ILLINOIS INSTITUTE OF TECHNOLOGY

MATH 564: REGRESSION PROJECT

MEDICAL COSTS ANALYSIS & PREDICTION THROUGH LINEAR REGRESSION

Name: Sai Nivas Rangaraju, Anmol Rao Karukonda
Aid: A20539492, A20554502
College: Illinois Institute of Technology
Date: Nov 14, 2024

Summary
Multiple Linear Regression model for Medical cost

1 Abstract

This study examines the factors influencing health insurance charges using data from 1,338 individuals. We analyzed various personal characteristics such as age, body mass index (BMI), number of children, and smoking status to understand their impact on insurance costs.

Our analysis revealed that age, BMI, number of children, and smoking status are significant predictors of insurance charges. Interestingly, gender and geographical region did not significantly affect insurance costs in our model.

Initially, our model explained about 75% of the variation in insurance charges. However, we discovered some statistical issues that required further refinement of our approach. After addressing these challenges, we developed a more sophisticated model that incorporated the interaction between BMI and smoking status.

This final model proved to be much more effective, explaining nearly 84% of the variation in insurance charges. A key finding was the substantial impact of smoking on insurance costs, with smokers paying significantly more than non-smokers. Moreover, we found that the effect of BMI on insurance charges is much more pronounced for smokers compared to non-smokers.

These findings highlight the complex relationship between personal characteristics and healthcare costs. Our research provides valuable insights for insurance companies developing pricing strategies and for policymakers seeking to understand and manage healthcare expenditures. It also underscores the potential financial benefits of lifestyle choices, particularly regarding smoking, on individual healthcare costs.

2 Introduction

In today's complex healthcare landscape, understanding the factors that influence insurance charges is crucial for both insurance providers and policyholders. This study aims to analyze and predict insurance charges using regression analysis, providing valuable insights into the intricate relationship between personal characteristics and healthcare costs.

Our analysis is based on a dataset of 1,338 individuals, containing information on age, sex, BMI (Body Mass Index), number of children, smoking status, region, and corresponding insurance charges. This diverse set of variables allows us to explore both demographic and lifestyle factors that may impact healthcare costs.

The primary objectives of this study are:

1. To identify the key predictors of insurance charges
2. To quantify the impact of each predictor on insurance costs
3. To develop a robust predictive model for insurance charges
4. To address any statistical issues that may affect the model's reliability

Our approach involves a step-by-step regression analysis, starting with a basic linear model and progressing to more sophisticated techniques. We begin by examining the relationship between all available predictors and insurance charges. Subsequently, we refine our model by addressing issues such as heteroscedasticity and potential interactions between variables.

Throughout this analysis, we face several challenges:

- Heteroscedasticity: Our initial models show signs of non-constant variance in the residuals, which violates one of the key assumptions of linear regression. This requires us to explore various transformation techniques and weighted least squares regression.

- Non-linear relationships: The complex nature of healthcare costs means that relationships between variables may not always be straightforward or linear. We need to consider potential interactions between variables, particularly between BMI and smoking status.
- Variable selection: With multiple predictors available, we must carefully consider which variables to include in our final model to balance predictive power with model simplicity.
- Interpretability: While our primary goal is to develop an accurate predictive model, we must also ensure that our findings are interpretable and actionable for stakeholders in the healthcare and insurance industries.

To address these challenges, we employ several problem-solving strategies:

1. Systematic model building: We start with a comprehensive model including all variables and progressively refine it based on statistical significance and diagnostic tests.
2. Transformation techniques: We explore various transformations, including logarithmic and square root transformations, to address issues of heteroscedasticity and non-linearity.
3. Weighted Least Squares (WLS) regression: We implement WLS as an alternative approach to handle heteroscedasticity when simple transformations prove insufficient.
4. Interaction analysis: We investigate potential interaction effects between variables, with a particular focus on the relationship between BMI and smoking status.
5. Diagnostic testing: We employ various diagnostic tools, including residual plots and the Breusch-Pagan test, to assess model assumptions and identify areas for improvement.

This report is structured to guide you through our analytical process. We start with an initial model fitting and interpretation, followed by a thorough examination of regression diagnostics. We then present our efforts to address the identified issues, culminating in a final model that provides the best fit for our data.

Key insights from our analysis include the significant impact of smoking status on insurance charges, the complex interaction between BMI and smoking, and the relative importance of age and number of children in predicting insurance costs.

By the end of this report, we aim to provide a clear understanding of the factors driving insurance charges and offer a reliable model for predicting these costs. This information can be valuable for insurance companies in developing pricing strategies, for policymakers in understanding healthcare cost drivers, and for individuals in making informed decisions about their health and insurance choices.

3 Data Description and Problem Statement

Our study aims to predict insurance charges and identify the key factors influencing these costs. This analysis is crucial for insurance companies, policymakers, and individuals alike, as it can inform pricing strategies, policy decisions, and personal health choices.

The dataset used in this analysis is sourced from kaggle: dataset. It contains information on 1,338 individuals and includes the following variables:

1. Age: A continuous variable representing the age of the primary beneficiary in years.
2. Sex: A categorical variable indicating the gender of the insurance contractor (male/female).
3. BMI (Body Mass Index): A continuous variable calculated as weight (kg) / (height (m))$\hat{2}$, providing a measure of body fat based on height and weight.
4. Children: An integer variable representing the number of children covered by the insurance plan.
5. Smoker: A categorical variable indicating whether the beneficiary is a smoker (yes/no).

6. Region: A categorical variable specifying the beneficiary's residential area in the US (northeast, southeast, southwest, northwest).
7. Charges: A continuous variable representing the individual medical costs billed by health insurance.

This dataset presents several challenges and opportunities for analysis:

- Diverse Variable Types: The dataset includes a mix of continuous, categorical, and discrete variables, requiring careful consideration in our modeling approach.
- Potential Non-Linear Relationships: Variables like age and BMI may have non-linear relationships with insurance charges, necessitating exploration of interaction terms and transformations.
- Heteroscedasticity: Initial analyses revealed issues with non-constant variance in the residuals, violating a key assumption of linear regression. This required us to explore various transformation techniques and weighted least squares regression.
- Skewed Distribution of Charges: The distribution of insurance charges is right-skewed, with the mean ($13,270) being significantly higher than the median ($9,382). This skewness posed challenges in our regression analysis and required consideration of data transformations.
- Regional Variations: The inclusion of the 'region' variable allows us to explore geographical differences in insurance charges, which could be influenced by factors such as local healthcare costs or regional health trends.
- Health Behavior Impact: The 'smoker' variable provides an opportunity to quantify the impact of a specific health behavior on insurance costs, potentially informing public health policies.

Our primary objectives in analyzing this dataset are:

1. To identify the most significant predictors of insurance charges.
2. To quantify the impact of each predictor, with a particular focus on modifiable factors like smoking status and BMI.
3. To develop a robust predictive model that can accurately estimate insurance charges based on an individual's characteristics.
4. To explore potential interaction effects between variables, particularly between BMI and smoking status.
5. To address statistical challenges such as heteroscedasticity and non-linear relationships in our modeling process.

By addressing these objectives, we aim to provide insights that can inform decision-making in the healthcare insurance industry, guide policy decisions aimed at managing healthcare costs, and help individuals understand the factors that influence their insurance premiums.

4 Proposed Methodology

Our approach to predicting insurance charges and identifying key influencing factors involved a systematic application of multiple regression techniques. We designed our methodology to address the complexities of the dataset and tackle potential statistical issues. Here's an overview of our approach:

Initial Linear Regression Model

We began with a comprehensive linear regression model that included all available predictors. This initial model allowed us to assess the significance of each predictor and identify potential issues in the data.

Model Refinement

Based on the results of the initial model, we refined our approach by:

- Removing non-significant predictors, particularly *sex* and *region*.
- Focusing on the most influential factors affecting insurance charges.

Diagnostic Testing

We conducted several diagnostic tests to ensure the validity of our model:

- Durbin-Watson test for autocorrelation.
- Visual inspection of residual plots.
- Breusch-Pagan test for heteroscedasticity.

These tests revealed issues with heteroscedasticity in our initial models, prompting further refinement.

Addressing Heteroscedasticity

To address heteroscedasticity, we explored several transformation techniques:

- Logarithmic Transformation
- Square Transformation
- Square Root Transformation
- Weighted Least Squares (WLS) Regression

Each transformation was evaluated using the Breusch-Pagan test to assess its effectiveness.

Multicollinearity Check

We used Variance Inflation Factor (VIF) analysis to check for multicollinearity, ensuring that our predictors were not overly correlated.

Interaction Analysis

To capture more complex relationships between variables, we explored potential interaction effects between all predictors. This comprehensive model identified significant interactions, particularly between *BMI* and *smoking status*.

Final Model

Based on the interaction analysis, our final model incorporated the significant interaction term between *BMI* and *smoking status*.

Model Validation

We validated our final model using:

1. Residual plots for visual inspection.
2. Breusch-Pagan test for heteroscedasticity.
3. R-squared and adjusted R-squared values for goodness of fit.

Conclusion: By following this systematic approach, we developed a robust model that predicts insurance charges with high accuracy. The final model incorporates significant interactions, providing insights into complex relationships while balancing predictive power and interpretability. This model can be a valuable tool for insurance pricing strategies and understanding the factors that drive healthcare costs.

5 Analysis and Results

Our systematic approach to analyzing the insurance charges dataset yielded several key findings and insights:

5.1 Initial Linear Regression Model

We began with a comprehensive linear regression model that included all available predictors. This initial analysis revealed that age, BMI, number of children, smoking status, and some regional differences were statistically significant predictors of insurance charges. Notably, sex was found to be a statistically insignificant predictor (p-value = 0.693).

5.2 Model Refinement

After removing the non-significant predictors (sex and region), our refined model explained approximately 74.97% of the variance in insurance charges ($R^2 = 0.7497$). This high R^2 value indicates that our selected predictors have strong explanatory power for insurance charges.

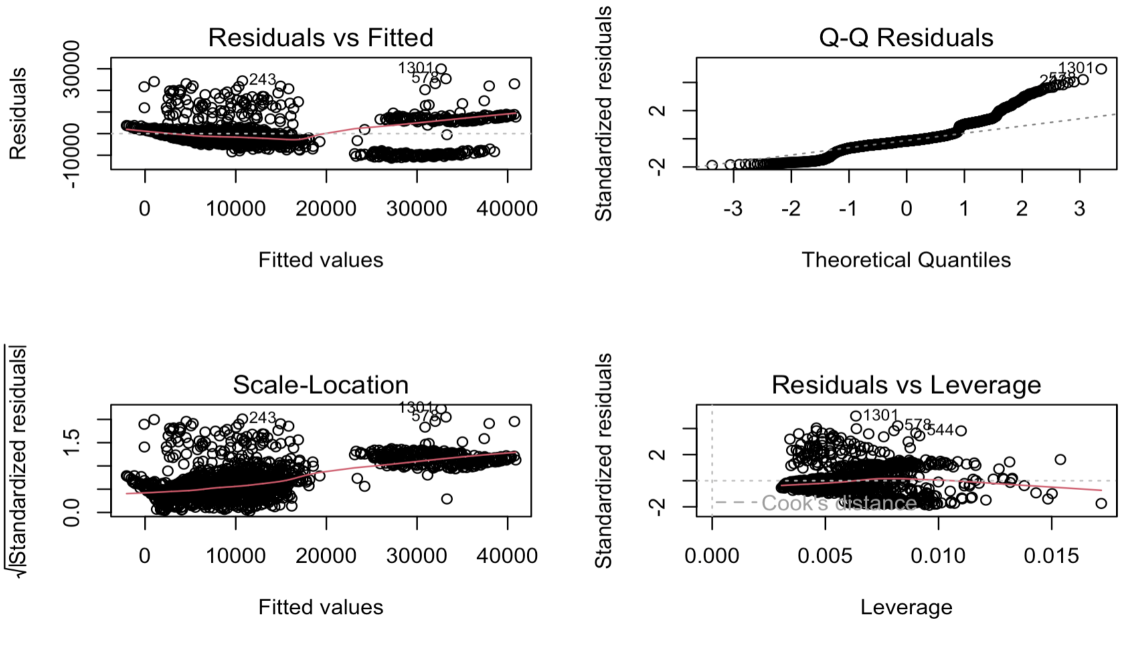


Figure 1: Initial Linear Regression's plot

5.3 Addressing Heteroscedasticity

The Breusch-Pagan test on our initial model revealed significant heteroscedasticity ($BP = 117.16$, p-value $< 2.2 \times 10^{-16}$). To address this issue, we explored several transformation techniques:

- **Logarithmic Transformation:** This improved the R^2 value to 0.7622 but did not fully resolve the heteroscedasticity issue ($BP = 80.917$, p-value $< 2.2 \times 10^{-16}$).
- **Square Transformation:** This approach also failed to adequately address the heteroscedasticity ($BP = 218.78$, p-value $< 2.2 \times 10^{-16}$).
- **Square Root Transformation:** While this showed some improvement, it still did not completely resolve the issue ($BP = 29.317$, p-value $= 6.741 \times 10^{-6}$).
- **Weighted Least Squares (WLS) Regression:** This method significantly improved the heteroscedasticity issue, as evidenced by the Breusch-Pagan test results ($BP = 1.2726 \times 10^{-5}$, p-value $= 1$).

The WLS model revealed that, on average, smokers pay approximately $22,725 more in insurance charges than non-smokers, holding all other factors constant.

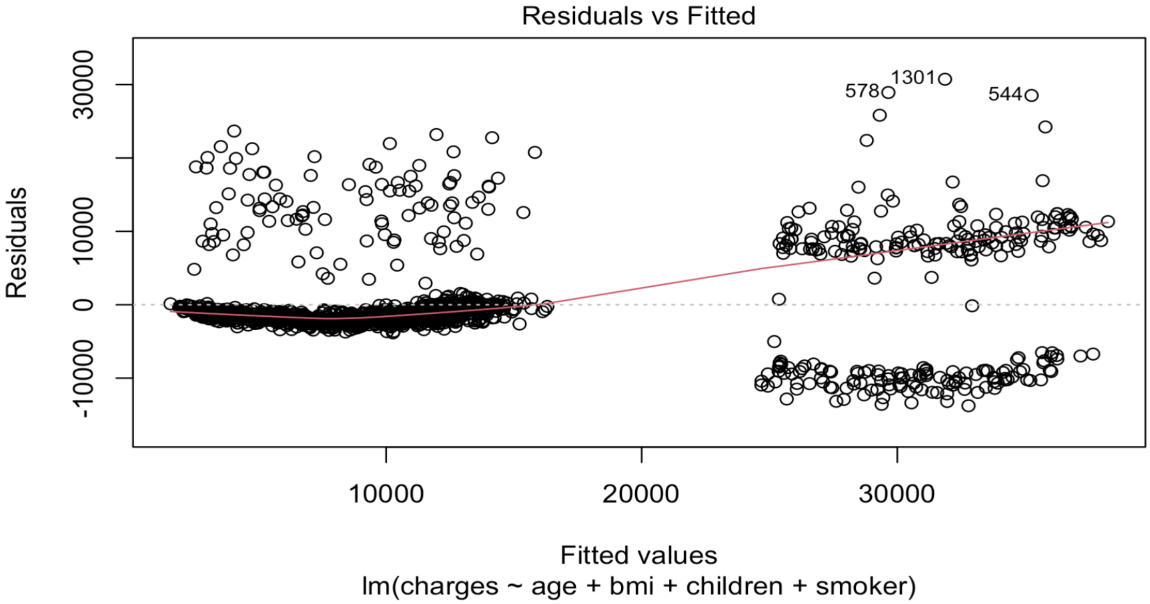


Figure 2: Weighted Least Squares Residual vs Fitted plot

5.4 Interaction Analysis

Further analysis revealed a significant interaction effect between BMI and smoking status. We incorporated this interaction into our final model, which resulted in a substantial improvement in predictive power.

5.5 Final Model

Our final model, incorporating the interaction between BMI and smoking status, explained 83.88% of the variance in insurance charges ($R^2 = 0.8388$). This represents a 28.5% improvement over the WLS model. Key findings from the final model include:

- **Age:** For each year increase in age, insurance charges increase by approximately $265, holding other factors constant.

- **Children:** Each additional child increases insurance charges by about \$509, all else being equal.
- **Smoking Status:** The effect of smoking on insurance charges is complex due to its interaction with BMI. For non-smokers, the base charge is \$2,729 less than for smokers.
- **BMI and Smoking Interaction:** For non-smokers, each unit increase in BMI is associated with a \$5.66 increase in charges. For smokers, each unit increase in BMI is associated with a much larger increase of \$1,439.44 (\$5.66 + \$1,433.78) in charges.

The final model equation is:

$$\text{charges} = -2729.002 + 264.948(\text{age}) + 5.656(\text{bmi}) + 508.924(\text{children}) - 20194.709(\text{smoker}) + 1433.788(\text{bmi} \times \text{smoker})$$

The Breusch-Pagan test on the final model ($BP = 7.0967$, p-value $= 0.2136$) indicates that we successfully addressed the heteroscedasticity issue, ensuring the reliability of our model's predictions.

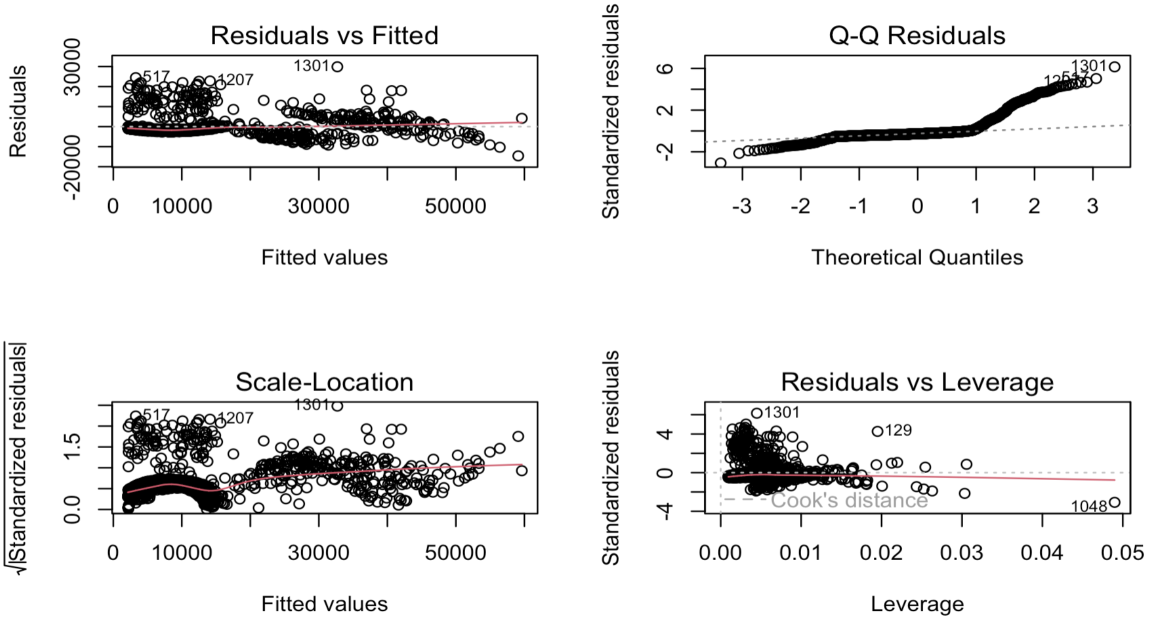
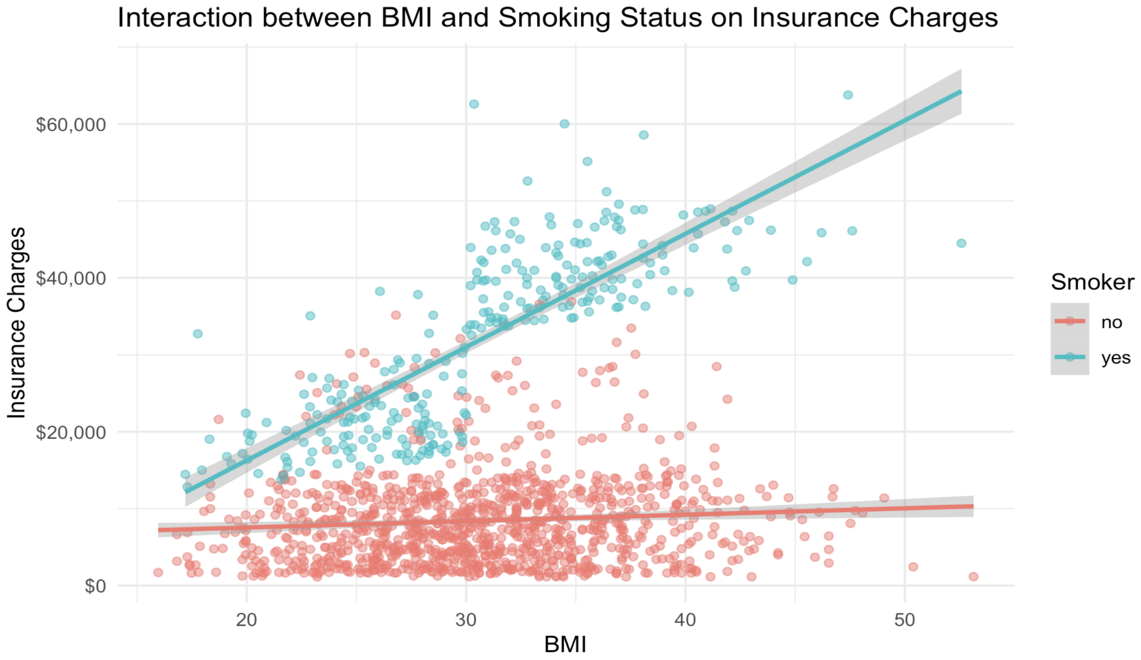


Figure 3: Interaction term model's diagnostic plots

These findings highlight the complex interplay between personal characteristics, particularly BMI and smoking status, in determining insurance charges. The substantial impact of smoking, especially in conjunction with higher BMI, underscores the importance of lifestyle factors in healthcare costs.



6 Conclusion

Our analysis of insurance charges using regression techniques has yielded several important insights:

6.1 Key Predictors

Age, BMI, number of children, and smoking status are significant predictors of insurance charges. Sex and region were found to be statistically insignificant in our analysis.

6.2 Model Performance

Our final model, which incorporates an interaction term between BMI and smoking status, explains 83.88% of the variance in insurance charges. This represents a substantial improvement over simpler models and underscores the complex relationship between these factors.

6.3 Impact of Smoking

Smoking status has the most dramatic effect on insurance charges. Our model indicates that smokers pay significantly more for insurance, with the effect amplified by BMI.

6.4 Age and Children

Both age and number of children are positively correlated with insurance charges:

- Each year of age adds approximately $265 to the charges.
- Each additional child increases charges by about $509.

6.5 BMI and Smoking Interaction

The effect of BMI on insurance charges is significantly different for smokers compared to non-smokers:

- For non-smokers, each unit increase in BMI is associated with only a $5.66 increase in charges.
- For smokers, each unit increase in BMI is associated with a much larger increase of $1,439.44.

6.6 Heteroscedasticity

Our initial models showed significant heteroscedasticity, which we successfully addressed through the incorporation of interaction terms. This highlights the importance of thorough diagnostic testing in regression analysis.

6.7 Implications

These findings have important implications for both the insurance industry and public health policy:

- **For insurers:** This model provides a more accurate way to predict insurance charges based on individual characteristics.
- **For policymakers:** The significant impact of smoking, especially in conjunction with higher BMI, underscores the potential healthcare cost savings that could be achieved through targeted public health interventions.
- **For individuals:** This analysis highlights the financial implications of lifestyle choices, particularly smoking, on healthcare costs.

6.8 Future Research Directions

Future research directions could include:

- Exploring additional variables that might influence insurance charges, such as specific health conditions or lifestyle factors.
- Investigating non-linear relationships between predictors and insurance charges.
- Conducting longitudinal studies to understand how changes in these factors over time affect insurance charges.

In conclusion, our analysis provides a robust model for predicting insurance charges and offers valuable insights into the complex interplay of factors that influence healthcare costs.

Bibliography and Credits

- Wickham H, Grolemund G (2023). *R for Data Science (2nd edition)*. O'Reilly Media. URL: `https://r4ds.hadley.nz/`.
- Powers, Christopher A. PharmD; Meyer, Christina M. MHS; Roebuck, M Christopher MBA; Vaziri, Baze MBA. Predictive Modeling of Total Healthcare Costs Using Pharmacy Claims Data: A Comparison of Alternative Econometric Cost Modeling Techniques. Medical Care 43(11):p 1065-1072, November 2005. — DOI: 10.1097/01.mlr.0000182408.54390.00
- Meehan, J., Chou, C., & Khasawneh, M. T. (2015). Predictive modeling and analysis of high-cost patients. IISE Annual Conference.Proceedings, , 2566-2575.
- Zhao, Y., Ash, A.S., Haughton, J. et al. Identifying Future High-Cost Cases Through Predictive Modeling. Dis-Manage-Health-Outcomes 11, 389–397 (2003). https://doi.org/10.2165/00115677-200311060-00005.

- Mullahy, John PhD. Econometric Modeling of Health Care Costs and Expenditures: A Survey of Analytical Issues and Related Policy Considerations. Medical Care 47(7_Supplement_1):p S104-S108, July 2009. — DOI: 10.1097/MLR.0b013e31819c9593.
- C. R. Stephens, H. Waelbroeck, and S. Talley. 2005. Predicting healthcare costs using GAs. In Proceedings of the 7th annual workshop on Genetic and evolutionary computation (GECCO '05). Association for Computing Machinery, New York, NY, USA, 159–163.
- Duncan, I., M. Loginov, and M. Ludkovski. 2016. "Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs." North American Actuarial Journal 20 (1): 65–87. doi:10.1080/10920277.2015.1110491.
- Liao, Xiyue, Ian Duncan, and Samuel O'Neill. 2023. "Alternative Predictive Models for Medicare Patient Cost." North American Actuarial Journal 28 (1): 126–38. doi:10.1080/10920277.2022.2161577.