Table of Contents

Group 12
Jigyasa Bardalaye 240019637
Noor Karaman 1900014984
Teanna Puthucheary 240063221

1. Identification of Variables

In this analysis, we identified expenditure on doctor visits as the response variable. By analysing the factors influencing healthcare expenditure, it allows us to draw meaningful conclusions in assessing the impact of socioeconomic and demographic factors. It can be used to inform policy decisions aimed at improving healthcare access and efficiency. As a continuous quantitative variable and a direct measure of health-related spending, it is well suited to be the dependent variable in a regression model.

The selection of independent variables is guided by intuition, existing literature and comparing multicollinearity between variables. The following are potential predictors of doctor visit expenditure:

- **Number of doctor visits:** This is directly linked to medical costs, with more visits typically resulting in higher expenditure.
- **Income:** A major determinant of healthcare access and spending. Higher income allows for more frequent or specialised care, increasing expenditure (Finkelstein et al., 2012)
- **General health:** In general, people in poorer health are more likely to require medical services, leading to higher expenditures.
- **Mental health:** Mental health is closely tied to general health, with individuals experiencing poor mental health often facing comorbid physical conditions. Studies like Scott et al. (2008) show that this comorbidity increases healthcare use and costs.
- **Hyperlipidemia and Hypertension:** It is crucial to understand the impact of chronic illness, particularly cardiovascular disease, as it serves as a key indicator of higher healthcare expenditures
- **Gender:** Gender differences in healthcare usage and spending are well-established. Bertakis (2000) found that women use healthcare services more frequently, while men tend to incur higher costs per visit.
- **Age:** Older individuals typically experience declining general health and require more frequent doctor visits, leading to higher healthcare expenditure. We expect the data will reflect this relationship.
- **BMI:** BMI serves as an indicator of overall health and chronic conditions that demand more frequent medical care. Higher BMI is linked to an increased risk of chronic diseases, such as diabetes and heart disease, resulting in more doctor visits and higher expenditures. Sturm (2003) found that individuals with elevated BMI incur greater healthcare costs.
- **Education:** Higher education levels often correlate with better access to healthcare resources and information (Ross & Wu, 1995).

The variables excluded from the analysis:

- **Region:** excluded due to its broad classification, restricting the ability to capture location-specific trends in healthcare expenditure. A more detailed state-level classification could provide more meaningful insights.
- **Number of non-doctor visits and expenditure:** if included it could obscure the factors directly affecting doctor visit expenditures.
- **Ethnicity**: is excluded due to insufficient relevance to direct expenditure.

A regression of the mentioned variables indicates a tolerable level of collinearity between the data seen in Figure 1.0, none of the VIFs are near the threshold of 5, indicating no cause for concern around collinearity.

```
> vif(lm1.fit)
      general        mental           bmi        income           age        gender     education  hypertension
     2.036539      1.787937      1.132620      1.205639      1.347750      1.042267      1.181678      1.405919
hyperlipidemia        dvisit
     1.351547      1.205464
```

Figure 1.0

2. Fitting Statistical Models

A Multiple Linear Regression (MLR) model will be employed to analyse the relationship between selected independent variables and expenditure on doctor visits. MLR is an effective analytical tool for understanding and forecasting the relationship between the dependent and multiple independent variables (Anderson et al., 2024). Figure 2.0 illustrates our fitted model, where the βs estimate the relationship between each independent variable and the dependent variable. Specifically, it indicates the change in expenditure associated with a one-unit increase in the corresponding independent variable, holding other predictors constant (Anderson et al., 2024).

$$dvexpend = \beta_0 + \beta_1 dvisit + \beta_2 income + \beta_3 general + \beta_4 mental + \beta_5 gender + \beta_6 education + \beta_7 hypertension + \beta_8 hyperlipidemia + \beta_9 age + \beta_{10} bmi + \epsilon$$

Figure 2.0

To assess the four key assumptions of Multiple Linear Regression - linearity, homoscedasticity, normality, and independence of residuals - we examine the plots shown in Figure 2.1, which are based on the fitted regression model.

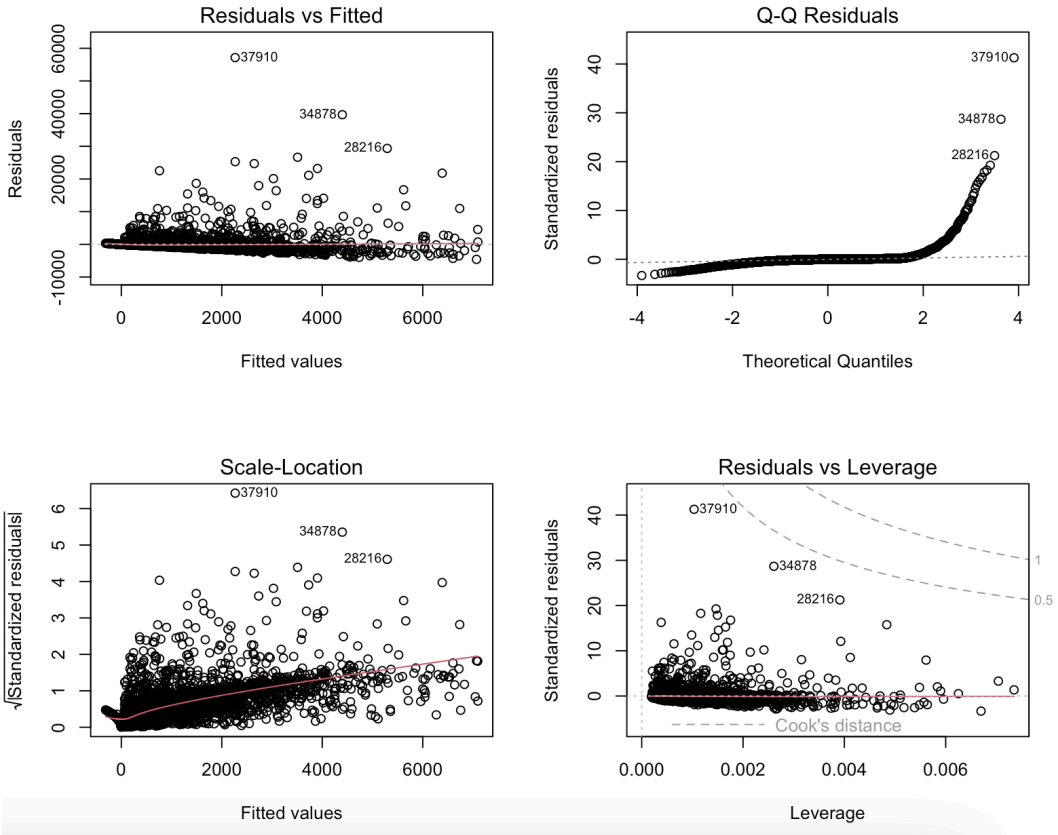

Figure 2.1

Jigyasa Bardalaye 240019637
Noor Karaman 1900014984
Teanna Puthucheary 240063221

Plot 1:
This graph plots the residual against the fitted values to assess linearity. There is no significant increase in variance with the mean, supporting the linearity assumption. The average residuals remain stable across fitted values, indicating a good model fit, displaying homoskedasticity. Although a few outliers are present, their impact is minimal due to the large dataset.

Plot 2:
The Q-Q plot compares standardised residuals against theoretical quantiles to evaluate normality. A systematic deviation from the straight line indicates a departure from normality, particularly at the higher quantile, suggesting that the assumption of normally distributed errors may not be valid, potentially affecting hypothesis testing

Plot 3:
A trend in average values indicates a violation of the constant variance assumption, though residuals scatter around the increasing line, suggesting variance increases at a constant rate. To help us with this issue we can try log transforming our dependent variable. The funnel shape indicates heteroskedasticity, suggesting that the variability of the residuals depends on the fitted values.

Plot 4:
Residuals plotted against leverage reveal influential points. The majority of points exhibit low residuals, while a few high-leverage points (*37910* and *34878*) have minimal impact on model fit, as they fall within Cook's distance (Cook, 1986), exerting a negligible effect on linearity as well as independence of residuals.

3. Interpretation of Results

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.214e+02  1.016e+02  -2.179  0.02932 *
general         5.352e+01  1.820e+01   2.940  0.00329 **
mental         -3.378e+01  1.803e+01  -1.874  0.06100 .
bmi            -3.656e+00  2.229e+00  -1.640  0.10105
income          1.142e-03  2.866e-04   3.985 6.78e-05 ***
age             1.686e+00  1.142e+00   1.477  0.13981
gender          5.048e+01  2.746e+01   1.838  0.06605 .
education       6.363e+00  5.033e+00   1.264  0.20618
hypertension   -9.312e+00  3.666e+01  -0.254  0.79950
hyperlipidemia -2.782e+01  3.731e+01  -0.746  0.45583
dvisit          2.430e+02  4.068e+00  59.732  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1385 on 10627 degrees of freedom
Multiple R-squared:  0.2931,    Adjusted R-squared:  0.2924
F-statistic: 440.6 on 10 and 10627 DF,  p-value: < 2.2e-16
```

Figure 3.0

By interpreting the results in Figure 3.0, the fitted regression equation as follows:

$$\widehat{dvexpend} = -221.4 + 243\,dvisit + 0.001\,income + 53.52\,general - 33.78\,mental - 3.66\,bmi + 0.001\,income + 1.69\,age + 50.48\,gender + 6.36\,education - -9.31\,hypertension - 27.82\,hyperlipidemia$$

Figure 3.1

The regression analysis reveals that **general health** is significant and acts as a predictor of doctor visit expenditures. A one-unit increase in the general health score (indicating poorer health) correlates with an average increase of $53.52 in healthcare spending. This aligns with the understanding that individuals with worse health tend to require more frequent or intensive medical care, leading to higher expenses.

Mental health is not statistically significant at 5% but is significant at 10% significance level. The coefficient shows a negative association between mental health and doctor visit expenditure. On the scale given 1 is perfect mental health and 5 being the worst, we can interpret the coefficient as if we increase the mental health figure by 1(i.e. our mental health gets worse) expenditure on doctor visits falls by $33.78, possibly due to reliance on non-physician mental health services.

The **BMI** coefficient is insignificant, while obesity typically correlates with higher health costs, its impact may be more directly reflected in other variables, such as the number of doctor visits or presence of diabetes.

Income is statistically significant at 5% and exerts a small but positive effect on expenditure. For every additional $1,000 in income, doctor visit expenditures increase on average by $1.14, indicating individuals are likely to utilise healthcare services more or seek more expensive treatments. High income elasticity supports Gerdtham and Jönsson's (2000), argument that healthcare is considered a luxury good.

The effect of **age** is positive yet this variable is statistically insignificant, indicating that age does not exert a strong independent influence on expenditure when accounting for other factors, such as health status and number of doctor visits.

The coefficient of **Gender,** which is significant**,** shows a positive effect on expenditure, suggesting that men spend about $50 more on doctor visits than women.

Education is not statistically significant, indicating that higher education levels do not directly impact doctor visit spending contradicting Ross and Wu (1995).

Both **hypertension** and **hyperlipidemia** are statistically insignificant and both have negative coefficients, indicating these chronic conditions do not independently raise doctor visit expenditure. This may be due to treatment costs typically being covered under separate healthcare expenses (eg. medication rather than visits or insurance).

The **number of doctor visits** is significant and the strongest predictor of expenditure, with each additional visit corresponding to an average increase of $243. This aligns with our intuitive expectation that more visits directly results in higher costs.

The model explains around 29% of the variability in doctor visit expenditures $(Adjusted\ R^2 = 0.29)$, which, although modest, is reasonable for healthcare data in the US that is often influenced by unmeasured factors like insurance plans. The F-statistic of 440.6 confirms the model's statistical significance, indicating that the independent variables collectively explain the variation in expenditures.

4. Limitations of the Model

The Multiple Linear Regression (MLR) model has limitations, especially its assumption of linearity between the dependent and independent variables. In reality, this often doesn't hold, leading to biassed predictions and missed patterns. Alternative models, such as the polynomial regressions (Miguez et al., 2018), may provide a more accurate representation of these relationships.

Non-normality of residuals, seen in deviations from the QQ plot (Plot 2), is a concern. Healthcare expenditures are often right-skewed due to a few individuals with very high costs, which, if unaddressed, can bias confidence intervals and p-values. The assumption of homoskedasticity was violated, as seen in the upward trend in the scale-location plot (Plot 3). Heteroskedasticity is common in healthcare data, with some individuals, like those with chronic conditions, incurring disproportionately high costs. This issue will be addressed in the next section.

A limitation of the data is a potential measurement error in self-reported variables, such as general and mental health, due to their subjective nature. Inaccurate reporting could undermine the estimated relationship between health status and expenditure, resulting in inconsistent findings.

Lastly, while the model explains little variance, this is common in healthcare data, where unobserved factors like insurance plans play a large role. Phelps and Newhouse (1974) show that insurance significantly affects medical costs in the US, and excluding such variables can lead to omitted variable bias, limiting the model's explanatory power.

5. Suggestions for Improving the Model

To enhance the accuracy and robustness of the regression model, several improvements could be made. Given the issues with heteroskedasticity and non-normality of residuals, applying a log transformation to expenditure could stabilise variance and bring the residuals closer to a normal distribution. Therefore, better reflect the assumptions of linear regression and improve the reliability of the estimates. If heteroskedasticity persists, heteroskedasticity-robust standard errors would offer more reliable hypothesis testing by adjusting standard errors without altering the model structure.

Incorporating additional variables, particularly insurance coverage, could further improve the model as insurance significantly influences healthcare expenditure in the US (Phelps & Newhouse, 1974). Other factors, such as access to healthcare and lifestyle variables like smoking or physical activity could also reduce omitted variable bias and provide a more comprehensive analysis.

Finally, exploring more complex non-linear models with quadratic or interaction terms, or even a Markov switching model (Kuan, 2002), with general health as the regime-switching parameter could further refine the analysis.

Implementing these improvements would likely lead to a more accurate and reliable model, offering better insights into the factors driving expenditure on doctor visits.

6. References

Anderson, D.R., Sweeney, D.J., Williams, Thomas A., *et al.* (2024) 'Chapter 15 Multiple Regression', in *Statistics for Business and Economics*. 6th Edition. Andover, Hampshire: Cengage Learning, pp. 458–468. Available at: https://www.vlebooks.com/Product/Index/3328625?page=0&startBookmarkId=-1 (Accessed: 20 October 2024).

Bertakis, K.D. (2000) Gender differences in the utilization of health care services - ..., Gender differences in the utilization of health care services . Available at: https://go.gale.com/ps/i.do?id=GALE%7CA60039859&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=00943509&p=AONE&sw=w (Accessed: 22 October 2024).

Cook, R.D., 1986. [Influential observations, high leverage points, and outliers in linear regression]: Comment. *Statistical Science*, *1*(3), pp.393-397.

Finkelstein, A. et al. (2012) The Oregon Health Insurance Experiment: Evidence from the first year*, OUP Academic. Available at: https://academic.oup.com/qje/article-abstract/127/3/1057/1923446 (Accessed: 22 October 2024).

Gerdtham, U.G. and Jönsson, B., 2000. International comparisons of health expenditure: theory, data and econometric analysis. In *Handbook of health economics* (Vol. 1, pp. 11-53). Elsevier.

Kuan, C.M., 2002. Lecture on the Markov switching model. *Institute of Economics Academia Sinica*, *8*(15), pp.1-30.

Miguez, F., Archontoulis, S. and Dokoohaki, H., 2018. Nonlinear regression models and applications. *Applied statistics in agricultural, biological, and environmental sciences*, pp.401-447.

Phelps, C. and Newhouse, J., 1974. Coinsurance and the demand for medical services .p48

Roland Sturm, P. (2003) Increases in clinically severe obesity in the United States, 1986-2000, Archives of Internal Medicine. Available at: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/216155 (Accessed: 22 October 2024).

Ross, C.E. and Wu, C.L., 1995. The links between education and health. *American sociological review*, pp.719-745.

Scott, K.M., Von Korff, M., Alonso, J., Angermeyer, M.C., Bromet, E., Fayyad, J., De Girolamo, G., Demyttenaere, K., Gasquet, I., Gureje, O. and Haro, J.M., 2009. Mental–physical comorbidity and its relationship with disability: results from the World Mental Health Surveys. *Psychological medicine*, *39*(1), pp.33-43.

7. Appendix

Using forward selection to determine which predictors are associated with the response variable, expenditure on doctor visits. As seen in Figure 7.0, the resulting equations are expected to closely mirror those in Figure 3.1, given the closely identical y-intercepts and β coefficients.

```
Call:
lm(formula = dvexpend ~ dvisit + income + general + mental +
    gender + bmi, data = expenditure)

Coefficients:
(Intercept)       dvisit       income      general       mental       gender          bmi
 -85.931894   243.879947     0.001307    53.103682   -34.876541    47.926627    -3.749453
```

Figure 7.0

The plots in Figure 7.1 closely resemble those in Figure 2.1, demonstrating the model's strong fit.

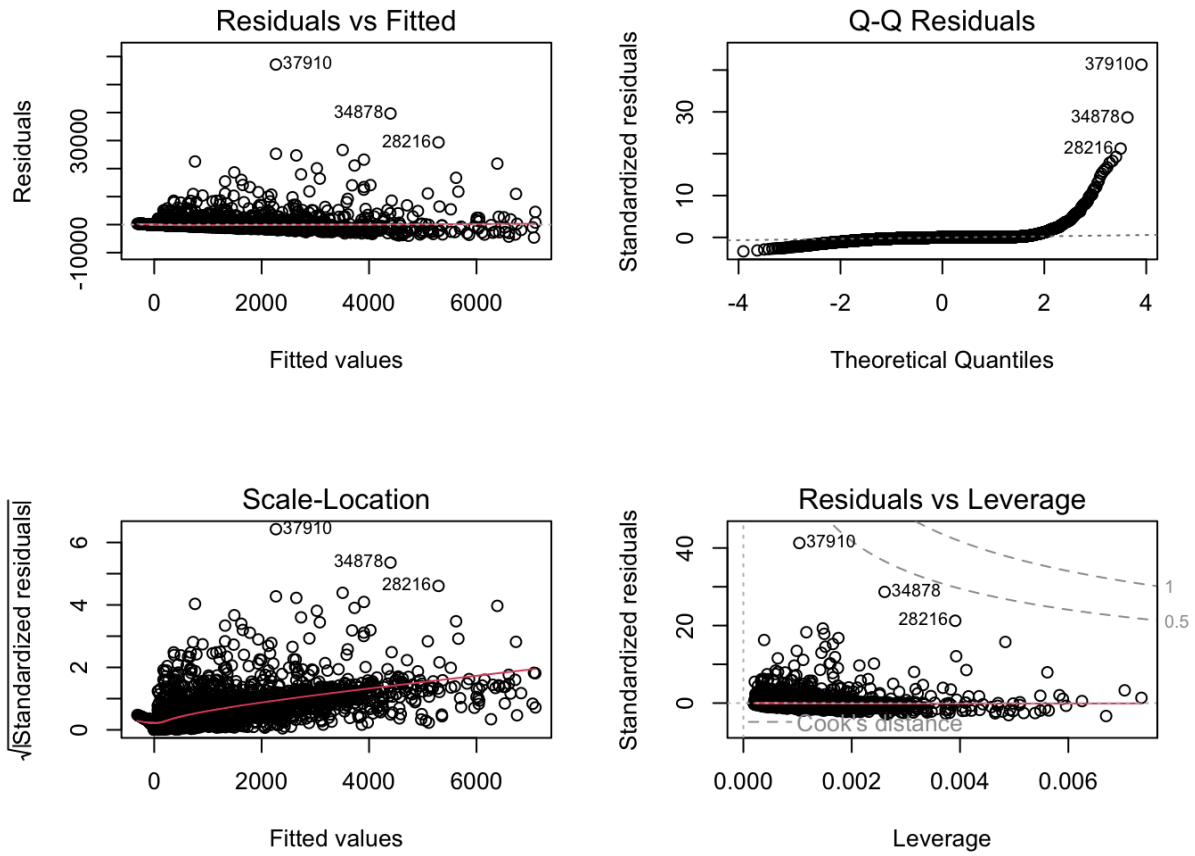

Figure 7.1

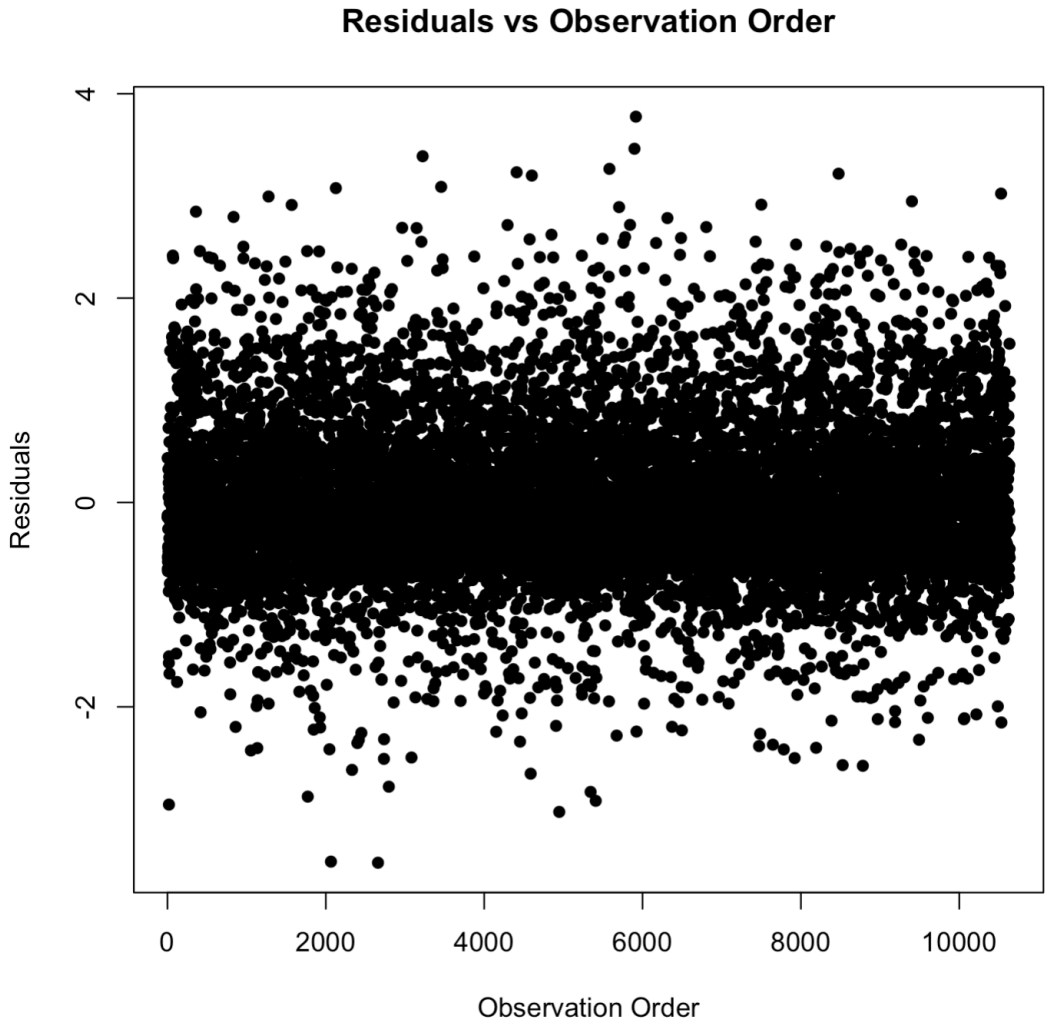

Figure 7.2

We inspect the plot to check the assumption of independence of residuals. The residuals are plotted against the order of observation, showing no discernible trend, with dispersion centred around 0 as expected. Therefore, we can assume the residuals are independently distributed.