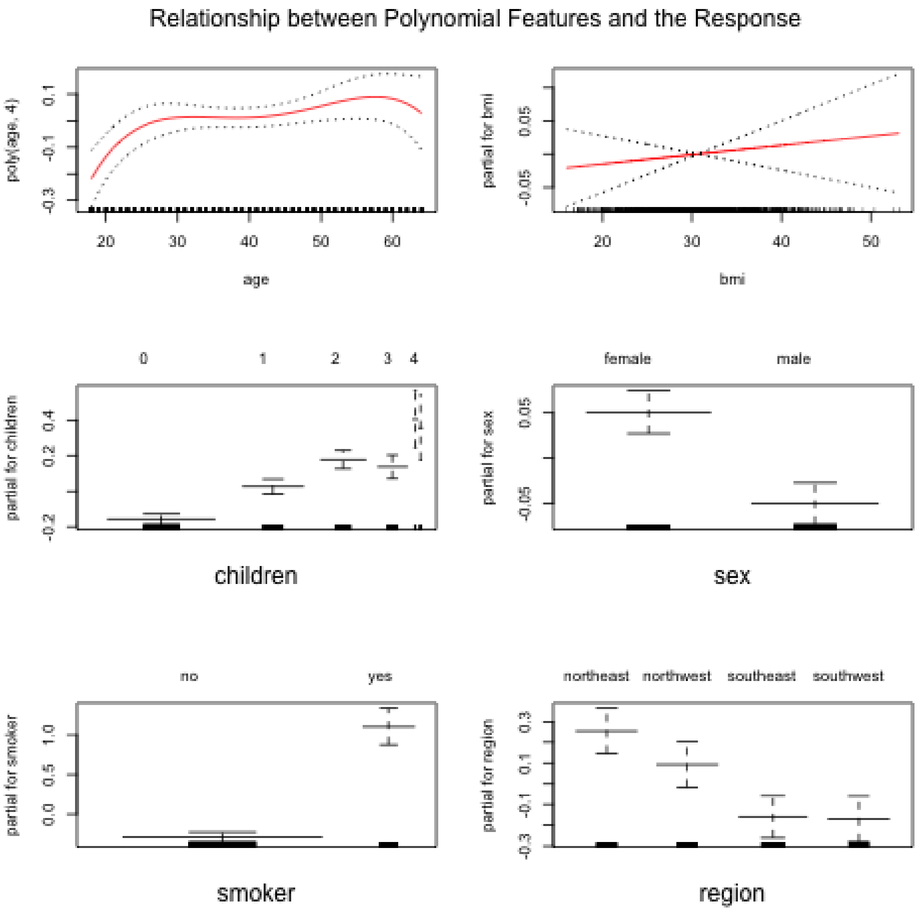

Figure 11: Relationship between Polynomial Features and the Response

*Note: For the selected polynomial model, Figure 9 shows the evaluation of the relationship between individual predictors and the response. This plot is different from the GAM generated plot only on variable BMI, as polynomial treats BMI linearly with medical charges. We can tell that without a natural spline fit, the visualization of BMI falls into the Runge's theorem of not being able to extrapolate on the boundary and thus, not better than a generalizable model with natural spline fit.

Table 16: Selected GAM

term	estimate	std.error	statistic	p.value
(Intercept)	8.028	0.169	47.551	0.000
smokeryes	1.375	0.145	9.488	0.000
ns(age, df = 5)1	0.194	0.068	2.866	0.004
ns(age, df = 5)2	0.215	0.082	2.608	0.009
ns(age, df = 5)3	0.230	0.090	2.571	0.010
ns(age, df = 5)4	0.481	0.132	3.653	0.000
ns(age, df = 5)5	0.089	0.097	0.912	0.362
regionnorthwest	-0.165	0.088	-1.865	0.062
regionsoutheast	-0.421	0.086	-4.873	0.000
regionsouthwest	-0.434	0.089	-4.885	0.000
ns(bmi, df = 14)1	0.530	0.136	3.907	0.000
ns(bmi, df = 14)2	0.434	0.174	2.494	0.013
ns(bmi, df = 14)3	0.671	0.162	4.139	0.000
ns(bmi, df = 14)4	0.678	0.171	3.966	0.000
ns(bmi, df = 14)5	0.590	0.170	3.475	0.001
ns(bmi, df = 14)6	1.008	0.171	5.890	0.000
ns(bmi, df = 14)7	0.768	0.172	4.472	0.000
ns(bmi, df = 14)8	1.059	0.172	6.171	0.000
ns(bmi, df = 14)9	0.956	0.173	5.512	0.000
ns(bmi, df = 14)10	1.113	0.170	6.541	0.000
ns(bmi, df = 14)11	1.213	0.169	7.170	0.000
ns(bmi, df = 14)12	1.319	0.157	8.377	0.000
ns(bmi, df = 14)13	1.668	0.358	4.657	0.000
ns(bmi, df = 14)14	1.510	0.232	6.516	0.000
children1	0.187	0.030	6.176	0.000
children2	0.330	0.034	9.629	0.000
children3	0.284	0.040	7.112	0.000
children4	0.567	0.083	6.844	0.000
children5	0.529	0.094	5.619	0.000
sexmale	-0.097	0.023	-4.194	0.000
smokeryes:sexmale	0.087	0.052	1.667	0.096
smokerno:age	0.033	0.002	17.747	0.000
smokeryes:age	NA	NA	NA	NA
smokerno:bmi	-0.051	0.004	-12.367	0.000
smokeryes:bmi	NA	NA	NA	NA
smokeryes:children1	-0.246	0.067	-3.677	0.000
smokeryes:children2	-0.308	0.071	-4.356	0.000
smokeryes:children3	-0.294	0.081	-3.636	0.000
smokeryes:children4	-0.551	0.237	-2.329	0.020
smokeryes:children5	-0.222	0.395	-0.562	0.574
regionnorthwest:age	0.002	0.002	1.108	0.268
regionsoutheast:age	0.007	0.002	3.413	0.001
regionsouthwest:age	0.007	0.002	3.290	0.001

*Note: In the selected GAM summary table, term ‘smokeryes:age’ and ‘smokeryes:bmi’ yield NA outputs. This is because of the same reason in the selected polynomial model output, but leave them here for comparison and clarity.