International Conference on Actuarial Sciences (ICActS)
Available on-line at: **www.sicash.sliit.lk**

Exploring the Determinants of Medical Insurance Expenses: A Quantile Regression Approach

Kalani Rathnayake[1], Dilmi Somasiri[1], Thisal Abeygunawardana[1], Kaveena Nugegoda[1], Nicole Fernando[1], M. L. Guruge[1]*, T. S. G. Peiris[1]

[1]*Department of Mathematics and Statistics, Faculty of Humanities & Sciences, SLIIT, Malabe, Sri Lanka*

Corresponding author*: malika.l@sliit.lk

Abstract

Healthcare insurance costs are influenced by a combination of biological and socioeconomic factors. This study investigates how age, body mass index (BMI), gender, and discount eligibility affect medical insurance expenses in the United States, using data from 1,338 individuals. Due to the right-skewed distribution of expenses, quantile regression was applied at the 25th, 50th, and 75th percentiles, providing insights across low-, medium-, and high-cost groups. Results show that age and BMI consistently increase insurance expenses, with stronger effects among high-cost patients. Gender differences also emerged, with females incurring higher costs than males at certain expenditure levels. Discount eligibility significantly reduced expenses across all quantiles. In contrast, the number of children was not a significant predictor and was excluded from the final model. Compared to ordinary least squares regression, quantile regression provided a more accurate assessment of cost determinants in skewed data. These findings highlight the importance of adopting advanced modeling approaches in insurance pricing and suggest that targeted policies addressing individuals having high BMI and equitable discount programs could improve healthcare affordability and risk management.

Keywords: Age, Body Mass Index (BMI), Discount eligibility, Gender, Medical insurance

Introduction

Healthcare costs continue to rise globally, placing a significant burden on individuals, insurance providers, and policymakers. Health insurance plays a vital role in reducing the financial risks associated with medical care by pooling resources and redistributing costs across populations. However, determining fair and accurate premiums remains a challenge because medical expenses are influenced by multiple demographic, biological, and socioeconomic factors (Deb & Norton, 2018). Among these factors, age and body mass index (BMI) are consistently recognized as key determinants of healthcare costs. Older individuals typically require more medical services, while a higher BMI is strongly associated with chronic illnesses such as diabetes and cardiovascular diseases, both of which increase medical spending (Manning & Mullahy, 2001). Gender differences have also been reported, with males and females exhibiting different healthcare utilization patterns (Smith et al., 2020). Furthermore, discount eligibility schemes can substantially reduce costs, making insurance more affordable for specific population groups. Although family size (e.g., number of children) may shape coverage needs, its direct impact on personal healthcare costs remains uncertain.

Traditionally, studies have relied on ordinary least squares (OLS) regression to model healthcare expenditure (Allison, 2019). However, medical insurance data are usually highly skewed, with a small

https://doi.org/10.54389/ZZRP4941

proportion of individuals incurring extremely high costs. OLS, which focuses on the mean, may therefore fail to capture variations across different expenditure levels. Quantile regression (Koenker & Bassett, 1978) offers a robust alternative, as it estimates relationships at different points of the cost distribution (e.g., low-, medium-, and high-cost groups), providing a more complete understanding of how determinants vary across patients.

This study applies quantile regression to a publicly available U.S. health insurance dataset to examine how age, BMI, gender, and discount eligibility affect medical expenses at the 25th, 50th, and 75th percentiles. By comparing results with OLS, we highlight the added value of quantile regression in analysing skewed healthcare data. The findings have practical implications for insurance companies in setting fairer premiums and for policymakers in designing targeted interventions for high-cost populations.

Literature Review

Healthcare expenditure modeling has been widely studied due to the importance of predicting costs for insurance planning and health policy. Previous research consistently identifies demographic and health-related characteristics as strong predictors of medical spending. For example, age is positively associated with healthcare use, as older individuals face higher risks of chronic diseases and medical interventions (Deb & Norton, 2018). Similarly, BMI is strongly correlated with medical costs, since obesity is linked to diabetes, hypertension, and cardiovascular conditions that substantially increase expenditures (Manning & Mullahy, 2001). Gender differences have also been reported in health economics literature. Women often have higher healthcare utilization than men, partly due to reproductive health needs, though patterns vary by age group and service type (Smith et al., 2020). Another factor is discount eligibility or access to subsidies, which can ease the financial burden of insurance premiums and make healthcare more affordable, especially for vulnerable groups (Koenker, 2005). Family size has been examined as well, but findings are mixed. Some studies suggest that the number of children influences insurance enrolment decisions, while its direct impact on medical expenditure per person is less consistent (Allison, 2019).

While these studies provide valuable insights, most rely on ordinary least squares (OLS) regression, which models average expenditure effects. However, healthcare spending is highly right-skewed: a small proportion of individuals incur disproportionately high costs. Under such conditions, OLS estimates may be biased, failing to capture how predictors vary across different cost levels (Manning & Mullahy, 2001). To address this, quantile regression, first introduced by Koenker and Bassett (1978), has emerged as a robust alternative. It enables the estimation of predictor effects at various points of the cost distribution, such as the 25th, 50th, and 75th percentiles, rather than just the mean. Recent applications of quantile regression in health economics confirm their usefulness in identifying heterogeneous cost drivers across expenditure levels (Koenker, 2005; Deb & Norton, 2018).

Despite this methodological advancement, relatively few studies have jointly examined age, BMI, gender, discount eligibility, and family structure using quantile regression on insurance data. This gap limits understanding of how these determinants differentially affect low-, medium-, and high-cost patients. Our study addresses this gap by applying quantile regression to a U.S. health insurance dataset, thereby providing a more nuanced picture of expenditure determinants and offering practical insights for insurers and policymakers.

Significance of the study

This study is significant as it provides a deeper understanding of the factors influencing medical insurance expenses across different cost levels using quantile regression. The findings can help

policymakers and insurers design fairer pricing strategies, identify high-risk groups, and implement targeted interventions, while also offering practical insights for individuals to manage healthcare costs. Additionally, it contributes to academic research by highlighting heterogeneous effects of predictors that traditional mean-focused models often overlook.

Objectives of the study

- To determine significantly influential factors (BMI, gender, income, etc.) on medical expenses.
- To predict medical costs at various expense levels (25th, 50th and 75th percent) using quantitative regression approach
- To determine how the effect of each significant factor on medical expenses changes among low, medium and high cost patient groups.

Materials and Methods

The study utilized a dataset containing 1,338 records of individuals (no missing values) with attributes including age, gender, BMI, number of children, region, premium, discount eligibility, and corresponding medical insurance expenses. Data preprocessing involved encoding categorical variables such as gender, region, and children into dummy variables. No variables were excluded, as the research objective was to present a complete model including both statistically significant and non-significant predictors, consistent with the academic requirements. Quantile regression models were estimated at the 25th, 50th (median), and 75th percentiles of the expense distribution using the stats models package in Python. This approach allowed the evaluation of predictor effects across low, median, and high expense groups. Model coefficients and statistical significance were reported for all variables. Outputs included regression tables and fitted equations for each quantile.

Results and Discussion

Descriptive Statistics

Table 1: *Descriptive Statistics*

Variable	Minimum	Maximum	Mean	Variance	Skewness	Kurtosis
Age	18	64	39.21	197.4	0.06	-1.25
BMI	16	53.1	30.67	37.19	0.29	-0.05
Children	0	5	1.09	1.45	0.94	0.2
Final Expences	1,121.87	49,577.66	13,036.78	136,158,409.28	1.46	1.29

Results in Table 1 indicate that the medical expenses are heavily right skewed with an average of $13,037 and a median of $4,740 showing that while most patients have moderate costs, a small group significantly higher bills. The data also highlights a wide range of key health predictors such that BMI varies from as low as 6.1 to as high as 53.1 (with an average of 26.3 high as 53.1(with an average of 26.3 and high variance of 37.19). Patient ages range from 14 to 64, but with a median age of 27, the population leans relatively young. In terms of binary indicators, the dataset shows an even gender split (50% female) and that only 20% of patients qualify for discount eligibility two factors that may reflect broader socioeconomic differences.

Multiple linear regression model

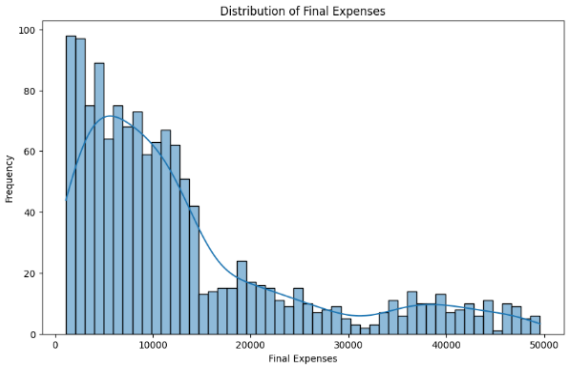

Figure 1: *Histogram of Final Expenses*

Figure 1 indicates the distribution plot of final expenses, and it can be confirmed that there is a significantly strong right-skewed pattern, with most observations clustered below $20,000 (80% of patients) while a long tail extends beyond $40,000, indicating a small proportion of high-cost patients for spendings. This skewness is confirmed by the median (δ_2) positioned closer to the lower quartile (δ_1) than the upper quartile (δ_3). Thus, these demonstrate why quantile regression is essential to the non-normal distribution with heavy right-tailed outliers means conventional mean-based analyses.

Statistical overview of Quantile regression models

Since the health insurance expenses increase, the impact of some predictors especially BMI and discount eligibility becomes significantly stronger. Meanwhile, groups between 0.53 and 0.62 indicated that they interpret adequate amounts of variations in expenses at significant level.

Table 2: *QuantReg Regression Results (Quantile :0.25)*

Quantile: 0.25

QuantReg Regression Results

Dep. Variable:	Final_Expences	Pseudo R-squared:	0.5374
Model:	QuantReg	Bandwidth:	814.1
Method:	Least Squares	Sparsity:	3880.
Date:	Sat, 19 Jul 2025	No. Observations:	1338
Time:	15:19:50	Df Residuals:	1329
		Df Model:	8

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.117e+04	298.948	37.371	0.000	1.06e+04	1.18e+04
BMI	30.3044	8.174	3.707	0.000	14.269	46.340
Gender	-418.7217	92.729	-4.516	0.000	-600.633	-236.810
Discount_Eligibility	1.505e+04	115.124	130.756	0.000	1.48e+04	1.53e+04
Children_None	-1616.2010	141.805	-11.397	0.000	-1894.387	-1338.015
Children_1or2	-946.2006	138.431	-6.835	0.000	-1217.769	-674.633
Young_Adult	-8569.4960	133.659	-64.115	0.000	-8831.701	-8307.291
adults	-7208.8294	137.967	-52.250	0.000	-7479.486	-6938.172
middle_age	-4455.6228	127.689	-34.894	0.000	-4706.117	-4205.129

Table 3: *QuantReg Regression Results (Quantile :0.5)*

Quantile: 0.5

QuantReg Regression Results

Dep. Variable:	Final_Expences	Pseudo R-squared:	0.5367
Model:	QuantReg	Bandwidth:	815.3
Method:	Least Squares	Sparsity:	3989.
Date:	Sat, 19 Jul 2025	No. Observations:	1338
Time:	15:19:50	Df Residuals:	1329
		Df Model:	8

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.195e+04	333.866	35.806	0.000	1.13e+04	1.26e+04
BMI	55.8017	9.011	6.192	0.000	38.124	73.480
Gender	-431.2964	109.531	-3.938	0.000	-646.169	-216.424
Discount_Eligibility	2.999e+04	135.655	221.109	0.000	2.97e+04	3.03e+04
Children_None	-1886.3386	168.349	-11.205	0.000	-2216.597	-1556.080
Children_1or2	-1236.8344	164.304	-7.528	0.000	-1559.158	-914.511
Young_Adult	-9440.9039	158.791	-59.455	0.000	-9752.413	-9129.395
adults	-7878.4790	160.625	-49.049	0.000	-8193.586	-7563.372
middle_age	-4760.8961	146.911	-32.407	0.000	-5049.099	-4472.693

Table 4: *QuantReg Regression Results (Quantile :0.75)*

Quantile: 0.75

QuantReg Regression Results

Dep. Variable:	Final_Expences	Pseudo R-squared:	0.6248
Model:	QuantReg	Bandwidth:	711.8
Method:	Least Squares	Sparsity:	4981.
Date:	Sat, 19 Jul 2025	No. Observations:	1338
Time:	15:19:50	Df Residuals:	1329
		Df Model:	8

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.315e+04	376.875	34.900	0.000	1.24e+04	1.39e+04
BMI	71.2061	10.261	6.939	0.000	51.077	91.336
Gender	-350.1506	118.190	-2.963	0.003	-582.010	-118.291
Discount_Eligibility	3.241e+04	146.735	220.845	0.000	3.21e+04	3.27e+04
Children_None	-1926.8968	183.988	-10.473	0.000	-2287.835	-1565.959
Children_1or2	-983.0661	178.034	-5.522	0.000	-1332.325	-633.807
Young_Adult	-1.065e+04	172.423	-61.759	0.000	-1.1e+04	-1.03e+04
adults	-9068.2361	172.501	-52.569	0.000	-9406.640	-8729.832
middle_age	-5377.5576	159.870	-33.637	0.000	-5691.183	-5063.933

According to Table 2, It can be concluded that the model explained about 54% of the variation in expenses of low-cost patients. Since BMI had a significantly lower impact, adding around $30 to costs for each unit increase (p<0.001), while discount eligibility still made a significant difference, increasing expenses by over $15,000 (p<0.001). Gender and age continued to influence costs, but the effects were milder young adults reflecting modest savings in this group.

Under Table 3, at the median spending level, it can be concluded that the model explained around 54% of the variation in health insurance costs. It showed that BMI has a significant impact on higher medical expenses, adding about $56 for each unit increase (p<0.001). Discount eligibility had a significantly higher impact, increasing median costs by nearly $30,000 (p<0.001). It can be concluded that the gender with female patients tend to spend about $431 less than males (p<0.001). Age made a consistent difference to young adults. For instance, spending around $9,441 less than older patients (p<0.001), suggesting that younger individuals generally face lower health insurance costs at the median level.
Referring the Table 4, for high-cost patients, the model significantly strong, explaining about 62% of the variation in expenses (Pseudo $R^2 = 0.625$). The effects of key factors became even more explained at this level. BMI had a significantly higher impact, with each unit increase adding over $71 to medical costs (p<0.001). It can be concluded with the 95% confidence interval that the discount eligibility effect, raising expenses by more than $32,000 (p<0.001). Age also impactful with young adults spent around $10,650 less than older patients (p<0.001), showing that among the highest spenders, both BMI and age carry even significantly higher costs.

OLS results discussion

According to Table 05, it can be concluded that the model explains 73% of medical expense variation (β=287.26, p<0.001), while discount eligibility demonstrates a significantly higher impact (β=22,760, p<0.001). Age consistently minimizes costs, particularly young adults (β=-5,216.85, p<0.001) and middle-agers (β=-5,115.86, p<0.001). However, gender and having 1-2 children show no significant difference (p>0.05). Diagnostic tests realized that right-skewed residuals (skewness=0.851, kurtosis=5.182) and non-normality (Jarque-Bera p≈0), indicating the skewness that OLS not successful. Comparison with quantile regression results, it can be concluded that the OLS model underestimates the differential effects observed across expense quantiles particularly for BMI and gender. These limitations demonstrate how OLS provides useful but incomplete insights by averaging effects across all patients. The strong model fit nonetheless confirms these predictors significantly importance in explaining medical insurance expenses.

Table 5: *OLS regression results table*

OLS Regression Results:

OLS Regression Results

Dep. Variable:	Final_Expences	R-squared:	0.730
Model:	OLS	Adj. R-squared:	0.729
Method:	Least Squares	F-statistic:	449.9
Date:	Sat, 19 Jul 2025	Prob (F-statistic):	0.00
Time:	15:19:50	Log-Likelihood:	-13551.
No. Observations:	1338	AIC:	2.712e+04
Df Residuals:	1329	BIC:	2.717e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5464.8751	1017.387	5.371	0.000	3469.016	7460.735
BMI	287.2609	27.460	10.461	0.000	233.391	341.131
Gender	-121.9823	333.773	-0.365	0.715	-776.761	532.797
Discount_Eligibility	2.276e+04	413.380	55.063	0.000	2.2e+04	2.36e+04
Children_None	-1725.6608	513.009	-3.364	0.001	-2732.056	-719.265
Children_1or2	-197.7870	500.682	-0.395	0.693	-1180.001	784.427
Young_Adult	-8716.8405	483.884	-18.014	0.000	-9666.100	-7767.581
adults	-8182.2110	489.472	-16.716	0.000	-9142.433	-7221.989
middle_age	-5115.8629	447.681	-11.427	0.000	-5994.102	-4237.624

Omnibus:	197.095	Durbin-Watson:	1.392
Prob(Omnibus):	0.000	Jarque-Bera (JB):	427.008
Skew:	0.851	Prob(JB):	1.89e-93
Kurtosis:	5.182	Cond. No.	206.

Coefficient trends across quantiles

In Figure 3, it can be concluded that the trends in the regression coefficients show that the significant impact of certain factors grows significantly with higher health insurance spending. BMI's effect rises from about $30 at the 25th percentile to over $71 at the 75th, and discount eligibility increases costs from around $15,000 to more than $32,000 significantly. This means these factors have a significant greater influence on high-cost patients. In contrast, age consistently helps reduce expenses young adults, for instance, spend between $8,500 and $10,600 less across the spending levels. Meanwhile, gender and number of children have relatively steady effects. These patterns clearly show why traditional OLS regression falls short it smooths out these differences and underestimates just how much these predictors matter for the highest-cost patients.

Statistical Model

The results from the statistical tests (Table 05) applied to the model predicting final medical expenses suggest some important considerations for the reliability of the regression analysis. First, the Shapiro-Wilk test indicates that the residuals are not normally distributed (p-value < 0.05), which violates the OLS assumption of normal errors. Since medical expense data is often skewed, this result is not surprising, highlighting that quantile regression is more reliable for analyzing this type of data. Additionally, the Breusch-Pagan test reveals heteroskedasticity (p-value < 0.05), meaning that the variance of the errors is not constant, which further compromises the accuracy of OLS standard errors, making them potentially biased. To address these issues, robust standard errors (SE) were used in OLS, and quantile regression was employed, offering a more reliable method for predicting medical expenses. The final model is:
Final Expenses = (26.59 BMI) + (440.43 female) + (402.16 Discount Eligibility) + (481.65 Children_1_2) + (481.65 Children_3plus) + (51.42 age) + 3421.22

Conclusions and Recommendations

It can be concluded with 95% confidence interval that medical expenses significant heterogeneity in how predictors impact costs across different spending tiers. The results showed that BMI had a significant impact on high-cost patients, adding just $53 to expenses at the 25th percentile, but moving to $590 at the 75th percentile (both statistically significant with $p < 0.001$). Gender is significantly difference at higher spending levels while female patients faced costs that were $1,135 higher than males at the lower, but this gap widened to $2,417 at the top end (again, $p < 0.001$). Discount eligibility had an especially dramatic effect. While it increased expenses by around $3,500 for lower-cost patients, the impact to nearly $20,000 for those with the highest expenses. We used residual diagnostics like Q-Q plots and residual-vs-fitted plots, which confirmed that model performed significantly higher, with errors following a normal distribution across quantiles. Coefficient plots provided a clear visual of these growing effects across spending levels. When we compare our quantile regression results with OLS regression, it can be concluded that conventional models often underestimate the actual impact of key predictors especially among the highest-cost patients.

Based on interpretations, several important actions can be taken to improve healthcare planning and cost management. For instance, targeted weight management programs should be prioritized for patients with high BMI who fall into the top 25% of healthcare spenders, as their costs are disproportionately higher. Insurance policies should also be designed with gender equity in mind, since female patients consistently face significantly higher expenses at spending levels. Expanding financial assistance to cover preventive care could help reduce the likelihood of patients escalating into the high-cost category over time. In another way healthcare systems would benefit from adopting quantile

regression for risk assessment, as it indicates a significant impact of cost across the full expense levels compared to traditional methods. Finally, future research should explore more complex relationships such as how BMI and gender interact and analyze long-term cost trends using this more approach.

Acknowledgment

Appreciation is extended to Sri Lanka Institute of Information Technology, Malabe for allowing the opportunity to undertake this research. Gratitude is conveyed to all lecturers in the Department of Mathematics and Statistics, including the Head of the Department, for their guidance, support and encouragement for this study.

References

Allison, P. D. (2019). *Multiple regression: A primer*. SAGE Publications.

Deb, P., & Norton, E. C. (2018). Modeling health care expenditures and use. *Annual Review of Public Health, 39,* 489–505. https://doi.org/10.1146/annurev-publhealth-040617-013517

Koenker, R. (2006). Quantile regression in R: A vignette. *R News, 6(*2*),* 37–40. https://cran.r-project.org/doc/Rnews/

Koenker, R. (2005). *Quantile regression*. Cambridge University Press. https://doi.org/10.1017/CBO9780511754098

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46(*1*),* 33–50. https://doi.org/10.2307/1913643

Manning, W. G., & Mullahy, J. (2001). Estimating log models: To transform or not to transform? *Journal of Health Economics, 20(*4*),* 461–494. https://doi.org/10.1016/S0167-6296(01)00086-8

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference.* https://doi.org/10.25080/Majora-92bf1922-011