Understanding the Determinants of Healthcare expenditure Amongst Medicare Beneficiaries

Amanny Abuthuraya

ECON 400: Econometrics

Professor Liliana Lawrence

July 18, 2024

Introduction

CMS.gov is a government website that provides data about Medicare and Medicaid health programs. These programs are type of government aid for US citizens, and recently have been extended to include immigrants in some states like Washington (Kamb, L., 2024). Medicare/Medicaid are health insurance programs based on either income, disability or age (USHHS, 2022). Every year, the Medicare Current Beneficiary Survey is conducted by the CMS and NORC. They select and contact random beneficiaries to interview them to fill out the survey (California Health Advocates, 2023)

Since 1991, the MCBS has been working towards the goal of improving Medicare and understanding the needs, costs and experiences of beneficiaries. By understanding the program's coverage, legislators and policymakers can focus on addressing the issues that arise. In this paper, my goal is to analyze the MCBS sample dataset to understand how a variety of variables like age, sex, race, income, chronic conditions, payments and different types of healthcare events effect the total healthcare expenditure.

Medicaid and Medicare are topics of interest to me, because I would like to broaden my scope of analysis and learn more about medical analysis. My analysis can be used in policy development for those associated with high spending trends and help develop care strategies and preventative measures for them.

To understand healthcare expenditure among beneficiaries of Medicare and Medicaid, it is important to review other papers that are relevant to the topic. I will be discussing two papers, "Quality, Health and Spending in Medicare Advantage and Traditional Medicare" published by AMJC and "Out-of-pocket health spending among Medicare beneficiaries: Which chronic diseases are most costly?" published by PLOS ONE, respectively.

"Quality, Health and Spending in Medicare Advantage and Traditional Medicare" compares Medicare Advantage (MA) and Traditional Medicare based on quality, health and cost outcomes (Xu, W. et al, 2021). It was found that MA plans tend to provide better quality care with lower spending costs when compared to TM. It suggests that the Medicare plan can affect the overall health expenditure.

Meanwhile, "Out-of-pocket health spending among Medicare beneficiaries: Which chronic diseases are most costly?", examines how supplemental costs, that are usually paid by "gap insurances", are associated with higher expenditure on healthcare (Johnston, K et al, 2020). Beneficiaries who use gap insurances tend to use their healthcare services more, even with Medicare.

After reviewing both research papers, it can be said that the type of insurance that is used by Medicare beneficiaries contributed to their spending trend. With the "higher end" of the line insurances having the most medical costs associated with them. Also, the health condition of the beneficiaries is another important factor in spending trends. Together, these findings suggest that the structure of insurance coverage is important when understanding spending patterns.

Data

The data is taken from CMS.gov, Medicare Current Beneficiary Survey – Cost Supplement. A public use file containing information on expenditures and payment sources allowing researchers to conduct analysis on Medicare beneficiaries living only in the community (cms.gov, 2024).

The dataset originally had over 30 columns, many of which were not needed for this research, so they have been removed. Also, the data had a few categorical data columns that were transformed into separate dummy variables. Meanwhile there were a lot of adjusted columns that were not of use and have been excluded from the original dataset. ID columns and other arbitrary columns were also removed before starting analysis.

Models & Methodologies

1) The main columns of the dataset are:
    - Independent variable: Totalpayment
    - Dependent variables:
        - Categorical variables: sex, race, income and age
        - Numerical variables: numberchroniccond, dentalevent, visionevent, hearingevent, homehealthevent, inpatientevent, medicalproviderevent, outpatientevent, prescribemedicine, medicarepayment, medicaidpayment, medicareadvantagepayment, privateinsurancepayment, outofpocketpayment, uncollectedliability and otherpayments

The word "event" in the variables refer to the places where the patient was treated, and "payment" refers to how their treatment were paid for. The categorical variables needed to be transformed. Each category (for example category 1 & 2) was transformed into a separate dummy variable (0,1) carrying their category name instead for easy analysis.

2) Methodologies
    - Min, Mean and Max:

```
      sex         numberchroniccond  dentalevent
 Min.   :0.0000   Min.   :1.000     Min.   :    0.0
 1st Qu.:0.0000   1st Qu.:2.000     1st Qu.:    0.0
 Median :0.0000   Median :2.000     Median :  143.2
 Mean   :0.4575   Mean   :2.348     Mean   :  819.0
 3rd Qu.:1.0000   3rd Qu.:3.000     3rd Qu.:  597.0
 Max.   :1.0000   Max.   :3.000     Max.   :22708.7
  visionevent       hearingevent    homehealthevent   inpatientevent
 Min.   :    0.0   Min.   :   0.0   Min.   :    0.0   Min.   :     0
 1st Qu.:    0.0   1st Qu.:   0.0   1st Qu.:    0.0   1st Qu.:     0
 Median :   30.0   Median :   0.0   Median :    0.0   Median :     0
 Mean   :  280.0   Mean   : 129.1   Mean   :  288.4   Mean   :  2034
 3rd Qu.:  189.7   3rd Qu.:   0.0   3rd Qu.:    0.0   3rd Qu.:     0
 Max.   :14083.4   Max.   :9615.3   Max.   :27825.2   Max.   :110040
```

```
 medicalproviderevent outpatientevent      prescribemedicine
 Min.   :     0.0      Min.   :     0.00   Min.   :     0.0
 1st Qu.:   685.9      1st Qu.:    21.54   1st Qu.:   253.5
 Median :  1904.2      Median :   254.54   Median :   896.6
 Mean   :  3984.6      Mean   :  2278.99   Mean   :  5333.7
 3rd Qu.:  4598.2      3rd Qu.:  1283.01   3rd Qu.:  4366.3
 Max.   :65844.6       Max.   :105355.57   Max.   :201894.3
  totalpayment     medicarepayment     medicaidpayment
 Min.   :     0   Min.   :     0.0   Min.   :    0.0
 1st Qu.:  2707   1st Qu.:   316.5   1st Qu.:    0.0
 Median :  6897   Median :  1961.5   Median :    0.0
 Mean   : 15148   Mean   :  8017.0   Mean   :  489.6
 3rd Qu.: 16116   3rd Qu.:  6337.2   3rd Qu.:    0.0
 Max.   :307916   Max.   :215227.4   Max.   :34316.1
```

```
 medicareadvantagepayment privateinsurancepayment outofpocketpayment
 Min.   :    0            Min.   :    0.0         Min.   :    0.0
 1st Qu.:    0            1st Qu.:    0.0         1st Qu.:  354.2
 Median :    0            Median :    0.0         Median : 1139.1
 Mean   : 2294            Mean   : 1095.2         Mean   : 2440.3
 3rd Qu.: 1156            3rd Qu.:  645.8         3rd Qu.: 2863.8
 Max.   :91750            Max.   :62549.5         Max.   :37203.2
 uncollectedliability  otherpayment          age_g1
 Min.   :    0.0      Min.   : -168.45   Min.   :0.0000
 1st Qu.:    0.0      1st Qu.:    0.00   1st Qu.:0.0000
 Median :    0.0      Median :    0.00   Median :0.0000
 Mean   :  303.5      Mean   :  507.81   Mean   :0.1678
 3rd Qu.:  109.9      3rd Qu.:   13.76   3rd Qu.:0.0000
 Max.   :24358.1      Max.   :41444.82   Max.   :1.0000
```

Since **otherpayment** has a negative value of -168.45, it has been removed because it isn't logical for a payment to be in negative and should be in positive form. Other than that, the maximum values are going to be kept because one medical bill could easily reach high amounts. It isn't something impossible depending on the patient's health.

- Correlation Plot (Graph 1 in Appendix)

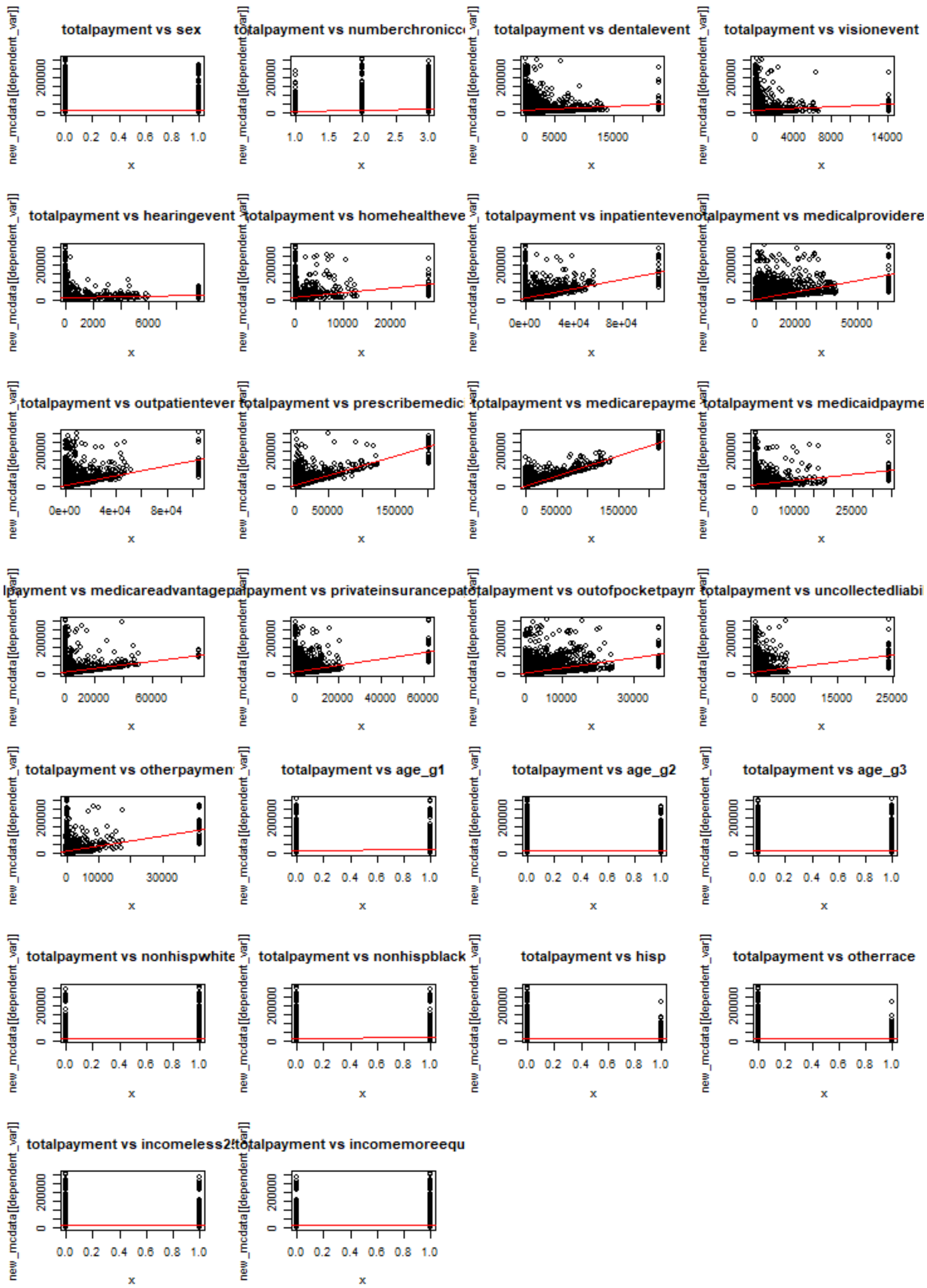
Correlation plots are used to check the multicollinearity issue within the data. Since the data frame was too big to be viewed within R Studio, it was saved as a CSV file to the local device. After going through the columns, there's no issue of severely high correlations (between the values of 0.8 & 0.9). We do have some issues with moderately high correlations, for example:

1- Correlation of **homehealthevent** and **prescribemedicine** = 0.717
2- Correlation of **inpatientevent** and **medicaidpayment** = 0.798
3- Correlation of **medicalproviderevent** and **totalpayment** = 0.717
4- Correlation of **prescribemedicine** and **totalpayment** = 0.717

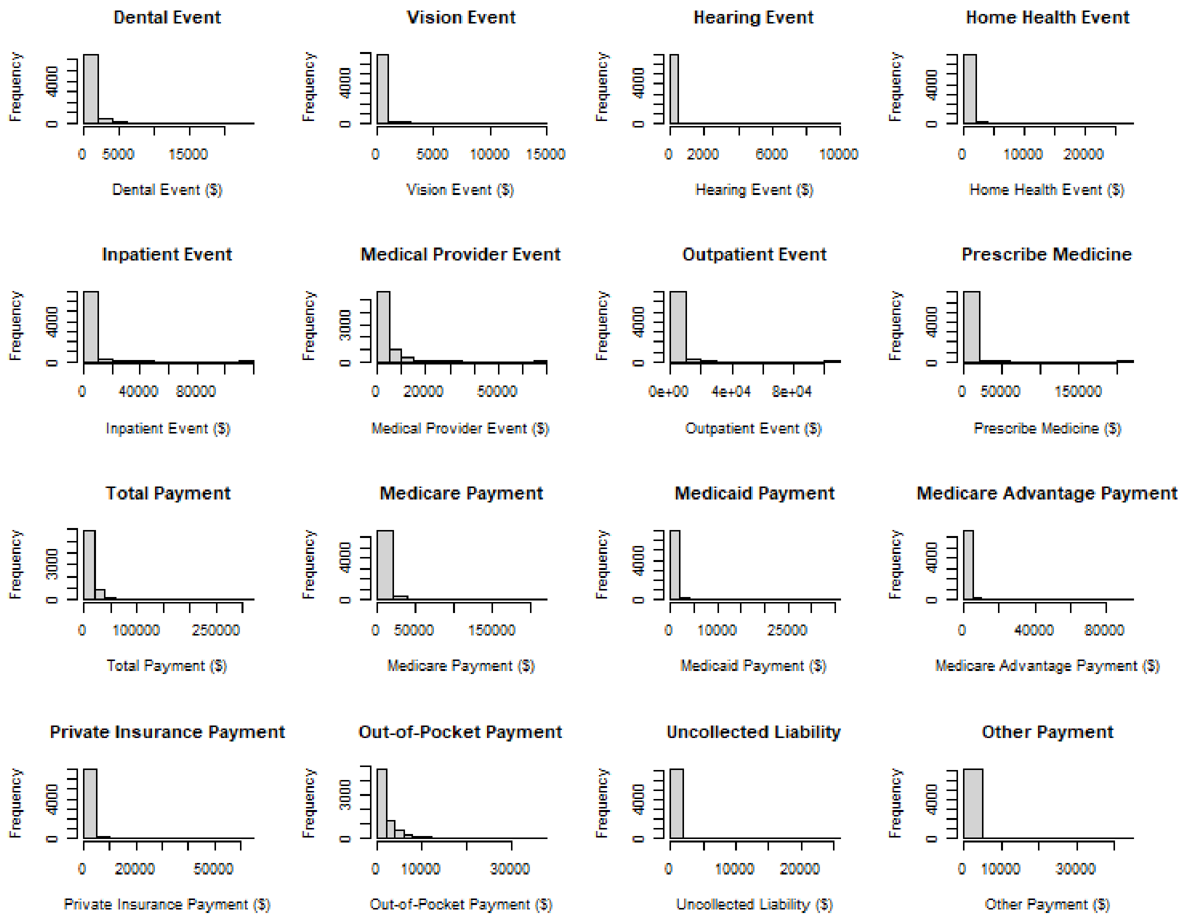
They could cause some problems to a degree, but VIFs of the data need to be checked before omitting any correlated variables.

- Linearity & Skewness

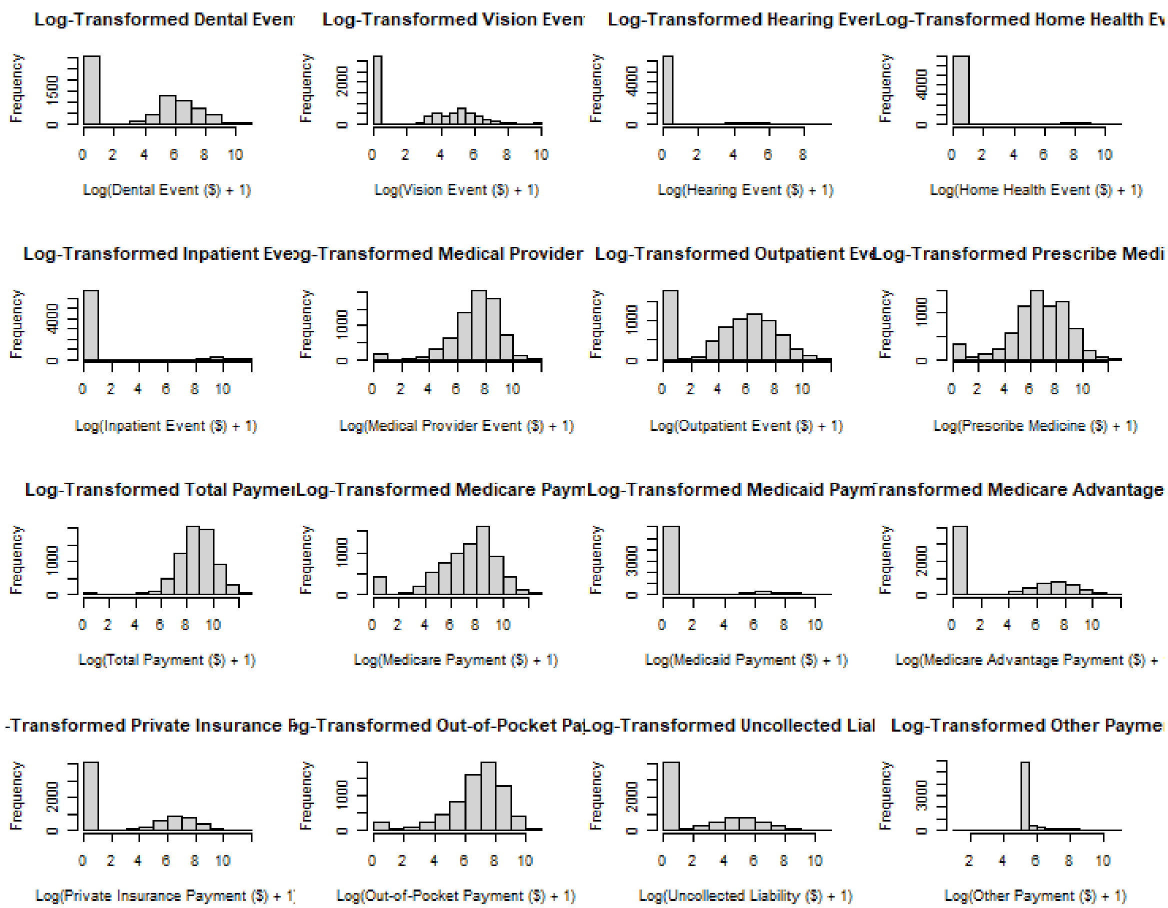
This is how the data looks like when plotting



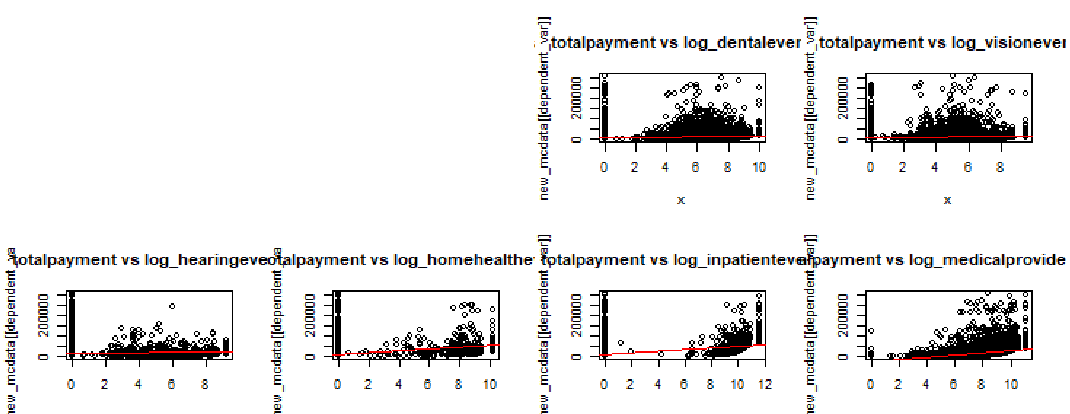
Since the data has some issues with linearity not being perfect, i.e medicareadvantagepayment vs outofpocketpayment, the data will have to be transformmed. To see where the issue lies, histograms are drawn:

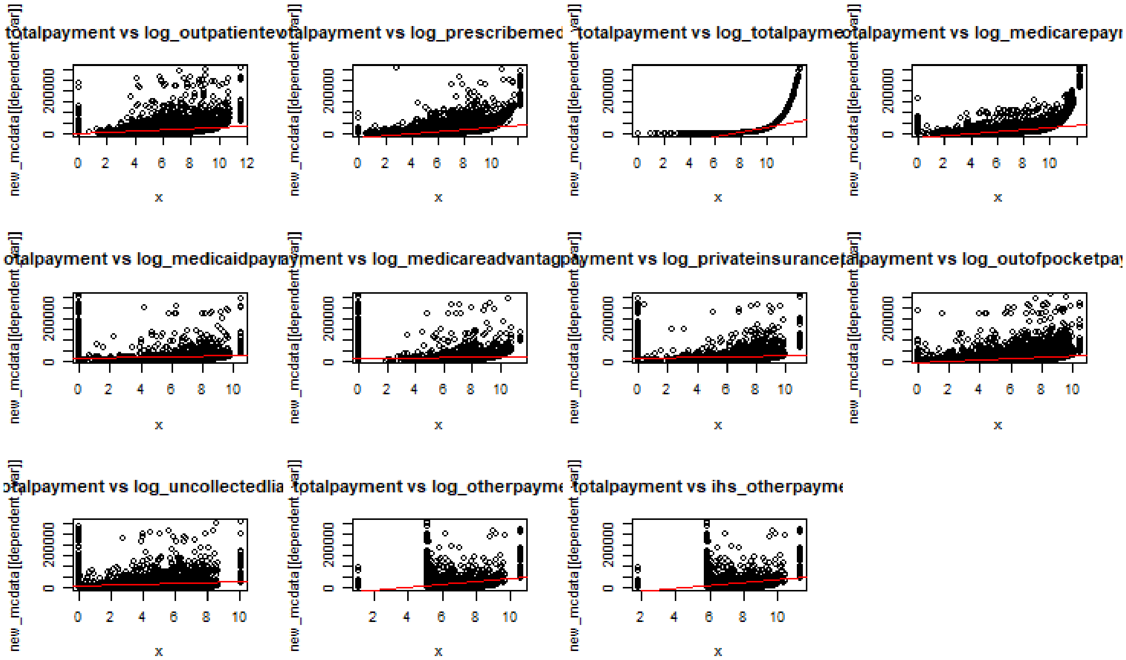


The issue happens to be an issue of the excess of zeros in the data which can be fixed with a log() function to help create a more normal distrbution. This is how the data looks like after applying the log() function:



After applying the log() function, the linearity can be checked again. It indicates that the function did help reduce the non-linearity between the variables. Even though it isn't perfect linearity, it can still be used in a regression model




totalpayment vs log_outpatientev
talpayment vs log_prescribemec
totalpayment vs log_totalpayme
otalpayment vs log_medicarepayı
otalpayment vs log_medicaidpayı
ayment vs log_medicareadvantag
ayment vs log_privateinsurance
talpayment vs log_outofpocketpa
otalpayment vs log_uncollectedlia
totalpayment vs log_otherpayme
totalpayment vs ihs_otherpayme
new_mcdata[[dependent_var]]
x
0
200000

- First OLS Model

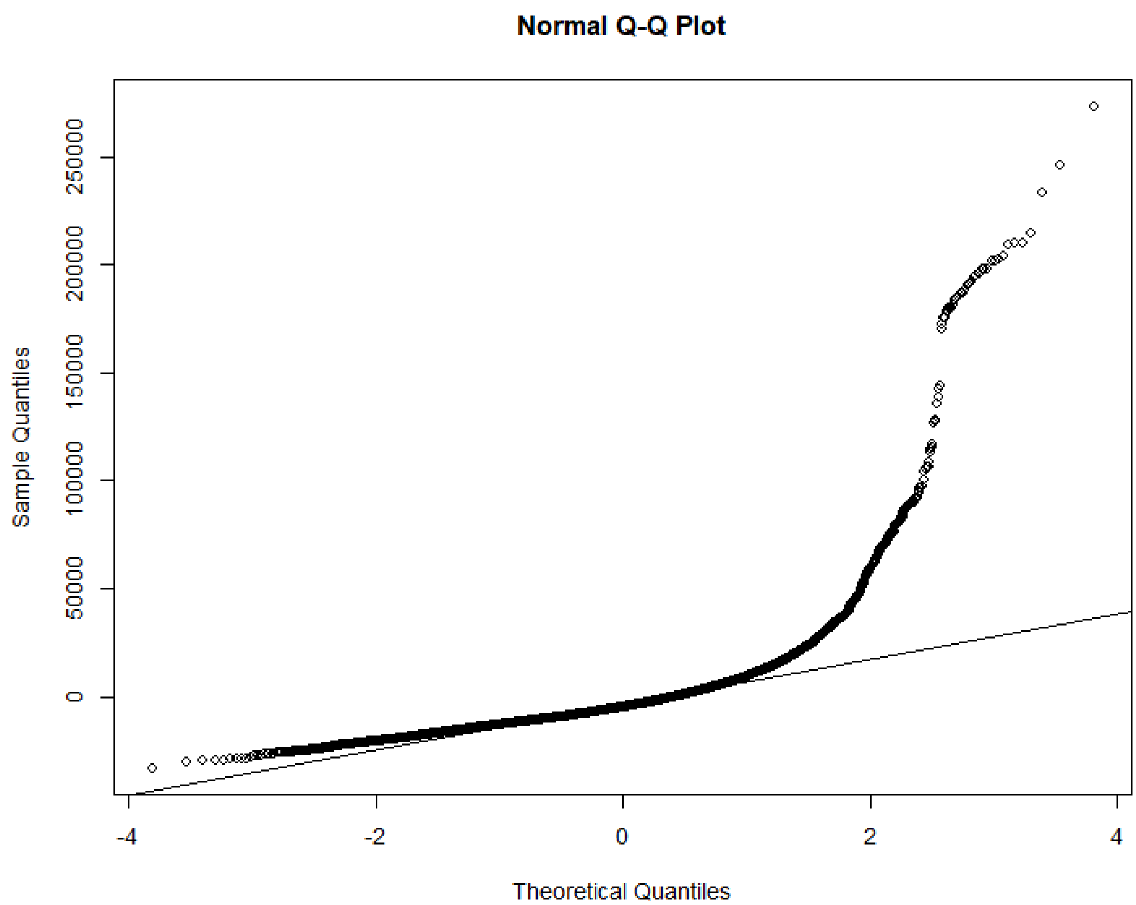
$$
\begin{aligned}
Totalpayment \\
&= B1\ sex\ +\ B2\ numberchroniccond\ +\ B3\ log\_dentalevent \\
&+\ B4\ log\_visionevent\ +\ B5\ log\_hearingevent \\
&+\ B6\ log\_medicalproviderevent\ + B7\ log\_outpatientevent \\
&+\ B8\ log\_prescribemedicine\ +\ B9\ log\_medicarepayment \\
&+\ B10\ log\_medicaidpayment\ +\ B11\ log\_medicareadvantagepayment \\
&+\ B12\ log\_privateinsurancepayment\ +\ B13\ log\_outofpocketpayment \\
&+\ B14\ log\_uncollectedliability\ +\ B15\ log\_otherpayment\ +\ B16\ age\_g1 \\
&+\ B17\ age\_g2\ + B18\ hisp\ +\ B19\ nonhispblack\ +\ B20 nonhispwhite \\
&+\ B21\ incomeless25
\end{aligned}
$$

Every variable was included in the model to see which variables are the most significant. Significance is based on the p-value of 0.05 or less. Not every categorical variable was included to prevent the dummy variable trap.

```
Call:
lm(formula = totalpayment ~ sex + numberchroniccond + log_dentalevent +
    log_visionevent + log_hearingevent + log_medicalproviderevent +
    log_outpatientevent + log_prescribemedicine + log_medicarepayment +
    log_medicaidpayment + log_medicareadvantagepayment + log_privateinsurancepayment +
    log_outofpocketpayment + log_uncollectedliability + log_otherpayment +
    age_g1 + age_g2 + hisp + nonhispblack + nonhispwhite + incomeless25,
    data = new_mcdata)

Residuals:
   Min     1Q Median     3Q    Max
-32973 -10507  -4759   3640 273324

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -56714.88    2405.32 -23.579  < 2e-16 ***
sex                             282.85     526.84   0.537 0.591362
numberchroniccond             -1528.16     406.12  -3.763 0.000169 ***
log_dentalevent                 -16.37      89.26  -0.183 0.854463
log_visionevent                -400.23     101.36  -3.949 7.93e-05 ***
log_hearingevent               -189.88     146.69  -1.294 0.195549
log_medicalproviderevent        976.79     218.71   4.466 8.08e-06 ***
log_outpatientevent             924.79      99.33   9.310  < 2e-16 ***
log_prescribemedicine          1696.46     168.37  10.076  < 2e-16 ***
log_medicarepayment            2152.06     151.95  14.163  < 2e-16 ***
log_medicaidpayment            1208.08     145.86   8.282  < 2e-16 ***
log_medicareadvantagepayment    681.58      95.17   7.162 8.72e-13 ***
log_privateinsurancepayment     649.42     114.90   5.652 1.65e-08 ***
log_outofpocketpayment          991.97     203.92   4.864 1.17e-06 ***
log_uncollectedliability        855.41     117.15   7.302 3.13e-13 ***
log_otherpayment               4094.30     301.70  13.571  < 2e-16 ***
age_g1                         5769.79     835.63   6.905 5.45e-12 ***
age_g2                          925.19     579.70   1.596 0.110534
hisp                            356.99    1410.10   0.253 0.800145
nonhispblack                   3702.02    1420.97   2.605 0.009198 **
nonhispwhite                    757.54    1185.66   0.639 0.522895
incomeless25                   -711.64     712.28  -0.999 0.317780
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

At the Alpha level of 0.05, the variables sex, log_dentalevent, log_hearingevent, age_g1, age_g2, hisp, nonhispwhite and incomeless25 are not significant and will be omitted from the model. The model also has an Adjusted R-Squared Value of 0.3168, which indicates how good a model can predict. Of course, this model can become better by omitting the insignificant variables and fixing further aspects of it.

Another aspect to check in the model is the residuals of the model. Based on this following graph, the model is an okay fit with some outliers at the tails. The upper tail has the most outliers, but this is probably due to the huge outliers in many of the different forms of payment in our data. An additional reason could be due to the existence of zero inflated values withing the model which we'll talk about later.



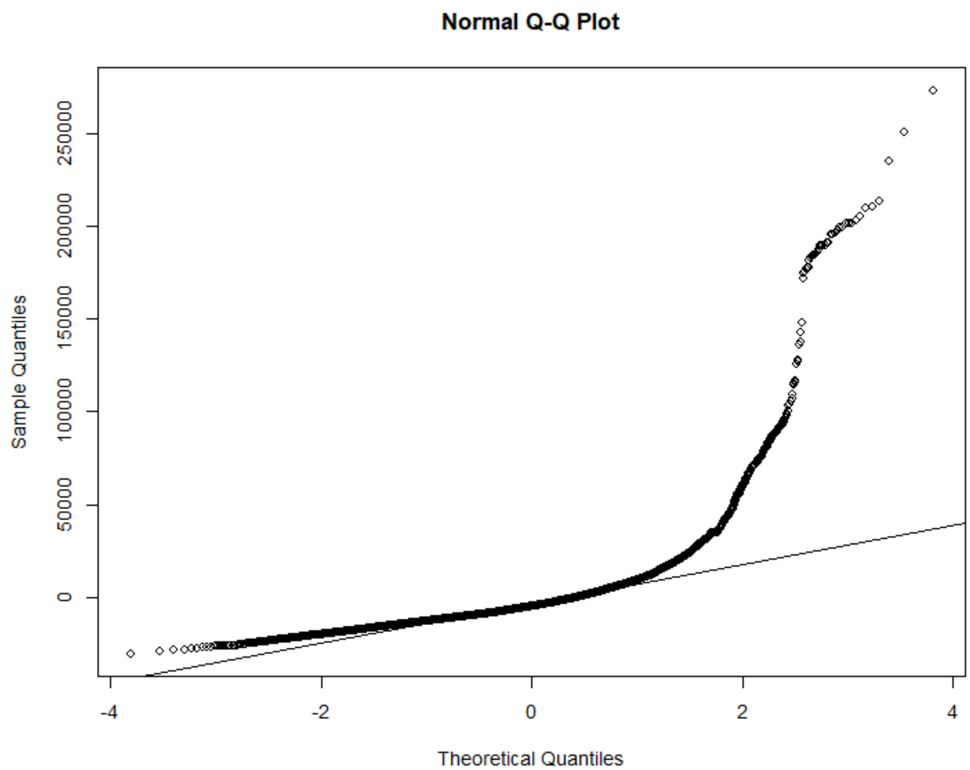
- Second OLS Model

$$Totalpayment = B1\, numberchroniccond + B2\, log\_visionevent + B3\, log\_medicalproviderevent + B4\, log\_outpatientevent + B5\, log\_prescribemedicine + B6\, log\_medicarepayment + B7\, log\_medicaidpayment + B8\, log\_medicareadvantagepayment + B9\, log\_privateinsurancepayment + B10\, log\_outofpocketpayment + B11\, log\_uncollectedliability + B12\, log\_otherpayment + B13\, nonhispblack$$

```
Call:
lm(formula = totalpayment ~ numberchroniccond + log_visionevent +
    log_medicalproviderevent + log_outpatientevent + log_prescribemedicine +
    log_medicarepayment + log_medicaidpayment + log_medicareadvantagepayment +
    log_privateinsurancepayment + log_outofpocketpayment + log_uncollectedliability +
    log_otherpayment + nonhispblack, data = new_mcdata)

Residuals:
   Min     1Q Median     3Q    Max
-30409 -10597  -4972   3612 273432

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -53682.23    1987.97 -27.003  < 2e-16 ***
numberchroniccond             -1973.69     399.73  -4.938 8.08e-07 ***
log_visionevent                -485.69     100.57  -4.829 1.40e-06 ***
log_medicalproviderevent        983.90     218.75   4.498 6.97e-06 ***
log_outpatientevent             972.11      99.28   9.792  < 2e-16 ***
log_prescribemedicine          1815.44     166.66  10.893  < 2e-16 ***
log_medicarepayment            2071.99     151.28  13.696  < 2e-16 ***
log_medicaidpayment            1396.16     134.64  10.370  < 2e-16 ***
log_medicareadvantagepayment    630.73      93.15   6.771 1.38e-11 ***
log_privateinsurancepayment     593.82     110.71   5.364 8.41e-08 ***
log_outofpocketpayment          925.35     192.32   4.811 1.53e-06 ***
log_uncollectedliability        882.07     117.10   7.533 5.56e-14 ***
log_otherpayment               4073.47     300.91  13.537  < 2e-16 ***
nonhispblack                   3587.75     901.21   3.981 6.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22030 on 7309 degrees of freedom
Multiple R-squared:  0.3138,    Adjusted R-squared:  0.3125
F-statistic: 257.1 on 13 and 7309 DF,  p-value: < 2.2e-16
```

After adjusting the variables in the model, all the variables seem to be significant to totalpayment. They are significant at the Alpha level of 0.05, which means that they have a 95% chance of predicting the variable totalpayment. Even though the model was created to become better, the Adjusted R-Squared decreased to 0.3125.

This is the residual plot of the model after the adjustment. There’s not much difference in the outliers. Since the outliers are causing a problem, another solution called Box-Cox Transformation will be used.



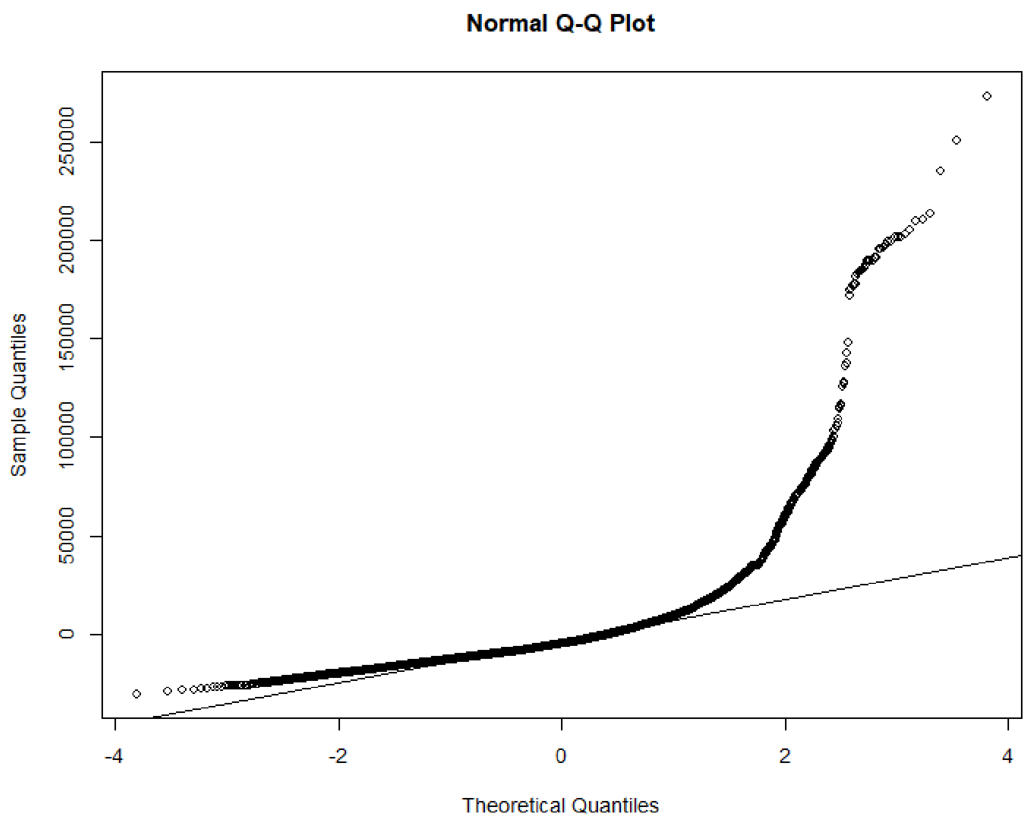
- Third OLS Model

$$
\begin{aligned}
(Totalpayment\hat{\ }best.lam) &= B0 + B1\, numberchroniccond + B2\ \log(visionevent) \\
&+ B3\ \log(medicalproviderevent) + B4\ \log(outpatientevent) \\
&+ B5\ \log(prescribemedicine) + B6\ \log(medicarepayment) \\
&+ B7\ \log(medicaidpayment) + B8\ \log(medicareadvantagepayment) \\
&+ B9 * \log(privateinsurancepayment) \\
&+ B10\ \log(outofpocketpayment) + B11\ \log(uncollectedliability) \\
&+ B12\ \log(otherpayment) + B13\ nonhispblack + \varepsilon
\end{aligned}
$$

If you want to understand why totalpayment^best.lam, you can watch Math et al's video on YouTube about the Box-Cox Transformation linked in the references of this paper. This is one of many possible solutions to the problem of zero-inflated values within our dataset. These are the results:

```
Call:
lm(formula = (totalpayment^best.lam) ~ numberchroniccond + log_visionevent +
    log_medicalproviderevent + log_outpatientevent + log_prescribemedicine +
    log_medicarepayment + log_medicaidpayment + log_medicareadvantagepayment +
    log_privateinsurancepayment + log_outofpocketpayment + log_uncollectedliability +
    log_otherpayment + nonhispblack, data = new_mcdata)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1956 -0.6904 -0.2566  0.4528  6.3602

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -1.141337   0.091697 -12.447  < 2e-16 ***
numberchroniccond            -0.062365   0.018438  -3.382 0.000722 ***
log_visionevent              -0.016764   0.004639  -3.614 0.000304 ***
log_medicalproviderevent      0.225721   0.010090  22.371  < 2e-16 ***
log_outpatientevent           0.088344   0.004579  19.293  < 2e-16 ***
log_prescribemedicine         0.195800   0.007687  25.470  < 2e-16 ***
log_medicarepayment           0.181348   0.006978  25.989  < 2e-16 ***
log_medicaidpayment           0.150238   0.006210  24.192  < 2e-16 ***
log_medicareadvantagepayment  0.089258   0.004297  20.774  < 2e-16 ***
log_privateinsurancepayment   0.080264   0.005107  15.717  < 2e-16 ***
log_outofpocketpayment        0.198580   0.008871  22.385  < 2e-16 ***
log_uncollectedliability      0.056850   0.005401  10.526  < 2e-16 ***
log_otherpayment              0.214636   0.013880  15.464  < 2e-16 ***
nonhispblack                  0.130993   0.041569   3.151 0.001633 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.016 on 7309 degrees of freedom
Multiple R-squared:  0.736,	Adjusted R-squared:  0.7355
F-statistic:  1567 on 13 and 7309 DF,  p-value: < 2.2e-16
```

All our model variables are still significant to the variable total payment. The only difference is that now our Adjusted R-Squared has the value of 0.7355 which indicates a much better fit this model is. It isn’t perfect but it is better than what we calculated in our first model. Our residuals though haven’t changed a lot either from our first model. Other solutions like the Poisson Regression could help fix the Adjust R-Squared of the model and give it an even better fit, but there’s implications that come with it. One of the implications is that the interpretation of the results is different due to the different type of regression being used. Due to that reason, this model will be chosen as an explanation of the variable totalpayment.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
0
50000
100000
150000
200000
250000
-4
-2
0
2
4

Conclusion

In conclusion, our final model is:

$$Transformed\ Total\ Payment\ =$$

$$-1.11596 + (-0.06229 * Number\ of\ Chronic\ Conditions) + (-0.01670 * Log\ of\ Vision\ Events) + (0.22373 * Log\ of\ Medical\ Provider\ Events) + (0.08858 * Log\ of\ Outpatient\ Events) + (0.19537 * Log\ of\ Prescribed\ Medicine) + (0.18112 * Log\ of\ Medicare\ Payment) + (0.14992 * Log\ of\ Medicaid\ Payment) + (0.08908 * Log\ of\ Medicare\ Advantage\ Payment) + (0.08028 * Log\ of\ Private\ Insurance\ Payment) + (0.19716 * Log\ of\ Out-of-Pocket\ Payment) + (0.05725 * Log\ of\ Uncollected\ Liability) + (0.21521 * Log\ of\ Other\ Payment) + (0.13121 * Non-Hispanic\ Black)$$

Where each unit increase in each variable causes the Transformed Total Payment to increase by the coefficient number next to it. Even though a variable like Number of Chronic Conditions has a negative effect on Total Payment, may seem counterintuitive, but the other variables that explain the payment explain why this number would be less. Another counterintuitive variable is Vision Events and the lowering of the Total Payment, this could be possibly due to the relatively lower cost associated with it. Meanwhile the rest of the variables, except Non-Hispanic Black, have a positive effect on the Total Payment. This suggests that the higher their values are, the higher the Total payment is. The last variable in our model, Non-Hispanic Black suggests that beneficiaries who identify as Black tend to have higher healthcare expenditure in comparison to all other races. This could be in relation to socio-economic and higher numbers of chronic conditions (Gaskin et al., 2021, p.144)

My model highlights the importance of factors like insurance coverage, chronic conditions and socioeconomic factors may play a role into beneficiaries' healthcare expenditures. By using the model to calculate their impact on expenditure, policies can be improved and preventative healthcare strategies can be implemented for those associated with high costs.

Resources

- Kamb, L. (2022, March 30). WA expanding health care options for undocumented immigrants. The Seattle Times. https://www.seattletimes.com/seattle-news/politics/wa-expanding-health-care-options-for-undocumented-immigrants/
- U.S. Department of Health and Human Services. (n.d.). Who is eligible for Medicaid? HHS.gov. Retrieved June 14, 2024, from https://www.hhs.gov/answers/medicare-and-medicaid/who-is-eligible-for-medicaid/index.html
- Centers for Medicare & Medicaid Services. (2021, November 17). Medicare Current Beneficiary Survey Cost & Supplement [Data set]. Data.cms.gov. https://data.cms.gov/medicare-current-beneficiary-survey-mcbs/medicare-current-beneficiary-survey-cost-supplement
- California Health Advocates. (2023, March 3). Have you been contacted for the Medicare Current Beneficiary Survey? Verify your participation & prevent fraud. California Health Advocates. https://cahealthadvocates.org/have-you-been-contacted-for-the-medicare-current-beneficiary-survey-verify-your-participation-prevent-fraud/
- Johnston, K. J., Pinska, L., Munro, H., Parker, D., Nguyen, H., & Quiter, E. (2019). Determining out-of-pocket expenditure towards care in the Medicare Current Beneficiary Survey. PLOS ONE. https://doi.org/10.1371/journal.pone.0222539
- Xu, W. Y., Fonseca, V., Finch, M. D., Kim, M. J., & Frett, B. (2021). Quality, health, and spending in Medicare Advantage and traditional Medicare. The American Journal of Managed Care (AMJC). https://doi.org/10.37765/ajmc.2021.88581
- Math et al. (2022, January 24). The derivative of a logarithm: why the log derivative is one over x [Video]. YouTube. https://www.youtube.com/watch?v=vGOpEpjz2Ks
- Gaskin, D. J., Dinwiddie, G. Y., & Chan, K. S. (2021). Racial and ethnic disparities in health care access and utilization under the Affordable Care Act. Medical Care Research and Review. https://doi.org/10.1177/1077558720965111

Appendix

- Graph 1 – Correlation Plot

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1		sex	numberch	dentaleve	visionever	hearingev	homeheal	inpatiente	medicalpı	outpatient	prescribe	totalpaym	medicarep	medicaidp	medicarea	privateins	outofpock	uncollecte	otherpayn	age_g1	age_g2	age_g3	nonhispw	nonhispbl	hisp	otherrace	incomeles	incomemoreequal25
2	sex	1	-0.08481	0.003845	-0.02498	0.02266	0.016331	0.011334	-0.01293	0.005618	-0.00288	-0.00462	0.000324	-0.02711	-0.02263	-0.00857	0.008659	0.018342	0.03337	0.059982	0.001863	-0.04671	-0.00781	-0.00138	-0.00875	0.029131	-0.08969	0.089692
3	numberch	-0.08481	1	0.015189	0.03769	0.028044	0.051705	0.066121	0.155007	0.078531	0.097703	0.160303	0.112101	0.049805	0.084969	0.064641	0.136769	0.022531	0.053907	-0.09225	-0.08986	0.155602	-0.03334	0.006276	0.026857	0.020218	0.028593	-0.02859
4	dentalever	0.003845	0.015189	1	0.025254	0.022251	-0.01653	-0.00458	0.052544	-0.01473	0.043057	0.11452	0.013402	-0.01042	0.075588	0.069284	0.40719	0.029985	0.04746	-0.07935	0.052366	0.008989	0.058278	-0.05371	-0.02598	-0.00711	-0.12433	0.124332
5	visionever	-0.02498	0.03769	0.025254	1	0.015885	0.004111	0.021352	0.122079	0.019932	0.001439	0.097452	0.065393	0.060713	0.026304	0.053189	0.100801	0.072346	0.013736	-0.04813	-0.03437	0.069133	0.037789	-0.03964	-0.02145	0.008106	-0.0563	0.056301
6	hearingev	0.02266	0.028044	0.022251	0.015885	1	-0.00948	-0.00202	0.066004	0.005334	-0.01673	0.044685	-0.0093	-0.00297	0.005342	0.020441	0.284601	0.044573	0.009345	-0.05796	-0.01814	0.060864	0.044086	-0.04062	-0.02255	-0.00142	-0.07721	0.077208
7	homeheal	0.016331	0.051705	-0.01653	0.004111	-0.00948	1	0.158091	0.128296	0.069483	0.021717	0.205969	0.192065	0.110698	-0.02286	0.057905	0.181271	0.044889	0.076893	-0.00418	-0.07027	0.070793	0.004812	-0.00393	0.000112	-0.00435	0.032334	-0.03233
8	inpatiente	0.011334	0.066121	-0.00458	0.021352	-0.00202	0.158091	1	0.227518	0.137295	0.06596	0.48575	0.362194	0.110945	0.413131	0.132454	0.133484	0.12992	0.060359	0.028839	-0.02953	0.006841	-0.00573	0.028113	-0.00876	-0.01422	0.007766	-0.00777
9	medicalpr	-0.01293	0.155007	0.052544	0.122079	0.066004	0.128296	0.227518	1	0.293597	0.112392	0.523814	0.364893	0.240869	0.257408	0.285197	0.367084	0.268546	0.047379	0.001126	-0.05133	0.048583	0.063434	-0.05432	-0.02738	-0.01449	-0.07307	0.073074
10	outpatient	0.005618	0.078531	-0.01473	0.019932	0.005334	0.069483	0.137295	0.293597	1	0.073626	0.446389	0.358567	0.146392	0.191231	0.243438	0.184009	0.214915	0.020265	0.040977	-0.02583	-0.00581	0.021255	0.002544	-0.02514	-0.01063	-0.01149	0.011492
11	prescriber	-0.00288	0.097703	0.043057	0.001439	-0.01673	0.021717	0.06596	0.112392	0.073626	1	0.717969	0.759649	0.106802	0.036055	0.091718	0.184673	0.074038	0.3934	0.128842	-0.0343	-0.06345	-0.02847	0.043459	-0.0031	0.002146	0.069493	-0.06949
12	totalpaym	-0.00462	0.160303	0.11452	0.097452	0.044685	0.205969	0.48575	0.523814	0.446389	0.717969	1	0.892092	0.238118	0.300737	0.333361	0.40513	0.258137	0.327136	0.10045	-0.06244	-0.01509	0.014764	0.017017	-0.02704	-0.01459	0.013673	-0.01367
13	medicarep	0.000324	0.112101	0.013402	0.065393	-0.0093	0.192065	0.362194	0.364893	0.358567	0.759649	0.892092	1	0.180624	-0.02267	0.147203	0.220298	0.123654	0.195373	0.112158	-0.05366	-0.03231	-0.0031	0.032799	-0.02006	-0.01015	0.066196	-0.0662
14	medicaidp	-0.02711	0.049805	-0.01042	0.060713	-0.00297	0.110698	0.110945	0.240869	0.146392	0.106802	0.238118	0.180624	1	0.022175	-0.0272	-0.07011	-0.02588	0.013011	0.170881	-0.05775	-0.07234	-0.12248	0.065191	0.101808	0.013833	0.219815	-0.21982
15	medicarea	-0.02263	0.084969	0.075588	0.026304	0.005342	-0.02286	0.413131	0.257408	0.191231	0.036055	0.300737	-0.02267	0.022175	1	-0.05558	0.05878	0.060902	0.023354	0.012874	-0.02173	0.011287	-0.03536	0.025508	0.027969	-0.00297	0.003278	-0.00328
16	privateins	-0.00857	0.064641	0.069284	0.053189	0.020441	0.057905	0.132454	0.285197	0.243438	0.091718	0.333361	0.147203	-0.0272	-0.05558	1	0.214786	0.197294	0.082362	0.011088	0.000435	-0.00872	0.070913	-0.04448	-0.04765	-0.0145	-0.1093	0.109304
17	outofpock	0.008659	0.136769	0.40719	0.100801	0.284601	0.181271	0.133484	0.367084	0.184009	0.184673	0.40513	0.220298	-0.07011	0.05878	0.214786	1	0.1894	0.151775	-0.08777	-0.025	0.089792	0.14449	-0.09564	-0.10082	-0.01777	-0.20865	0.208652
18	uncollecte	0.018342	0.022531	0.029985	0.072346	0.044573	0.044889	0.12992	0.268546	0.214915	0.074038	0.258137	0.123654	-0.02588	0.060902	0.197294	0.1894	1	0.102985	0.002325	-0.0199	0.01742	0.04611	-0.00466	-0.04485	-0.02277	-0.07211	0.072111
19	otherpayn	0.03337	0.053907	0.04746	0.013736	0.009345	0.076893	0.060359	0.047379	0.020265	0.3934	0.327136	0.195373	0.013011	0.023354	0.082362	0.151775	0.102985	1	0.009774	-0.01934	0.011303	0.027053	-0.00665	-0.03071	-0.00213	-0.05326	0.053261
20	age_g1	0.059982	-0.09225	-0.07935	-0.04813	-0.05796	-0.00418	0.028839	0.001126	0.040977	0.128842	0.10045	0.112158	0.170881	0.012874	0.011088	-0.08777	0.002325	0.009774	1	-0.33828	-0.42304	-0.13063	0.147869	0.01474	0.03889	0.364699	-0.3647
21	age_g2	0.001863	-0.08986	0.052366	-0.03437	-0.01814	-0.07027	-0.02953	-0.05133	-0.02583	-0.0343	-0.06244	-0.05366	-0.05775	-0.02173	0.000435	-0.025	-0.0199	-0.01934	-0.33828	1	-0.70959	0.010832	-0.02448	0.018724	-0.01425	-0.1726	0.172597
22	age_g3	-0.04671	0.155602	0.008989	0.069133	0.060864	0.070793	0.006841	0.048583	-0.00581	-0.06345	-0.01509	-0.03231	-0.07234	0.011287	-0.00872	0.089792	0.01742	0.011303	-0.42304	-0.70959	1	0.087384	-0.08715	-0.02907	-0.01539	-0.10688	0.106882
23	nonhispw	-0.00781	-0.03334	0.058278	0.037789	0.044086	0.004812	-0.00573	0.063434	0.021255	-0.02847	0.014764	-0.0031	-0.12248	-0.03536	0.070913	0.14449	0.04611	0.027053	-0.13063	0.010832	0.087384	1	-0.56573	-0.58533	-0.40506	-0.30803	0.308033
24	nonhispbl	-0.00138	0.006276	-0.05371	-0.03964	-0.04062	-0.00393	0.028113	-0.05432	0.002544	0.043459	0.017017	0.032799	0.065191	0.025508	-0.04448	-0.09564	-0.00466	-0.00665	0.147869	-0.02448	-0.08715	-0.56573	1	-0.10832	-0.07496	0.200595	-0.20059
25	hisp	-0.00875	0.026857	-0.02598	-0.02145	-0.02255	0.000112	-0.00876	-0.02738	-0.02514	-0.0031	-0.02704	-0.02006	0.101808	0.027969	-0.04765	-0.10082	-0.04485	-0.03071	0.01474	0.018724	-0.02907	-0.58533	-0.10832	1	-0.07756	0.203257	-0.20326
26	otherrace	0.029131	0.020218	-0.00711	0.008106	-0.00142	-0.00435	-0.01422	-0.01449	-0.01063	0.002146	-0.01459	-0.01015	0.013833	-0.00297	-0.0145	-0.01777	-0.02277	-0.00213	0.03889	-0.01425	-0.01539	-0.40506	-0.07496	-0.07756	1	0.058263	-0.05826
27	incomeles	-0.08969	0.028593	-0.12433	-0.0563	-0.07721	0.032334	0.007766	-0.07307	-0.01149	0.069493	0.013673	0.066196	0.219815	0.003278	-0.1093	-0.20865	-0.07211	-0.05326	0.364699	-0.1726	-0.10688	-0.30803	0.200595	0.203257	0.058263	1	-1
28	incomemo	0.089692	-0.02859	0.124332	0.056301	0.077208	-0.03233	-0.00777	0.073074	0.011492	-0.06949	-0.01367	-0.0662	-0.21982	-0.00328	0.109304	0.208652	0.072111	0.053261	-0.3647	0.172597	0.106882	0.308033	-0.20059	-0.20326	-0.05826	-1	1

- Comparison of OLS models

	Dependent variable:		
	totalpayment		
	(1)	(2)	(3)
sex	282.853		
	(526.839)		
numberchroniccond	-1,528.159***	-1,973.691***	-1,973.691***
	(406.121)	(399.728)	(399.728)
log_dentalevent	-16.373		
	(89.261)		
log_visionevent	-400.230***	-485.691***	-485.691***
	(101.355)	(100.572)	(100.572)
log_hearingevent	-189.881		
	(146.688)		
log_medicalproviderevent	976.790***	983.899***	983.899***
	(218.710)	(218.750)	(218.750)
log_outpatientevent	924.787***	972.108***	972.108***
	(99.330)	(99.276)	(99.276)
log_prescribemedicine	1,696.461***	1,815.444***	1,815.444***
	(168.373)	(166.661)	(166.661)
log_medicarepayment	2,152.065***	2,071.990***	2,071.990***
	(151.947)	(151.279)	(151.279)
log_medicaidpayment	1,208.082***	1,396.163***	1,396.163***
	(145.862)	(134.637)	(134.637)
log_medicareadvantagepayment	681.581***	630.726***	630.726***
	(95.166)	(93.149)	(93.149)
log_privateinsurancepayment	649.423***	593.822***	593.822***
	(114.903)	(110.713)	(110.713)
log_outofpocketpayment	991.968***	925.353***	925.353***
	(203.923)	(192.321)	(192.321)
log_otherpayment	4,094.299***	4,073.465***	4,073.465***
	(301.697)	(300.908)	(300.908)
age_g1	5,769.789***		
	(835.635)		
age_g2	925.195		
	(579.699)		
hisp	356.994		
	(1,410.102)		
nonhispblack	3,702.017***	3,587.755***	3,587.755***
	(1,420.965)	(901.215)	(901.215)
nonhispwhite	757.543		
	(1,185.663)		
incomeless25	-711.642		
	(712.283)		
Constant	-56,713.880***	-53,681.230***	-53,682.230***
	(2,405.316)	(1,987.973)	(1,987.973)
Observations	7,323	7,323	7,323
R2	0.319	0.314	0.314
Adjusted R2	0.317	0.313	0.313
Residual Std. Error	21,966.410 (df = 7301)	22,034.100 (df = 7309)	22,034.100 (df = 7309)
F Statistic	162.642*** (df = 21; 7301)	257.057*** (df = 13; 7309)	257.057*** (df = 13; 7309)

- Code:

```
# All the needed libraries
 library("stargazer")
 library("readxl")
 library("zoo")
 library("dplyr")
 library("lubridate")
 library("purrr")
 library("ggplot2")
 library("moments")
 library("car")
 library("lmtest")
 library("corrplot")
 library("fastDummies") #used for changing categorical data into dummy variables
 library("gt") #used for formatting tables nicely when knitting
 library("GGally")
 library("psych")
 library("moments") #for testing skewness
 library("nortest")
 #installing packages
 #install.packages("gt")
 #install.packages("fastDummies")
 #install.packages("GGally")
 #install.packages("corrplot")
 #install.packages("psych")
 #install.packages("nortest")
 library(caret)

#Setting working directory for the data
 setwd("D://Bellevue College//'24 Spring//ECON 400//Week 8 - Term Paper")
```

```
  mcdata <- read_excel("cspuf2021.xls")



#In this section, I'll be formatting the dataset for further usage


  #renaming variables for easy use https://www.geeksforgeeks.org/how-to-rename-multiple-
columns-in-r/
    new_names <- c("age", "sex", "race",
                "income", "numberchroniccond", "dentalevent", "visionevent",
                "hearingevent", "homehealthevent", "inpatientevent",
                "medicalproviderevent","outpatientevent", "prescribemedicine",
                "totalpayment", "medicarepayment","medicaidpayment",
                "medicareadvantagepayment", "privateinsurancepayment", "outofpocketpayment",
                "uncollectedliability", "otherpayment")
    names(mcdata)<- new_names

  #Starting with the age variable
    # Changing categorical variables to separate dummy variables that represent each group
    # Create new columns that represent each new group
      mcdata$age_g1<-c(0)
      mcdata$age_g2<-c(0)
      mcdata$age_g3<-c(0)

  #to delete the mistake columns
    #mydata2 = select(mcdata, -28, -29)
    #mcdata<-mydata2
    #rm(mydata2)

  #Dividing the age variable amongst the 3 new columns
    mcdata$age_g1 <- ifelse(mcdata$age == 1, 1, 0)
```

```
  mcdata$age_g2 <- ifelse(mcdata$age == 2, 1, 0)
  mcdata$age_g3 <- ifelse(mcdata$age == 3, 1, 0)

#sex variable
  # Changing categorical variables to 0 & 1 dummy variables within the same column
  # value of 1 = male, value of 0 = female
    mcdata$sex <- ifelse(mcdata$sex == 1, 1, 0)
    print(head(mcdata$sex, n=5))

#race variable:
  #1:Non-Hispanic white
  #2:Non-Hispanic black
  #3:Hispanic
  #4:Other
  # Create new columns that represent each new group
    mcdata$nonhispwhite<-c(0)
    mcdata$nonhispblack<-c(0)
    mcdata$hisp<-c(0)
    mcdata$otherrace<-c(0)
  #Assigning dummy variables to each column
    mcdata$nonhispwhite <- ifelse(mcdata$race == 1, 1, 0)
    mcdata$nonhispblack <- ifelse(mcdata$race == 2, 1, 0)
    mcdata$hisp <- ifelse(mcdata$race == 3, 1, 0)
    mcdata$otherrace <- ifelse(mcdata$race == 4, 1, 0)

#income variable
  # 1:<$25,000
  # 2:>=$25,000
  # Create new columns that represent each new group
    mcdata$incomeless25<-c(0)
```

```
  mcdata$incomemoreequal25<-c(0)
 #Using if else statement to assign values
  mcdata$incomeless25 <- ifelse(mcdata$income == 1, 1, 0)
  mcdata$incomemoreequal25 <- ifelse(mcdata$income == 2, 1, 0)

# Removing any possible N/A values
  mcdata <- na.omit(mcdata)

# Dropping the categorical columns after switching them to dummy columns
  #https://www.listendata.com/2015/06/r-keep-drop-columns-from-data-frame.html
   drops <- c("age", "race", "income")
  # Creating a new data set after the changes
   new_mcdata= mcdata[ , !(names(mcdata)%in% drops)]


#Visualizing/viewing the data

   # Checking for correlation then saving it as a csv file due to the 22*22 design
   cor_matrix <- cor(select_if(new_mcdata, is.numeric))
   cor_matrix
   write.csv(cor_matrix, "cor_matrix.csv", row.names = TRUE)



   #Testing for linearity of total payment vs each variable
   par(mfrow = c(4, 4))
   dependent_var <- "totalpayment"
   independent_vars <- setdiff(names(new_mcdata), dependent_var)

   # Create a function to plot linearity
   par(mfrow = c(4,4))
```

```
plot_linearity <- function(x, y, plot_title) {
  plot(y ~ x, data = new_mcdata, main = plot_title, xlab = deparse(substitute(x)), ylab = deparse(substitute(y)))
  abline(lm(y ~ x, data = new_mcdata), col = "red")
}
# Loop through each independent variable and plot linearity
for (var in independent_vars) {
  plot_title <- paste(dependent_var, "vs", var)
  plot_linearity(new_mcdata[[var]], new_mcdata[[dependent_var]], plot_title)
}

#Viewing issue with linearity via histograms
#Histograms
#https://www.geeksforgeeks.org/histograms-in-r-language/
#https://bookdown.org/dli/rguide/histogram.html
# Dental event
hist(new_mcdata$dentalevent, main = "Dental Event", xlab = "Dental Event ($)")
# Vision event
hist(new_mcdata$visionevent, main = "Vision Event", xlab = "Vision Event ($)")
# Hearing event
hist(new_mcdata$hearingevent, main = "Hearing Event", xlab = "Hearing Event ($)")
# Home health event
hist(new_mcdata$homehealthevent, main = "Home Health Event", xlab = "Home Health Event ($)")
# Inpatient event
hist(new_mcdata$inpatientevent, main = "Inpatient Event", xlab = "Inpatient Event ($)")
# Medical provider event
hist(new_mcdata$medicalproviderevent, main = "Medical Provider Event", xlab = "Medical Provider Event ($)")
# Outpatient event
```

hist(new_mcdata$outpatientevent, main = "Outpatient Event", xlab = "Outpatient Event ($)")

Prescribe medicine

hist(new_mcdata$prescribemedicine, main = "Prescribe Medicine", xlab = "Prescribe Medicine ($)")

Total payment

hist(new_mcdata$totalpayment, main = "Total Payment", xlab = "Total Payment ($)")

Medicare payment

hist(new_mcdata$medicarepayment, main = "Medicare Payment", xlab = "Medicare Payment ($)")

Medicaid payment

hist(new_mcdata$medicaidpayment, main = "Medicaid Payment", xlab = "Medicaid Payment ($)")

Medicare Advantage payment

hist(new_mcdata$medicareadvantagepayment, main = "Medicare Advantage Payment", xlab = "Medicare Advantage Payment ($)")

Private insurance payment

hist(new_mcdata$privateinsurancepayment, main = "Private Insurance Payment", xlab = "Private Insurance Payment ($)")

Out-of-pocket payment

hist(new_mcdata$outofpocketpayment, main = "Out-of-Pocket Payment", xlab = "Out-of-Pocket Payment ($)")

Uncollected liability

hist(new_mcdata$uncollectedliability, main = "Uncollected Liability", xlab = "Uncollected Liability ($)")

Other payment

hist(new_mcdata$otherpayment, main = "Other Payment", xlab = "Other Payment ($)")

#Numerical data is skewed to the right due to most of them having one heavy outlier

#Trying to eliminate skewness by using either log or sqrt forms

+1 is added to handle the 0 and any possible negative values

#Sqrt was tested, but didn't handle the skewness issue so I stuck with taking the log of all the payments that were in $

Dental event

new_mcdata$log_dentalevent <- log(new_mcdata$dentalevent + 1)

hist(new_mcdata$log_dentalevent, main = "Log-Transformed Dental Event", xlab = "Log(Dental Event ($) + 1)")

Vision event

new_mcdata$log_visionevent <- log(new_mcdata$visionevent + 1)

hist(new_mcdata$log_visionevent, main = "Log-Transformed Vision Event", xlab = "Log(Vision Event ($) + 1)")

Hearing event

new_mcdata$log_hearingevent <- log(new_mcdata$hearingevent + 1)

hist(new_mcdata$log_hearingevent, main = "Log-Transformed Hearing Event", xlab = "Log(Hearing Event ($) + 1)")

Home health event

new_mcdata$log_homehealthevent <- log(new_mcdata$homehealthevent + 1)

hist(new_mcdata$log_homehealthevent, main = "Log-Transformed Home Health Event", xlab = "Log(Home Health Event ($) + 1)")

Inpatient event

new_mcdata$log_inpatientevent <- log(new_mcdata$inpatientevent + 1)

hist(new_mcdata$log_inpatientevent, main = "Log-Transformed Inpatient Event", xlab = "Log(Inpatient Event ($) + 1)")

Medical provider event

new_mcdata$log_medicalproviderevent <- log(new_mcdata$medicalproviderevent + 1)

hist(new_mcdata$log_medicalproviderevent, main = "Log-Transformed Medical Provider Event", xlab = "Log(Medical Provider Event ($) + 1)")

Outpatient event

new_mcdata$log_outpatientevent <- log(new_mcdata$outpatientevent + 1)

hist(new_mcdata$log_outpatientevent, main = "Log-Transformed Outpatient Event", xlab = "Log(Outpatient Event ($) + 1)")

Prescribe medicine

new_mcdata$log_prescribemedicine <- log(new_mcdata$prescribemedicine + 1) hist(new_mcdata$log_prescribemedicine, main = "Log-Transformed Prescribe Medicine", xlab = "Log(Prescribe Medicine ($) + 1)")

Total payment

new_mcdata$log_totalpayment <- log(new_mcdata$totalpayment + 1)

hist(new_mcdata$log_totalpayment, main = "Log-Transformed Total Payment", xlab = "Log(Total Payment ($) + 1)")

Medicare payment

new_mcdata$log_medicarepayment <- log(new_mcdata$medicarepayment + 1)

hist(new_mcdata$log_medicarepayment, main = "Log-Transformed Medicare Payment", xlab = "Log(Medicare Payment ($) + 1)")

Medicaid payment

new_mcdata$log_medicaidpayment <- log(new_mcdata$medicaidpayment + 1)

hist(new_mcdata$log_medicaidpayment, main = "Log-Transformed Medicaid Payment", xlab = "Log(Medicaid Payment ($) + 1)")

Medicare Advantage payment

new_mcdata$log_medicareadvantagepayment <- log(new_mcdata$medicareadvantagepayment + 1)

hist(new_mcdata$log_medicareadvantagepayment, main = "Log-Transformed Medicare Advantage Payment", xlab = "Log(Medicare Advantage Payment ($) + 1)")

Private insurance payment

new_mcdata$log_privateinsurancepayment <- log(new_mcdata$privateinsurancepayment + 1)

hist(new_mcdata$log_privateinsurancepayment, main = "Log-Transformed Private Insurance Payment", xlab = "Log(Private Insurance Payment ($) + 1)")

Out-of-pocket payment

new_mcdata$log_outofpocketpayment <- log(new_mcdata$outofpocketpayment + 1)

hist(new_mcdata$log_outofpocketpayment, main = "Log-Transformed Out-of-Pocket Payment", xlab = "Log(Out-of-Pocket Payment ($) + 1)")

Uncollected liability

new_mcdata$log_uncollectedliability <- log(new_mcdata$uncollectedliability + 1)

hist(new_mcdata$log_uncollectedliability, main = "Log-Transformed Uncollected Liability", xlab = "Log(Uncollected Liability ($) + 1)")

Other payment

```
new_mcdata$otherpayment <- new_mcdata$otherpayment + abs(min(new_mcdata$otherpayment)) + 1
new_mcdata$log_otherpayment <- log(new_mcdata$otherpayment)


new_mcdata$ihs_otherpayment <- asinh(new_mcdata$otherpayment)
hist(new_mcdata$log_otherpayment, main = "Log-Transformed Other Payment", xlab = "Log(Other Payment ($) + 1)")


#Extracting the column and rows separately from the dim() function
dims<-dim(new_mcdata)
num_rows<-dims[1]
num_columns<-dims[2]
#Printing each one separately
print_rows<- print(paste0("Number of rows: ", num_rows))
print_columns<- print(paste0("Number of columns: ", num_columns))


#Trying to eliminate skewness by using either log or sqrt forms
# +1 is added to handle the 0 and any possible negative values
#Sqrt was tested, but didn't handle the skewness issue so I stuck with taking the log of all the payments that were in $
# Dental event
new_mcdata$log_dentalevent <- log(new_mcdata$dentalevent + 1)
hist(new_mcdata$log_dentalevent, main = "Log-Transformed Dental Event", xlab = "Log(Dental Event ($) + 1)")
# Vision event
new_mcdata$log_visionevent <- log(new_mcdata$visionevent + 1)
hist(new_mcdata$log_visionevent, main = "Log-Transformed Vision Event", xlab = "Log(Vision Event ($) + 1)")
# Hearing event
```

```
new_mcdata$log_hearingevent <- log(new_mcdata$hearingevent + 1)

hist(new_mcdata$log_hearingevent, main = "Log-Transformed Hearing Event", xlab = "Log(Hearing Event ($) + 1)")

# Home health event

new_mcdata$log_homehealthevent <- log(new_mcdata$homehealthevent + 1)

hist(new_mcdata$log_homehealthevent, main = "Log-Transformed Home Health Event", xlab = "Log(Home Health Event ($) + 1)")

# Inpatient event

new_mcdata$log_inpatientevent <- log(new_mcdata$inpatientevent + 1)

hist(new_mcdata$log_inpatientevent, main = "Log-Transformed Inpatient Event", xlab = "Log(Inpatient Event ($) + 1)")

# Medical provider event

new_mcdata$log_medicalproviderevent <- log(new_mcdata$medicalproviderevent + 1)

hist(new_mcdata$log_medicalproviderevent, main = "Log-Transformed Medical Provider Event", xlab = "Log(Medical Provider Event ($) + 1)")

# Outpatient event

new_mcdata$log_outpatientevent <- log(new_mcdata$outpatientevent + 1)

hist(new_mcdata$log_outpatientevent, main = "Log-Transformed Outpatient Event", xlab = "Log(Outpatient Event ($) + 1)")

# Prescribe medicine

new_mcdata$log_prescribemedicine <- log(new_mcdata$prescribemedicine + 1)

hist(new_mcdata$log_prescribemedicine, main = "Log-Transformed Prescribe Medicine", xlab = "Log(Prescribe Medicine ($) + 1)")

# Total payment

new_mcdata$log_totalpayment <- log(new_mcdata$totalpayment + 1)

hist(new_mcdata$log_totalpayment, main = "Log-Transformed Total Payment", xlab = "Log(Total Payment ($) + 1)")

# Medicare payment

new_mcdata$log_medicarepayment <- log(new_mcdata$medicarepayment + 1)

hist(new_mcdata$log_medicarepayment, main = "Log-Transformed Medicare Payment", xlab = "Log(Medicare Payment ($) + 1)")

# Medicaid payment

new_mcdata$log_medicaidpayment <- log(new_mcdata$medicaidpayment + 1)
```

```
hist(new_mcdata$log_medicaidpayment, main = "Log-Transformed Medicaid Payment", xlab = "Log(Medicaid Payment ($) + 1)")

# Medicare Advantage payment

new_mcdata$log_medicareadvantagepayment <- log(new_mcdata$medicareadvantagepayment + 1)

hist(new_mcdata$log_medicareadvantagepayment, main = "Log-Transformed Medicare Advantage Payment", xlab = "Log(Medicare Advantage Payment ($) + 1)")

# Private insurance payment

new_mcdata$log_privateinsurancepayment <- log(new_mcdata$privateinsurancepayment + 1)

hist(new_mcdata$log_privateinsurancepayment, main = "Log-Transformed Private Insurance Payment", xlab = "Log(Private Insurance Payment ($) + 1)")

# Out-of-pocket payment

new_mcdata$log_outofpocketpayment <- log(new_mcdata$outofpocketpayment + 1)

hist(new_mcdata$log_outofpocketpayment, main = "Log-Transformed Out-of-Pocket Payment", xlab = "Log(Out-of-Pocket Payment ($) + 1)")

# Uncollected liability

new_mcdata$log_uncollectedliability <- log(new_mcdata$uncollectedliability + 1)

hist(new_mcdata$log_uncollectedliability, main = "Log-Transformed Uncollected Liability", xlab = "Log(Uncollected Liability ($) + 1)")

# Other payment



new_mcdata$otherpayment <- new_mcdata$otherpayment + abs(min(new_mcdata$otherpayment)) + 1

new_mcdata$log_otherpayment <- log(new_mcdata$otherpayment)


new_mcdata$ihs_otherpayment <- asinh(new_mcdata$otherpayment)

hist(new_mcdata$log_otherpayment, main = "Log-Transformed Other Payment", xlab = "Log(Other Payment ($) + 1)")
```

```
#Testing for linearity of total payment vs each variable again after changes
par(mfrow = c(4, 4))
dependent_var <- "totalpayment"
independent_vars <- setdiff(names(new_mcdata), dependent_var)

# Create a function to plot linearity
par(mfrow = c(4,4))
plot_linearity <- function(x, y, plot_title) {
  plot(y ~ x, data = new_mcdata, main = plot_title, xlab = deparse(substitute(x)), ylab = deparse(substitute(y)))
  abline(lm(y ~ x, data = new_mcdata), col = "red")
}
# Loop through each independent variable and plot linearity
for (var in independent_vars) {
  plot_title <- paste(dependent_var, "vs", var)
  plot_linearity(new_mcdata[[var]], new_mcdata[[dependent_var]], plot_title)
}


#Regression Model
  # ols model

    # Variables age_g3, otherrace and incomemoreequal25 have been removed due to perfect collinearity with their opposite variables


     ols1 <- lm(formula = totalpayment ~ sex + numberchroniccond + log_dentalevent+ log_visionevent+ log_hearingevent+
              log_medicalproviderevent +log_outpatientevent + log_prescribemedicine+ log_medicarepayment+
```

```
                 log_medicaidpayment+ log_medicareadvantagepayment+
log_privateinsurancepayment + log_outofpocketpayment+

log_uncollectedliability+log_otherpayment+age_g1+age_g2+hisp+nonhispblack+nonhispwhite+
incomeless25
                , data = new_mcdata)
        summary(ols1)
        r1<-residuals(ols1)
        qqnorm(r1)
        qqline(r1)


        ols2 <- lm(formula = totalpayment ~  numberchroniccond + log_visionevent +
log_medicalproviderevent +
                 log_outpatientevent + log_prescribemedicine + log_medicarepayment +
log_medicaidpayment +
                 log_medicareadvantagepayment + log_privateinsurancepayment +
log_outofpocketpayment +
                 log_uncollectedliability + log_otherpayment + nonhispblack, data = new_mcdata)
        summary(ols2)
        r2<-residuals(ols2)
        qqnorm(r2)
        qqline(r2)


        library("MASS")
        new_mcdata$totalpayment <- new_mcdata$totalpayment + 1
        bc <- boxcox(ols2, lambda = seq(-3, 3))
        best.lam = bc$x[which(bc$y==max(bc$y))]
```

```
# Update the model formula with the best lambda value
fullmodel.inv <- lm((totalpayment^best.lam) ~ numberchroniccond + log_visionevent +
log_medicalproviderevent +
                    log_outpatientevent + log_prescribemedicine + log_medicarepayment +
log_medicaidpayment +
                    log_medicareadvantagepayment + log_privateinsurancepayment +
log_outofpocketpayment +
                    log_uncollectedliability + log_otherpayment + nonhispblack, data =
new_mcdata)
summary(fullmodel.inv)
r3<-residuals(ols3)
qqnorm(r3)
qqline(r3)

stargazer(ols1,ols2, ols3,  type = "text")
```