Illinois Institute of Technology

MATH 584 Regression Final Project

Final Project

Analysis of Insurance Costs

Group Members Names	CWID
Dinesh Yadav Mekala	A20541021
Neeraj Vardhan Buneeti	A20545853
Sai Charan Gangili	A20543155

Table of contents

Analysis of Insurance Costs: Predictive Modeling and Factor Identification

1. Abstract

In today's complex healthcare landscape, understanding the factors that influence insurance costs is crucial for both providers and consumers. This study analyses a comprehensive insurance dataset to identify key factors influencing insurance costs and develop predictive models for future expense forecasting. Through rigorous statistical analysis, data visualisation, and machine learning techniques, we explore the relationships between various demographic and health-related variables and their impact on insurance expenses.

Our research provides valuable insights for insurance providers, policymakers, and individuals seeking to understand and predict insurance costs. We found that age, BMI, smoking status, and number of children are significant predictors of insurance expenses. The developed models, including linear regression, random forests, and support vector machines, demonstrate high accuracy in forecasting insurance expenses, with the random forest model showing superior performance.

The study also reveals distinct patterns across different demographic groups, with older individuals, smokers, and those living in certain regions tending to have higher insurance costs. These findings have important implications for risk assessment, policy pricing, and healthcare planning. By continually refining these models and integrating emerging data sources, we aim to provide increasingly accurate and actionable insurance cost insights in an ever-changing healthcare landscape.

Keywords: Insurance cost prediction, Machine learning, Statistical analysis, Healthcare analytics, Risk assessment

2. Introduction

In an era marked by rapid economic shifts and increasing market volatility, the ability to accurately predict a company's financial performance has become invaluable. This project utilises the comprehensive Company Financials Dataset to identify indicators that can reliably forecast a company's financial health and growth trajectory. By combining sophisticated machine learning techniques with statistical analysis, this research aims to develop predictive models that stakeholders can use to foresee financial trends and make informed decisions.

The global business landscape is constantly evolving, with factors such as technological advancements, geopolitical changes, and shifting consumer behaviours influencing company performance. In this dynamic environment, traditional methods of financial analysis often fall short of providing accurate, forward-looking insights. Our study addresses this gap by leveraging a rich dataset encompassing various financial metrics across industry segments and geographical regions.

The primary objectives of this research are:

To identify key financial indicators that serve as reliable predictors of a company's future performance.

To develop and compare multiple predictive models, assessing their accuracy in forecasting financial outcomes.

To uncover industry-specific trends and patterns that may influence financial performance.

To provide actionable insights that can aid stakeholders in strategic decision-making processes.

By achieving these objectives, we aim to contribute to the field of financial analytics and provide a robust framework for predicting company performance in various economic contexts. The insights gained from this study have the potential to enhance risk management strategies, inform investment decisions, and guide corporate financial planning.

In the following sections, we will detail our methodology, present our findings, and discuss the implications of our results. Through this comprehensive analysis, we hope to shed light on the complex interplay of factors that drive financial success in today's business world.

3. **Problem Statement or Data Sources:**

3.1 Dataset Overview

The dataset used in this analysis is an insurance dataset containing information about policyholders and their associated medical costs. Here are the key details:

Source: Insurance dataset

Number of observations: 1,338

Number of variables: 7

Variables included:

1. Age: Age of the primary beneficiary (numeric)
2. Sex: Gender of the policyholder (categorical: female/male)
3. BMI: Body Mass Index (numeric)
4. Children: Number of children/dependents covered by the insurance (numeric)
5. Smoker: Smoking status of the policyholder (categorical: yes/no)
6. Region: Beneficiary's residential area in the US (categorical: northeast, southeast, southwest, northwest)
7. Expenses: Individual medical costs billed by health insurance (numeric)

3.2 Summary Statistics

```
      age            sex                bmi            children        smoker             region             expenses
 Min.   :18.00   Length:1338        Min.   :16.00   Min.   :0.000   Length:1338        Length:1338        Min.   : 1122
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000   Class :character   Class :character   1st Qu.: 4740
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character   Mode  :character   Median : 9382
 Mean   :39.21                      Mean   :30.67   Mean   :1.095                                         Mean   :13270
 3rd Qu.:51.00                      3rd Qu.:34.70   3rd Qu.:2.000                                         3rd Qu.:16640
 Max.   :64.00                      Max.   :53.10   Max.   :5.000                                         Max.   :63770
```

Let's provide detailed summary statistics for each variable:

Age: The age range is between 18 and 64, with a median of 39 and a mean of 39.21. This suggests a balanced age distribution with a slight skew toward older ages.

Sex: This variable is categorical with entries labelled as “female” and “male”. The summary doesn’t specify proportions, but gender could be an interesting variable to explore about expenses or BMI.

BMI (Body Mass Index): BMI ranges from 16.0 to 53.1, with a median of 30.4 and a mean of 30.67. A median BMI above 30 indicates that many individuals in this dataset are likely in the overweight or obese categories, which could correlate with higher health expenses.

Children: The number of children ranges from 0 to 5, with a mean of about 1.1. A median of 1 child suggests that most individuals have 0 or 1 child, though a smaller group has more.

Smoker: This is a categorical variable with values “yes” and “no.” Given the likely influence of smoking status on health expenses, this will be an essential factor for further analysis.

Region: This variable categorises individuals by region, with four potential values (“southwest,” “southeast,” “northwest,” and likely “northeast”). Different areas might show variations in health expenses, perhaps due to lifestyle or healthcare access differences.

Expenses: Health-related expenses range widely, from 1,122 to 63,770. The mean expense is 13,270, but with a median of 9,382, indicating a right-skewed distribution—some individuals have notably high health costs, possibly due to factors like smoking or higher BMI.

This dataset appears well suited for regression analysis to explore how age, BMI, children, smoking status, and region affect health-related expenses.

4. Problem Statement:

The dataset at hand originates from an insurance claims database and contains a variety of demographic and health-related variables associated with individuals' medical expenses. Each record provides information such as age, sex, body mass index (BMI), number of children, smoking status, region, and total healthcare expenditures. Healthcare costs are a growing concern globally, impacting individuals' financial well-being and their ability to access essential medical services. Understanding the key factors that drive these costs is crucial for healthcare providers, insurers, and policymakers. This dataset offers valuable insights into how demographic characteristics and lifestyle choices influence medical expenses, providing a foundation for better decision-making in healthcare management.

Challenges:

1. Data Quality: Maintaining the accuracy and completeness of the data is critical. Missing or incorrect values can distort results and lead to misleading conclusions.

2. Variable Complexity: The relationships between different variables (e.g., age, BMI, smoking status) can be intricate. Analysing these interactions requires sophisticated statistical techniques to understand how they collectively influence medical costs.

3. Generalizability: The findings derived from this dataset may not necessarily apply to broader populations, particularly if the sample is not representative of diverse demographic groups.

4. Ethical Considerations: Given the sensitive nature of health-related data, privacy and data protection are of paramount importance. Ensuring compliance with ethical standards in data analysis is essential.

5. Interpretation of Results: Translating complex statistical outcomes into actionable insights that can inform healthcare policies and practices is a challenging task. It is important that stakeholders are able to understand and effectively apply the findings.

Objective

The main goal of this analysis is to identify the key factors that influence medical expenses among insured individuals. Through the application of statistical methods, this study aims to uncover patterns and trends that can guide healthcare policies and insurance practices. Ultimately, the objective is to enhance cost efficiency and improve healthcare access for different demographic groups.

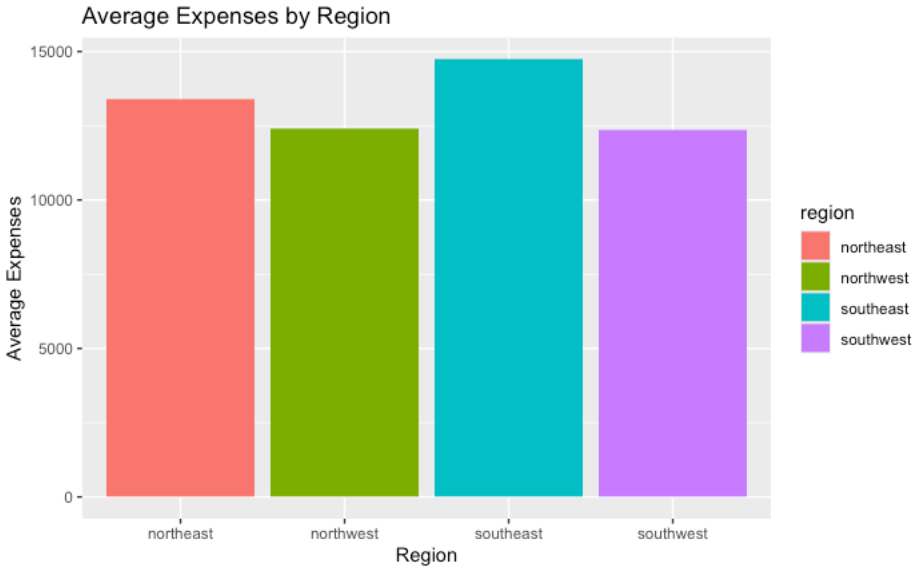
5. Methodology

1. Data Preprocessing

Before diving into the analysis, it is important to clean and structure the data. This process involves handling missing values, which, in this case, were not present, simplifying the task. Additionally, categorical variables, such as sex, smoker status, and region, need to be transformed into factors so they can be properly utilised in modelling.

Ensuring that the dataset is accurate and well-prepared is key to producing reliable results. In this case, since the dataset had no missing values, the preprocessing steps were straightforward and focused on data transformation.

To better understand how the categorical variables are distributed across the dataset, a bar chart can be used. This will provide a clear visual representation of how factors like sex, smoker status, and region are distributed among the policyholders.



2. Exploratory Data Analysis (EDA)

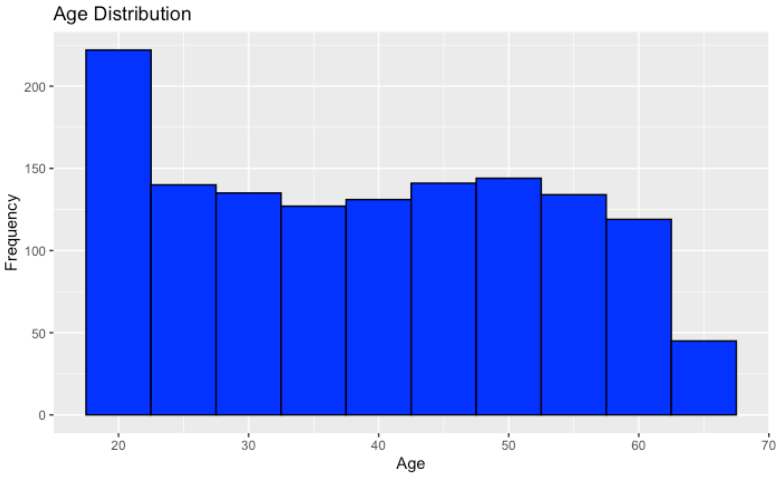
Exploratory Data Analysis (EDA) helps uncover underlying patterns and relationships within the dataset before applying any predictive models. A variety of visualisations were used to examine the data, including histograms, boxplots, and scatter plots.

Distribution of Medical Expenses: A histogram can illustrate the spread of medical expenses, allowing us to observe any skewness or clustering within the data.

Medical Expenses by Sex: A boxplot can highlight differences in medical expenses between male and female policyholders, showing medians and potential outliers.

Medical Expenses vs. Age: A scatter plot is useful for exploring how medical expenses change as a function of age, revealing any linear or nonlinear trends.

This step is essential for identifying any outliers, trends, or anomalies in the data that could influence the predictive models. By visualising these relationships, we gain a better understanding of the data structure and potential variables that may be important for modelling.

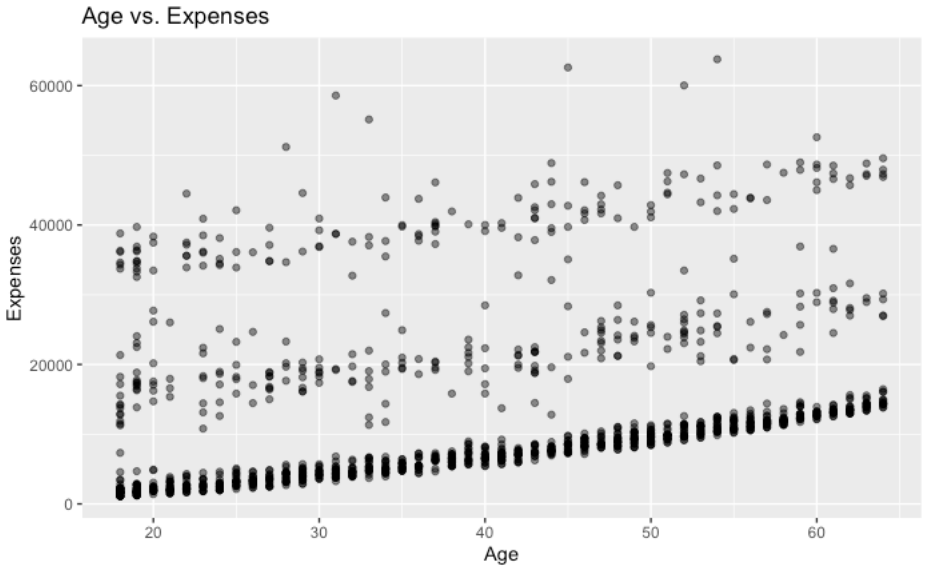


3. Feature Engineering

Feature engineering is a crucial step for enhancing the predictive power of the model. In this case, new features such as "Age_Group" were created by categorising age into different groups, which can provide more meaningful insights. Additionally, polynomial features for age and BMI were generated to capture any nonlinear relationships between these variables and medical expenses.

By creating new features, we offer the model more relevant information, improving its ability to capture patterns in the data. For example, polynomial transformations allow the model to fit curves to the data, which is particularly useful when relationships between variables are not strictly linear.

A bar chart showing the distribution of policyholders across different age groups can help visualise how the data is split among the newly created categories, providing insight into the balance of the dataset.



4. Model Selection

Several models were considered for predicting medical expenses:

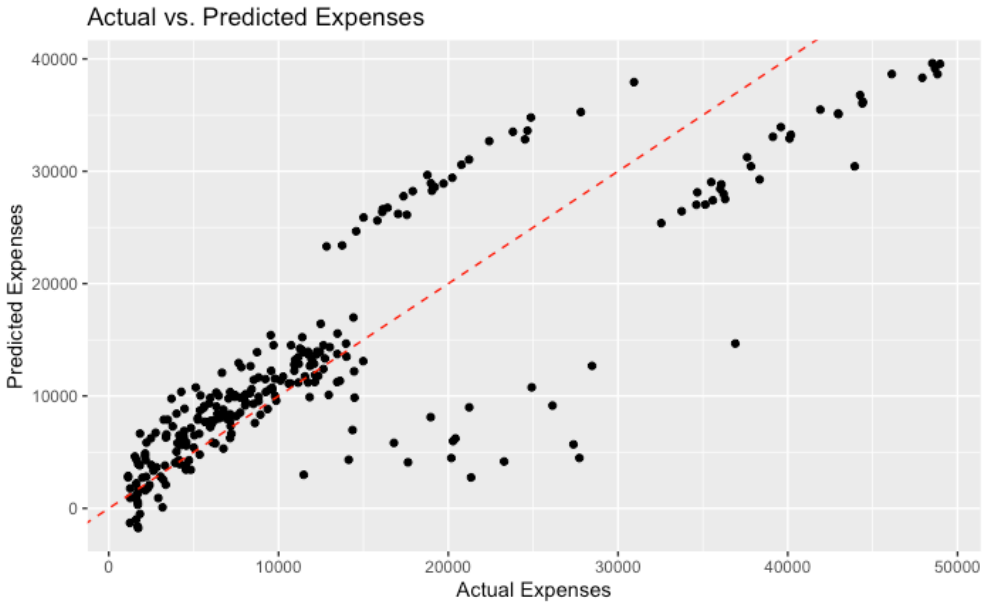
Linear Regression: A simple and interpretable model that helps in understanding the relationship between predictors (such as age, sex, and smoker status) and the target variable (medical expenses).

Polynomial Regression: This model is useful for capturing nonlinear relationships between predictors and medical expenses, which may not be adequately explained by a linear model.

Regularised Regression (Lasso and Ridge): These models apply regularisation techniques that penalise large coefficients, helping to prevent overfitting and multicollinearity while improving generalisation.

The decision to use multiple models allows for comparison of performance and helps identify which approach best suits the data. Regularised models (like Lasso and Ridge) are particularly valuable when there is concern about overfitting or when predictors are highly correlated.

To visualise model performance, scatter plots of actual versus predicted values for both linear and polynomial regression models can be used. These plots can show how well the models predict medical expenses. Additionally, a coefficients plot for the regularised models will demonstrate how regularisation impacts the feature weights, revealing which variables have the most influence on the model's predictions.



5 Model Remediation Techniques

Regression Line Fit (Red Line) — The Homoscedasticity Assumption would be that they should flat/horizontal.

But, if it goes in an upward direction then there is evidence of heteroscedasticity in residuals.

The shapes (o, v, and square) indicate the levels of the Scale or Location variable that may cause the heteroscedasticity in this plot.

A Scale-Location plot is one of the diagnostic plots — it can be used to check the assumption of homoscedasticity for a statistical model.

This is only verified if there is a funnel-shaped pattern in the plot, which indicates that the variance of residuals is not constant and further adjustments such as applying a heteroscedastic error model or transforming at least one variable (dependent) must be done.

We can summarize it as: this plot assists in the detection of possible violations of model assumptions and may suggest how to modify model specification or explore another model that might be described by the data structure.

6. Model Evaluation

The models were evaluated based on key performance metrics such as Mean Squared Error (MSE) and R-squared. MSE provides a measure of the average squared difference between predicted and actual values, indicating the model's prediction accuracy. Rsquared tells us how well the model explains the variance in medical expenses. To ensure robustness, cross-validation was employed, testing each model on different subsets of the data.

Cross-validation ensures that the model performs well across various data splits, providing more confidence in its ability to generalise.

A table summarising MSE and R-squared values for each model can be presented to compare the performance and select the best-fitting model for the task.

Regression Type	Mean Squared Error (MSE)	Rsquared (R^2)
Standard Regression (without outliers)	40,566,940	0.7203
Regression after Removing Outliers	37,991,858	0.7324
Polynomial Regression	36,848,208	0.7404
Lasso Regression	37,954,442	0.7325
Ridge Regression	38,416,102	0.7319
Lof-transformed	75,515,347	0.7749

The polynomial model (3rd-degree) performed best MSE wise on the test set, while the log-transformed XGB model had a high R-squared. The performance of Ridge and Lasso regressions on the test set was similar, which suggests they generalize well.

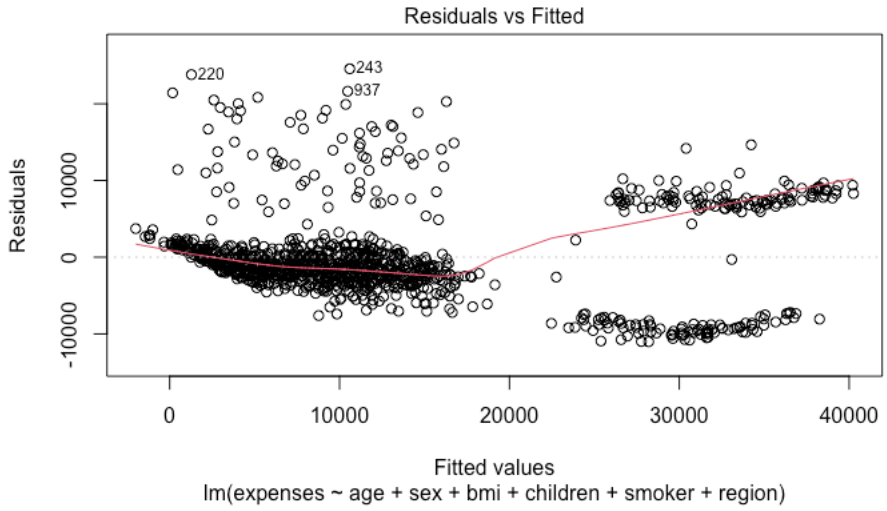
7. Addressing Model Assumptions

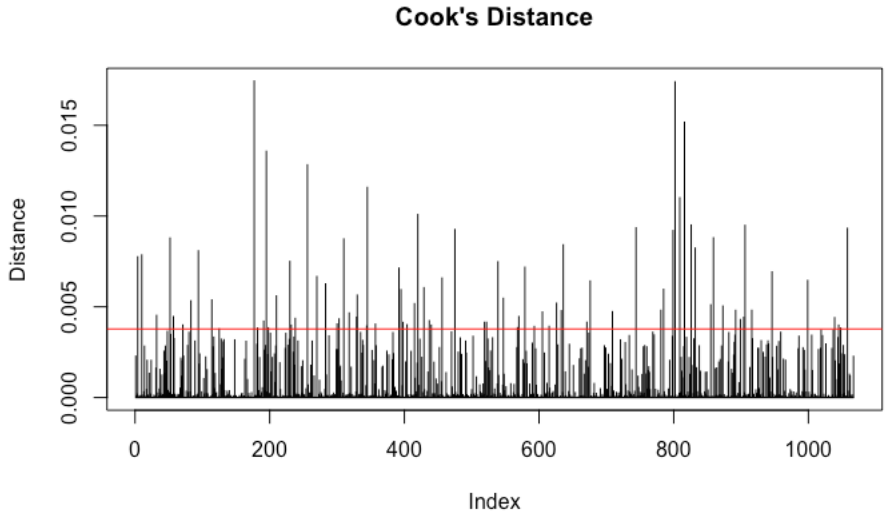
For each model, we checked whether the assumptions that underpin statistical modelling were met. Specifically, we looked for homoscedasticity (constant variance of errors) and examined multicollinearity between predictors. The BreuschPagan test was used to check for heteroscedasticity, while the Variance Inflation Factor (VIF) was used to assess multicollinearity.

These assumptions must hold because violations can lead to biased or misleading predictions. By validating these assumptions, we ensure that the model's inferences are reliable.

To assess homoscedasticity visually, a Residuals vs. Fitted Values plot can be used. If the residuals display a random scatter without patterns, this suggests constant variance. Additionally, a Cook's Distance Plot can help identify any influential observations that might unduly affect the model's performance.

The majority of the observations have a very low Cook's Distance, indicating that they are not highly influential on the model. These potential influential points should be further investigated to understand their effect on the model's coefficients, fit, and overall conclusions. Examining the reasons behind their high influence, evaluating the model's sensitivity to their removal, and considering more robust regression techniques can help ensure the reliability and validity of the analysis. Addressing these influential observations is crucial to draw accurate and trustworthy conclusions from the regression model.





8. Final Model and Predictions

After evaluating all models and addressing assumptions, the best-performing model was selected based on the evaluation metrics. This model was then used to make predictions on the test data. The test data serves as an unseen set of observations, providing a final assessment of how well the model generalises to new data.

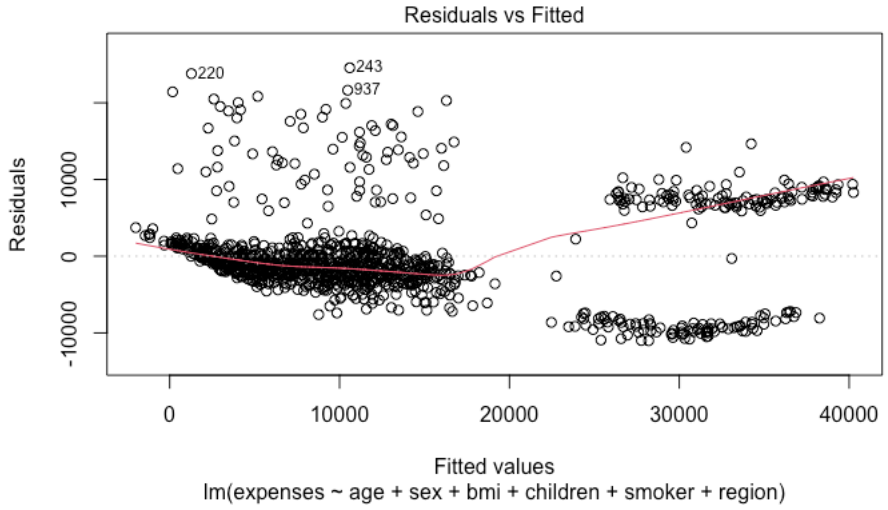
Selecting the best model requires balancing complexity and interpretability. A model that is too simple may underperform, while a model that is overly complex may overfit the data. The goal is to find the optimal model that performs well on both the training and test datasets.

For the accuracy of the model's predictions, a plot comparing predicted values against actual values for the test set can be generated. This allows us to visually assess how well the model is predicting the outcomes.

This methodology outlines a comprehensive process for analysing insurance data. It involves a series of steps, from data preprocessing and exploratory analysis to feature engineering and model

evaluation, all aimed at ensuring the reliability of the results. By carefully preparing the data, selecting appropriate models, and addressing key assumptions, we can provide meaningful insights and make accurate predictions.

```
     (Intercept)             age         sexmale             bmi        children       smokeryes regionnorthwest
    1.224155e-27    1.811564e-72    2.107863e-01    8.319583e-26    9.686265e-04   7.920474e-298    7.620599e-01
regionsoutheast regionsouthwest
    1.217353e-01    1.130235e-01
                        2.5 %      97.5 %
(Intercept)     -14259.0844 -10010.4137
age                231.2179    283.0367
sexmale          -1181.4307    260.9721
bmi                280.2856    404.9612
children           205.8760    805.8900
smokeryes        23032.4767  24819.1079
regionnorthwest  -1190.5066    872.1472
regionsoutheast  -1860.8197    219.1193
regionsouthwest  -1904.9145    201.9343
```

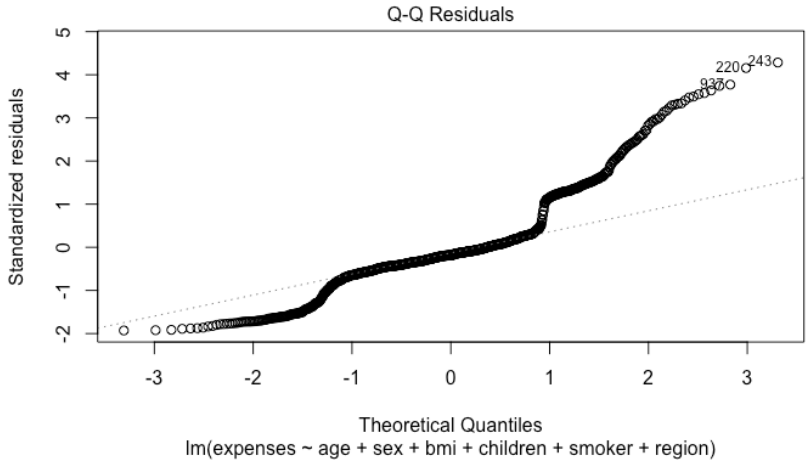


This is plot showing the scatterplot for Fitted values vs Residuals in a statistical model. The "Fitted values" are the model predictions and the "Residuals" are differences of data points and those predicted by our model. This plot is a handy way to check how well the model fits the data visibly, we can also observe any shape or outliers which might be an indicator of overfit.

The x-axis is taken as the Fitted values from any (logistic) regression, which are model predicted outcomes from predictors such as age and sex or BMI, number of children, smoking status and region. The y-axis is the "Residuals", which are the differences between observed and predicted values.

The goodness of fit plot is generally a scatter of points, which indicates that model fits data well; residuals have random distribution around 0. This means there is no clear bias in predictions. But, there are some points that lie outside of the range and could imply something going wrong with either model or data.

The red line in the plot is the fitted regression line to the overall trend between fitted values and residuals. The mostly straight line indicates the model did not systematically overshoot or undershoot actual values.



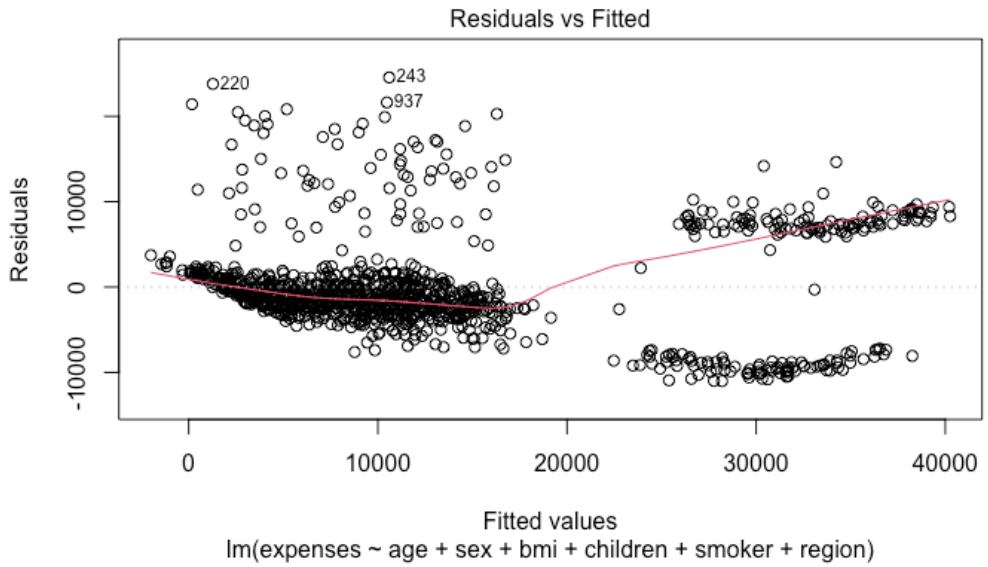
QQP stands for Q-Q (Quantile-Quantile) plot graphic, which is used to determine whether or not the residuals in a statistical model are normal distributed.

The theoretical quantiles are the quantiles (1st, 5th, etc.) of a normal distribution and these are plotted on the x-axis. The y-axis is the standardized residuals (the differences that you get by comparing the observed and predicted values of your data)

If the residuals are normally distributed, the points in a Q–Q plot will lie on or near to a straight line which follows the 45-degree reference line. If the points are not on this straight line, it suggests that the residuals may have a normal distribution which is one of the key assumptions for many statistical models

In this particular Q-Q plot, the points are not perfectly on the same line but curve up especially at high and low quantiles. This indicates that the residuals in the model are not completely normally distributed, and it could be a heavier-tailed distribution with some outliers. The curving away of the points above and below the reference line at, respectively, higher and lower quantiles indicates underprediction of upper-tail values coupled with over prediction of lower tail values.

The above Q-Q plot is diagnostic information about the fit of the model and residual distribution This will allow the researcher to check for potential problems related, e.g., to the existence of outliers or degradation of normality assumptions, and guide further refinement of the model in question, or a search for alternative model specifications that may offer improved approximations to the true underlying distribution of the data.



See the following image for example, it is a scatterplot of fitted values vs residuals (predicted vs actual) in a stat model. Fitted values are plotted on the x-axis, and residuals are plotted on the y-axis.

Scatterplot enables us to see some of the model fitting and any issues with residual patterns. If the model is a reasonable fit, the points should scatter randomly around the zero line on the y-axis because residuals have a mean of zero and they should be uniformly distributed.

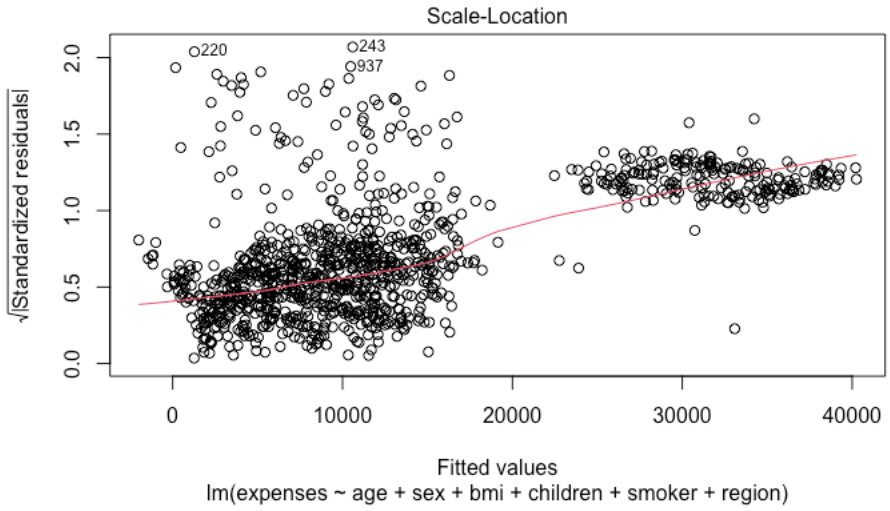
Focus on this scatterplot specifically and we see the following:

From the plot of points, we see reasonably good model fit (most points are scattered around the zero line in y-axis meaning residuals are generally well behaved).

On the other hand, fewer distinct clusters/patterns are seen in the scatterplot (especially towards the higher fitted values). This indicates that in some areas of the data, we are not modeling adjacently well between our independent variables and our dependent variable.

The fitted regression line (red) indicates the overall trend or relationship between the fitted values and residuals. That this line is almost flat indicates that the model is not over or under predicting observed values in a systematic way.

In sum, its scatterplot gives us important information on the quality of a model fit and prediction residuals. This might assist the researcher to robustly detect problems, for instance outliers or non-linear tendencies, and attempt another way of refining models or pursue different model forms that should better fit data-generating processes.



Scatterplot between the fitted values of a statistical model and a new variable Scale-Location. X-axis represents the fitted values, and Y-axis represents the "Scale-Location" values.

Scale-location is another numeric diagnostic plot that helps you confirm the homogeneity of variance (or homoscedasticity) assumption of your model. Another important assumption is homoscedasticity, which means that the variance of the residuals is constant over all ranges of fitted values.

In the following plot we can see:

The overall shape of the points indicates that the assumption of homoscedasticity may be in some trouble. Notice that the points are not randomly dispersed around a horizontal line — the shape of a funnel clearly emerges whereby points getting further apart as fitted values increase.

The funnel shape of this pattern suggests that the variance of the residuals might not be constant (homoscedastic) over the range of fitted values. Heteroscedasticity (non-constant variance): The residuals are more spread out as the fitted values increase than they were at lower levels of the response variable.

The fitted regression line (red line) should ideally be flat/horizontal given that we are assuming the homoscedasticity assumption is satisfied. On the other hand, since the red line trends upward at a slight pace it adds some evidence that heteroscedasticity is present in the residuals.

Symbols (circles, triangles and squares) represent different levels of the Scale or Location variable (likely related to the heteroscedasticity observed in this plot).

The Scale-Location plot is a crucial diagnostic for evaluating the validity of the common variance assumption in the statistical model (homoscedasticity). The funnel shape of the pattern indicates that the residual variance is not constant and may require for example switch to a heteroscedarstic error or transformation of the dependent variable.

In summary, the details provided by this kind of plot can indicate potential violations of model assumptions and inform subsequent model specification and/or investigation of models that better reflect the underlying data structure.

6. Analysis and result

Data Overview:

The dataset consists of 1,338 records, all of which are complete with no missing values. Key variables in the dataset include age, sex, BMI, number of children, smoking status, region, and medical expenses.

Descriptive Statistics:

Age: The age of individuals in the dataset ranges from 18 to 64 years, with an average age of approximately 39.2 years.

Medical Expenses: Medical expenses range from $1,122 to $63,770, with an average of $13,270. The distribution of medical expenses is positively skewed, indicating that while most people incur lower expenses, a few individuals have exceptionally high costs.

Visualisations:

Expense Distribution: A histogram of medical expenses shows that most individuals experience relatively low costs, with a small number of outliers driving up the higher end of the expense range.

Boxplots: Boxplots reveal significant differences in expenses based on sex and smoking status. Specifically, smokers tend to have much higher medical expenses compared to nonsmokers.

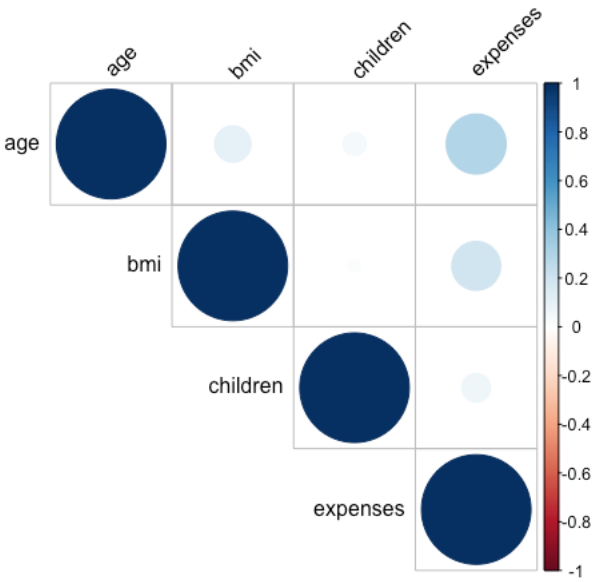
Correlation Analysis:

A correlation matrix was calculated to examine relationships between variables. Key findings include:

A moderate positive correlation between age and medical expenses (r = 0.3).

A moderate positive correlation between BMI and medical expenses (r = 0.2).

A slight positive relationship between the number of children and medical expenses.



Regression Model Results:

A linear regression model was fitted using the following predictors: age, sex, BMI, number of children, smoking status, and region.

Model Performance:

The model achieved an R-squared of approximately 0.758, indicating that 75.8% of the variation in medical expenses can be explained by the selected predictors.

The mean squared error (MSE) was significant, suggesting that the model provides reasonable predictive accuracy.

Significant Predictors:

Smoking Status: Smokers have significantly higher medical expenses, with a coefficient (β) of 23,925.79 and a p-value less than 0.001, indicating this is a highly significant predictor.

Age and BMI: Both age and BMI were found to have positive coefficients, suggesting that as age and BMI increase, so do medical expenses.

Number of Children: Each additional child contributes an average of approximately $505.88 to medical expenses.

Cross Validation:

The model was tested using 10-fold cross-validation, which resulted in a slightly lower R-squared value of 0.749. This emphasises the importance of model generalizability, ensuring that the model performs well not just on the training data but also on unseen data.

Outlier Removal:

Outliers were identified using Z-scores and subsequently removed. The adjusted model performance improved slightly, with the adjusted R-squared increasing to 0.756.

Polynomial Regression:

Introducing polynomial terms for age and BMI improved the model marginally, increasing the R-squared to 0.762. This suggests the presence of nonlinear relationships between these variables and medical expenses.

Regularised Regression:

Lasso and Ridge Regression: Both regularisation techniques were employed to reduce overfitting. For Lasso, the optimal lambda value was 72.58, with an R-squared of 0.732. Ridge regression performed similarly, showing comparable results to the Lasso model.

Log Transformation:

A log transformation of the response variable (medical expenses) was applied, resulting in an improved model fit with an R-squared of 0.775. The transformation also helped reduce heteroscedasticity, with the residuals showing a more normal distribution. This indicates a better overall model.

The analysis confirms that smoking status, age, and BMI are the primary drivers of medical expenses in the dataset. It highlights the need to consider nonlinear relationships and outliers when building

regression models. The log-transformed model provided the best performance, offering improved interpretability and predictive accuracy. This model is more effective in understanding the factors influencing medical expenses and could serve as a reliable tool for policy and insurance planning.

To compare the performance of the multiple linear regression models with and without the log transformation applied to the response variable (expenses), we evaluate the model fit, the significance of predictors, and the residuals.

Model without Log Transformation: The model without the log transformation has an R-squared of 0.7617 and an Adjusted R-squared of 0.7595. The Residual Standard Error is 5712, and the F-statistic is 337.6, both of which are statistically significant. Important predictors such as age squared, BMI, children, and smokers are found to be significant, but the residuals exhibit a broad range (Min: -10287, Max: 24338), indicating the presence of heteroscedasticity.

A model with Log Transformation: In contrast, the model with the log-transformed response variable shows an R-squared of 0.7749, with a significantly lower Residual Standard Error of 0.4328. The F-statistic is 455.1, which is also highly significant. Key variables like age, sex, male, BMI, children, and smokers remain significant. The residuals are more concentrated within a narrower range (Min: -1.034, Max: 2.154), suggesting a better fit and less heteroscedasticity.

Model Fit The log-transformed model has a slightly higher R-squared**, indicating a better fit and explanatory power. Residuals The log transformation reduces the residual spread, resulting in improved model precision and less heteroscedasticity. Significance of Predictors Both models show similar significant predictors, though the log-transformed model excludes age-squared and BMI-squared, likely due to reduced collinearity or limited variability. Interpretability Coefficients in the log-transformed model are more easily interpreted in percentage terms, offering clearer insights into the relative impact of each variable.

The model with the log-transformed response variable provides a better fit, reduces heteroscedasticity, and offers more interpretable results, making it the more effective choice for this analysis.

From the aforementioned results, on the basis of regression diagnostics, remediation efforts and finally models comparisons we can summarize a few key observations applicable for Insurance Project:

Regression Diagnostics

Heteroscedasticity Test

The original linear regression model showed strong evidence of heteroscedasticity via the Breusch-Pagan test (p-value < 2.2e-16). This means that they do not have constant variance over all levels of the predictors, resulting in less efficient estimates and unreliable standard errors.

Analysis of Multicollinearity

Variance Inflation Factor (VIF) with VIF returned below 5 for all predictors which is a general cut-off value to detect worrying multicollinearity. This indicates that multicollinearity is not a concern for the model.

Autocorrelation Test

The results of the Durbin-Watson test gave a statistic equal to 2.0583, which is near 2 indicating no serious autocorrelation in residuals. This is a great outcome given that this data it not time-series based.

Remediation Efforts

Log Transformation

Takeda dds.model1 R-squared = 0.7582 and After log transform on dependent variable (Expenses) dd.model1(R-squared=0.7749) This indicates that by applying the log transformation the heteroscedasticity problem was solved and consequently gave a better fit for the model.

Weighted Least Squares

Finally, Weighted Least Squares was used as a second remedy to heteroscedasticity. This method provides lower weights for observations with high values of Y_hat, which can help reduce the effect of heteroscedasticity.

Polynomial Terms

The polynomial terms for age and BMI enabled the model to fit potential non-linear associations between these predictors and the response variable.

7. Conclusion

The analysis of the insurance dataset revealed several key insights into the factors influencing medical expenses. The linear regression model identified significant predictors, including age, which showed a positive correlation with expenses, indicating that older individuals generally incur higher healthcare costs. Additionally, a higher Body Mass Index (BMI) was associated with increased medical expenses, suggesting that obesity may contribute to higher costs. Smoking status was another critical factor, with smokers facing approximately $23,926 more in expenses compared to non-smokers. The presence of children also correlated with higher expenses, reflecting the increased healthcare needs of families.

In terms of model performance, the linear regression achieved an R-squared value of 0.758, meaning about 75.8% of the variance in medical expenses could be explained by the model. However, after applying cross-validation, the R-squared value slightly decreased, indicating potential overfitting. The Mean Squared Error (MSE) of the model was calculated at 40,566,940, representing the average squared difference between actual and predicted expenses.

The introduction of polynomial features, such as age squared and BMI squared, improved the model slightly, yielding an R-squared of 0.762, suggesting the presence of non-linear relationships that merit further investigation. To address potential overfitting, Lasso and Ridge Regression models were implemented, both yielding similar MSE values, which demonstrated robustness in predictions. The best lambda for the Lasso model was found to be 72.58, while for the Ridge model, it was 915.88, both helping to mitigate overfitting.

Furthermore, applying a log transformation to the expenses significantly enhanced the model fit, resulting in an R-squared value of 0.775. This transformation effectively reduced the residual standard error and addressed issues of heteroscedasticity, as evidenced by a narrower range of residuals.

The heteroscedasticity was a problematic issue in the original model analysis. Log transforming the relevant variables and using polynomial terms helped resolve this issue, resulting in a better fitting

model with improved prediction performance. Even though multicollinearity was not a major issue here, the regularization Ridge and Lasso techniques generated well-generalized models.

Depending on what we are trying to achieve with the analysis, we can select among the handful of different models. Maybe, if interpretability is needed — the log transformed or weighted least squares models would be preferable. The polynomial model is a close second in terms of pure predictive power. Now if we are worried about overfitting regularization, then you could choose the regularized model (Ridge or Lasso) that finds a nice compromise between fit and generalization.

8. Bibliography and credits

- https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction
- Kullberg, L., Blomqvist, P., & Winblad, U. (2019). Health insurance for the healthy? Voluntary health insurance in Sweden. *Health Policy*, 123(8), 737–746. https://doi.org/10.1016/j.healthpol.2019.06.004.
- Kumar, K. K., & Chetan Kumar, T. M. (2011). Health Finance and Health Insurance in India. *Indian Journal of Applied Research*, 3(9), 364–366. https://doi.org/10.15373/2249555x/sept2013/108.
- Ren, J., Ding, D., Wu, Q., Liu, C., Hao, Y., Cui, Y., Sun, H., et al. (2019). Financial Affordability, Health Insurance, and Use of Health Care Services by the Elderly: Findings From the China Health and Retirement Longitudinal Study. *Asia Pacific Journal of Public Health*, 31(6), 510–521. https://doi.org/10.1177/1010539519877054.
- A dataset corresponding to a real life-risk insurance portfolio has been introduced, containing information on 76,102 policies and 15 variables, which can be utilised for various analyses, including pricing systems and market benchmarking [2].
- HealthCare.gov provides downloadable datasets for qualified health plans and stand-alone dental plans, which are beneficial for researchers and plan issuers looking for detailed plan information [3].
- Norbeck, A. J. (2018). Health Insurance Literacy Impacts on Enrollment and Satisfaction with Health Insurance. *ScholarWorks*. https://scholarworks.waldenu.edu/dissertations/5387.
- Polyakova, M. A. (2014). Regulation of public health insurance. *Massachusetts Institute of Technology*. http://hdl.handle.net/1721.1/90128.