Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: 1. Sadeem Al Gaaod (200013927) 2. Victor Linkevich (210028599) 3. Duong Hoang Lai (240034730) 4. Vyan Shekh Akhmad (210026105)	**GROUP NUMBER:** 1
MSc in: Business Analytics	
Module Code: SMM634	
Module Title: Analytics Methods for Business	
Lecturer: Rosalba Radice	**Submission Date:** **23/10/2024**

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Contents

Part 1: Identification Response Variable

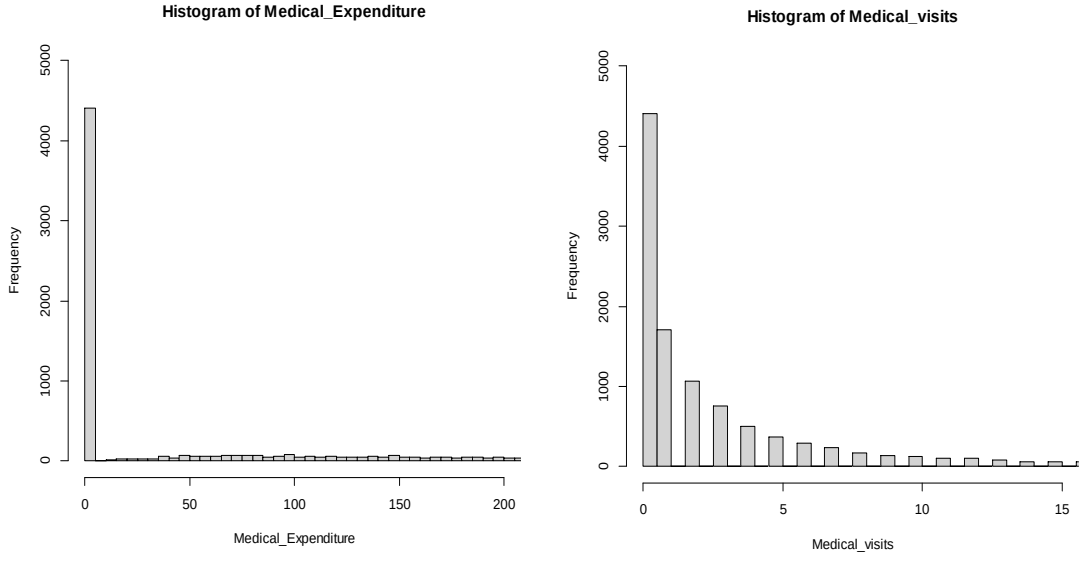
This report examines the factors influencing income using data from 10,638 individuals from the USA in 2012. The dataset includes variables covering a wide range of health, demographic, and socioeconomic factors.

Table 1, Preliminary Variable classification and analysis

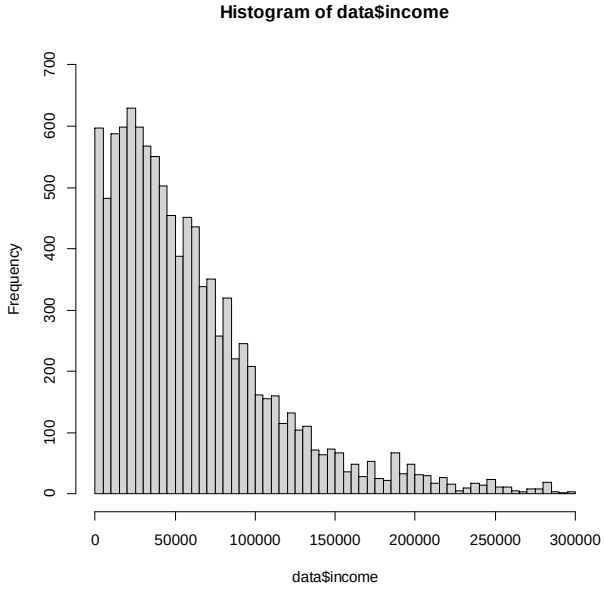
VARIABLE	TYPE	UNIQUE VALUES	DESCRIPTION
GENERAL	Ordinal	5	General Health (1= good, 5=bad)
MENTAL	Ordinal	5	Mental Health same as above
BMI	Continuous	398	Healthy range: 18.5-24.9
INCOME	Continuous	3733	Income in US Dollars
AGE	Discrete	48	Age from 18 to 65
GENDER	Binary	2	0= Female, 1=Male
ETHNICITY	Ordinal	4	1 white, 2 black, 3 native American, 4 others
EDUCATION	Discrete	18	Years in Education
REGION	Ordinal	4	1 Northeast, 2 Midwest, 3 South, 4 West
HYPERTENSION	Binary	2	0 = no, 1 = diagnosed
HYPERLIPIDEMIA	Binary	2	0 = no, 1 = diagnosed
DVISIT	Discrete	30	Doctor visits
NDVISIT	Discrete	29	Non-doctor visits
DVEXPEND	Continuous	1894	Doctor expenditure (USD)
NDVEXPEND	Continuous	1078	Non-Doctor expenditure (USD)

In general, a continuous numerical variable will accommodate the assumptions underlying a linear regression more than another type of variable as the variable can take any number on a fitted prediction line. In addition, the distribution of errors within a continuous variable tends to more closely resemble a normal distribution.

Figure 1, Histograms of total medical expenditures and total medical visits



Medical Expenditure is the sum of variables *dvexpend + ndvexpend*, whereas *Medical visits* is the sum of variables *dvisit + ndvisit*. These histograms illustrate that out of the 10,638 observations we have almost 45% of these variables equal to 0. This huge skewness to 0 will have severe impacts on our regression such as: heteroskedasticity, non-normal errors, and highly biased predictions.



For this exercise we have chosen income, as seen in the income histogram above the variable is more distributed in contrast to the medical visits and expenditure variables. This response variable will provide key insights to how health and social metrics impact the income of an individual. We want to analyse the relationship between all our variables statistically, therefore we are looking at all possible covariates and go through the elimination process in part 2.

Part 2: Fitting a General Linear Regression Model

Assumptions of a Linear Regression

A Multiple-Linear regression can be displayed where β_i are the coefficients of the linear relationship between Y_i and X_i, and ϵ_i is the residuals:

$$Y_i = \beta_{0i} + \beta_{1i}X_{1i} + \beta_{2i}X_{2i} + \dots \beta_{ni}X_{ni} + \epsilon_i \,,$$

A linear regression is one of the simplest econometric models that can be easily applied and interpreted. The assumptions associated with the general linear regression to help justify our decision, using the LINE acronym utilized by (Berenson, Levine and Szabat, 2015):

Linearity: the relationship between the response variable and co-variates is linear.

Independence of Errors: Where ϵ_i is not correlated with one another.

Normality of error: Where ϵ_i is normally distributed with a mean of 0 for each X_i unit.

Equal variance: Residual (ϵ_i) variance is constant.

Multicollinearity

Multicollinearity is when two or more independent variables are highly correlated leading to insignificant betas and an inflated standard error.

We have employed the Variance Inflation Factor (VIF) calculated by:

$$VIF(X_{i1}) = \frac{1}{1 - R_i^2},$$

The R_i^2 represents the R-squared value for the independent variable regressed on all other predictors. This means that as the X_{i1} is more correlated with other predictors, R_i^2 will rise, if the VIF score is greater than 5 is considered too high (Berenson, Levine and Szabat, 2015). In our case Mental health at level 5 had a VIF score of 21.2, and General Health at level 4 had a VIF score of 10.1 which were removed.

Variable selection

After the VIF, the model is saturated with insignificant variables. The Stepwise AIC is a model selection procedure where variables are incrementally added and selects the lowest Akaike Information Criterion (AIC).

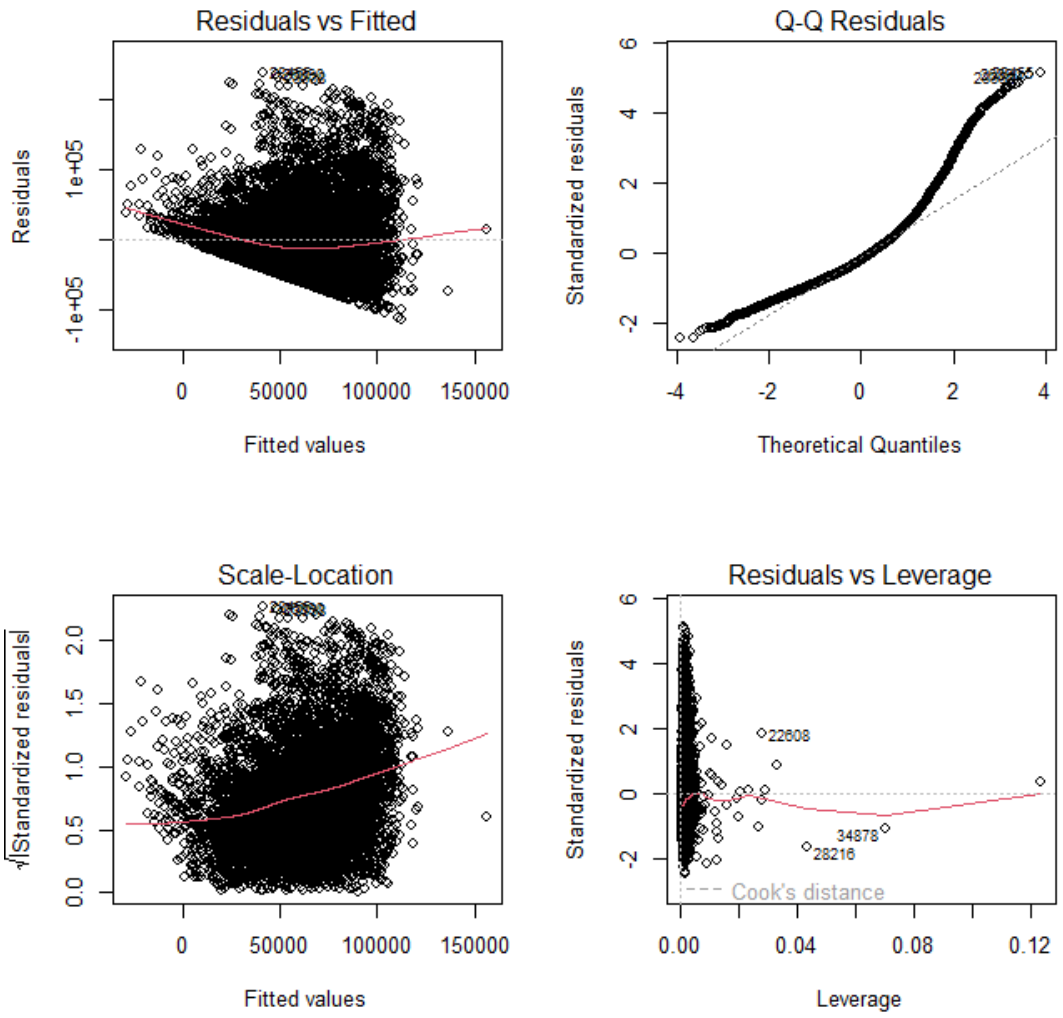
$$AIC = (2 * \ Number\ of\ parameters) - 2 * \ln(likelihood\ function),$$

AIC tries to find an optimal solution between a complex and goodness of fit. After running this on our regression, we remain with 18 variables. To contrast we created a parsimonious model that used the theoretically most impactful significant indicators which were: bmi, age, education, general health (excellent and fair), and Black ethnicity (Anderson et al., 2018).

Part 3: Summary and Interpretation of the Results

Post diagnostic testing

The graphs below represent the post diagnostic residual tests for the initial (complex) model.



Residual vs Fitted: This plot tests linearity, and if we have a constant variance. In this case we see that the data generally is linear around zero with random variance.

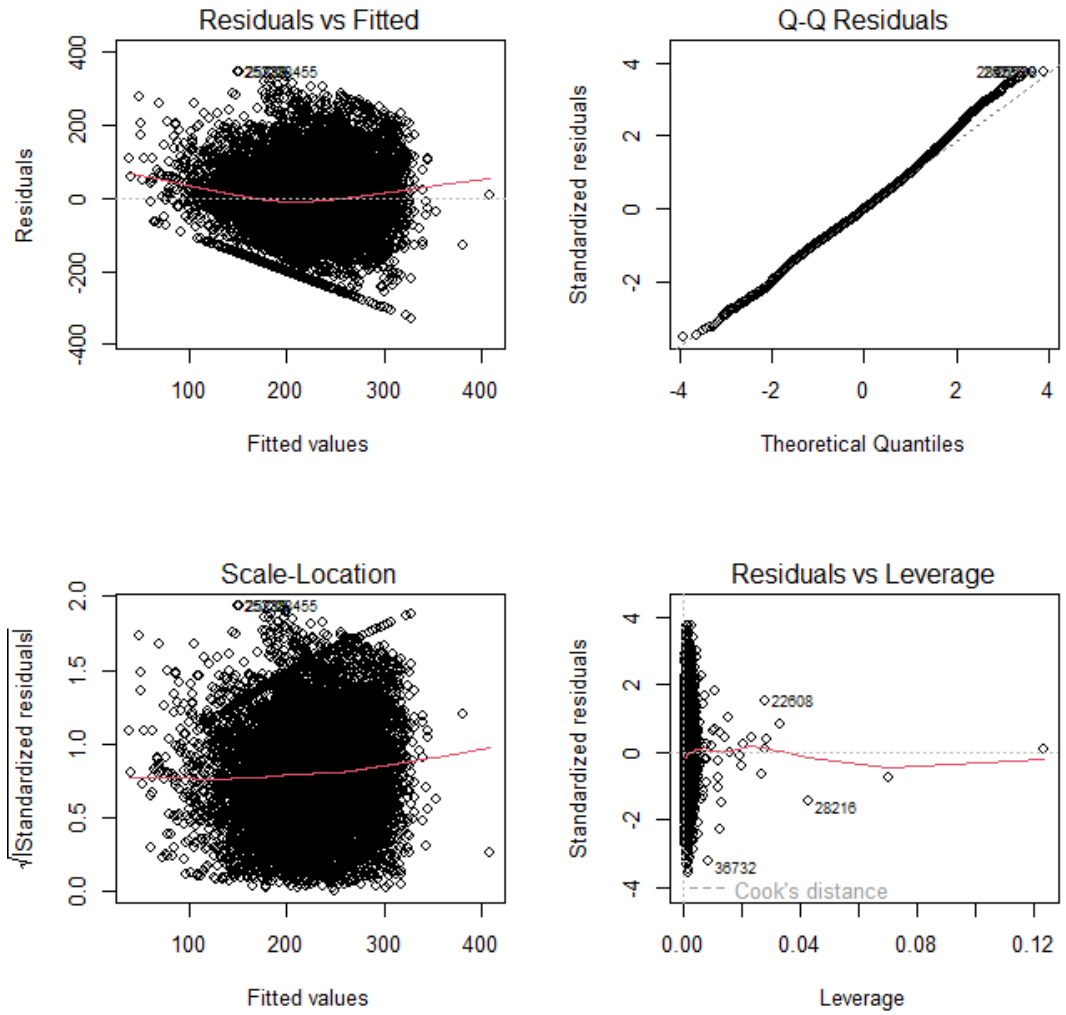
Q-Q Residual plot: Tests for Normality of our model. We can see that our residuals are not normally distributed and there is a heavy right tail.

Scale-Location and the Residual vs Leverage plots: Finds whether we have any influential points outside of the Cook's distance or outliers that are impacting our data. The plots show that no point seems to exit the cook's distance or a severe outlier disrupting our data (Berenson, Levine and Szabat, 2015).

To solve the problem of normality in our residual Q-Q plot, the income curve must be smoothed, and since we have a portion of our data as '0' we decided to use the square root of income rather than the log of income for this function. Moreover, to capture any non-linear relationships in

the data, we ran a loop for all variables in the model to be plotted against the residuals but found no non-linear relationships present (Baldock, 2014).

Updated model with the square root of income has the following post-diagnostic plots:



As we can see from the plots above, the Q-Q residuals are now much more normally distributed adhering to the assumption of a linear regression, in addition there was a minor improvement in the linearity under the Residual vs fitted plot. However, the coefficients become harder to interpret as a result.

Summary output of income regressions

	Dependent variable:	
	income	
	(Initial)	(Parsimonious)
bmi	−383.021***	−455.211***
	(74.887)	(73.609)
age	462.789***	473.544***
	(37.932)	(34.360)
gender	5,812.593***	
	(908.299)	
education	5,496.354***	5,745.683***
	(160.115)	(158.943)
hypertension	−3,323.916***	
	(1,230.525)	
hyperlipidemia	3,913.333***	
	(1,242.348)	
dvexpend	1.112***	
	(0.280)	
gen excellent	5,989.388***	8,048.986***
	(1,104.161)	(1,062.049)
gen fair	−10,824.290***	−15,002.880***
	(1,693.167)	(1,618.713)
men good	−5,924.020***	
	(1,129.475)	
men fair	−14,349.920***	
	(2,093.392)	
eth white	−7,509.029***	
	(1,589.521)	
eth black	−20,497.250***	−16,360.960***
	(1,839.906)	(1,128.259)
reg mw	−4,521.362***	
	(1,248.975)	
reg s	−7,011.800***	
	(1,035.207) 2	
Constant	−5,798.781	−16,277.260***
	(3,600.595)	(3,168.285)
Adjusted R^2	0.187	0.171
Residual Std. Error	46,403.580 (df = 10622)	46,840.020 (df = 10631)
F Statistic	163.658*** (df = 15; 10622)	367.219*** (df = 6; 10631)

Above is the summary of initial full model and Parsimonious model for income, square root of income summary can be found in Appendix B.

From the model summary above, we can see that both models only contain variables that are significant at the 1% confidence level. The difference in parsimonious model contains 9 fewer variables than our initial model, this led to a decrease of the adjusted R squared by 1.6%. In addition, we note that the F statistic proves the models contain at least 1 variable that are non-zero.

From the Beta coefficients, both models display remarkably similar relationships between income and the independent variables. We found that there was a significant positive relationship between income with age, and education. For example, for our initial and parsimonious models, age has a beta coefficient of approximately 462, and 473 respectively. This means that for every 1-year increase in age, income on average moves $462 or $473 upward. This is a logical relationship as older people make more income due to experience until retirement age which is outside our data maximum (65 years old). Whereas bmi exhibited a negative relationship indicating that for each unit increase in bmi (worse health for most data), income would fall by approximately $383 or $455.

The interpretation of dummy variables such as gen_excellent which is a general health score of 5, contains coefficients of approximately 5,989 and 8,048 respectively. For this dummy variable our reference category was the lowest health score of 5. The models indicate that compared to the reference category, someone with a very high general health score will on average have a higher income of $5,989 or $8,048. However, this relationship we found isn't completely linear across the scores. Looking at the general health score of fair (level 4), we see that these individuals are expected to make less than the reference category by a quantity of $10,800 or $15,000. These results can indicate that being on either extreme of the general health category may indicate a higher income, jobs that have high incomes can be extremely stressful and thus this relationship can be difficult to interpret.

Moreover, we incorporated the ANOVA technique between both models in R which is an F-test with the null hypothesis that all variables not shared in both models have coefficients equal to 0. However, we found that we rejected the F-test for all remaining variables indicating high significance in the covariates that we removed from the initial model (Baldock, 2014).

Part 4: Limitations

Nature of the data set:

The data set is an empirical social science research, which includes data relating to human behaviour that is usually unpredictable and varies drastically depending on unrecorded factors such as mood, feelings, personal preferences, etc. This usually results in oversimplifying some relationships, potentially leading to incomplete conclusions.

Zero-inflated data set:

During data exploration process, we noticed that almost half of the data set have values of 0 on certain variables such as: dvisit, ndvisit, dvexpend and ndvexpend. This could be an implication of data quality control problem during data gathering process, impacting the integrity of the data set. Moreover, with a lot of zeros in the variables, it could also lead to some issues when conducting regression analysis:

- Reduced Variability: A high number of zeros reduced the overall variability of the independent variable, leading to less reliable estimates of regression coefficients.
- Heteroscedasticity: Since the variables with high presences of zeros correlate with the spread of dependent variable (income), it resulted in heteroscedasticity (non-constant variance of residuals), violating one of the key assumptions of linear regression.
- Interpretation of Coefficients: The interpretation of regression coefficients become more complex. Inflated standard errors and less precise coefficient estimates.
- Reduced Statistical Power: A large number of zeros reduced the statistical power of the analysis, resulting in low R squared and adjusted R squared, making it harder to detect significant effects or relationships.

Limitation of the analysis:

Regression analysis focuses on linear relationships and may overlook complex interactions or non-linear associations

After running stepAIC method to refine the model, initially there were many variables included in the model, yet it is difficult to remove variables as they impact the explainability of the prediction model. While our parsimonious model addresses this to some extent, the risk remains, particularly when attempting to generalize findings to other datasets.

The coefficients of regression models can be difficult to interpret, especially in the presence of interaction terms or when dealing with categorical variables.

Using the square root of income improved model fit but made interpreting the beta coefficients more complex. The transformation means the coefficients reflect changes in the square root of income, making it harder to directly understand their impact on actual income.

Part 5: Suggestions for Improvement

Segmentation of Data: The zeros in the data set could represent a separate segment of our population. It might be useful to analyse the zeros and non-zeros separately or consider interaction terms to capture potential hidden behaviours patterns.

Examine Patterns: Investigate whether there are patterns or systematic reasons for the zeros, such as specific characteristics of the observations (e.g., demographic factors). This can inform further modelling approach

Zero-inflated Data Analysis: to cope with the substantial number of zeros in the data, we could further look into Zero-inflated regression analysis model.

Utilizing Shrinkage methods for Variable Selection: Given the high number of variables and the potential overfitting identified, future models should incorporate shrinkage methods like LASSO (Least Absolute Shrinkage and Selection Operator) and Ridge regression, using them we can select the most significant variables and improve model interpretability.

Addressing data quality issues: To mitigate the high proportion of zeros in variables such as dvisit and dvexpend future data collection efforts should be focused on data quality as well as handling missing or skewed data.

Integration of external data sources: Regional economic indicators and environmental factors (pollutions levels) could be mentioned in the future research report.

Longitudinal Analysis: The dataset in the current analysis is cross-sectional, meaning it captures data only at a single point in time, which limits the ability to assess longitudinal changes. Conducting a longitudinal analysis would allow us to track individual's health and income over extended periods, which would provide deeper insights into how health factors affect income trajectories over time.

Model Comparison: One way to better measure the predictive performance of models is by using the mean squared error (MSE) instead of only relying on ANOVA and adjusted R-squared metrics

References:

Anderson, D.R., Sweeney, D.J., Williams, T.A., Camm, J.D. and Cochran, J.J. (2018). *Statistics for business & economics*. Boston, Ma: Cengage Learning.

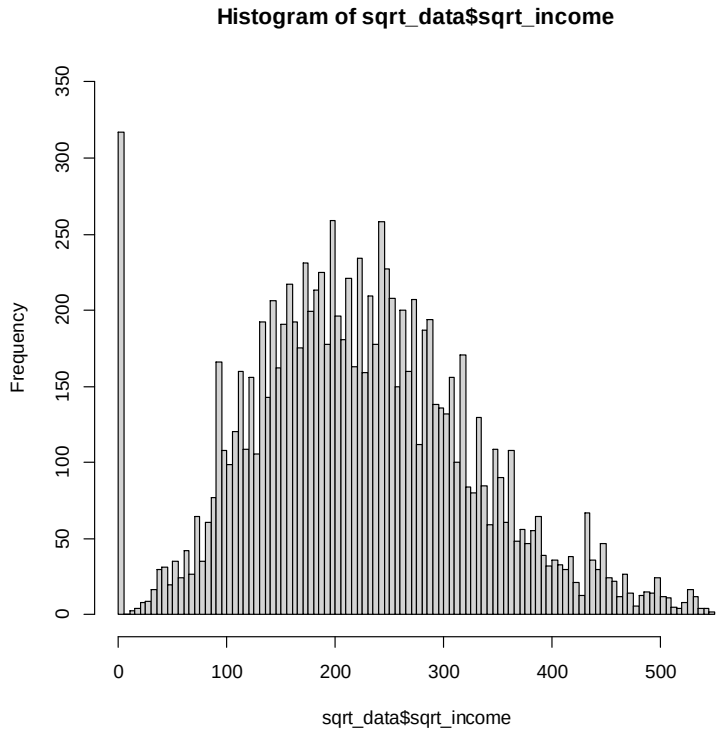
Baldock, S. (2014). *Using R for Statistics*. Berkeley, Ca: Apress.

Berenson, M.L., Levine, D. and Szabat, K.A. (2015). *Basic Business Statistics PDF eBook, Global Edition*. Pearson Higher Ed.

Ozili, P.K. (2023). The acceptable R-square in empirical modelling for social science research. *The acceptable R-square in empirical modelling for social science research*. [online] Available at: https://ideas.repec.org/p/pra/mprapa/115769.html [Accessed 21 Jan. 2024].

Appendices:

Appendix A: Histogramm of sqrt(income)



Appendix B: Square root of income output

	Dependent variable:	
	sqrt income	
	(Initial)	(Parsimonious)
bmi	−0.653***	
	(0.149)	
age	1.009***	0.934***
	(0.075)	(0.067)
gender	13.845***	14.125***
	(1.807)	(1.808)
education	11.051***	11.596***
	(0.318)	(0.317)
hypertension	−7.430***	
	(2.448)	
hyperlipidemia	6.749***	
	(2.472)	
dvexpend	0.002***	
	(0.001)	
gen excellent	11.218***	
	(2.197)	
gen fair	−26.031***	−29.617***
	(3.369)	(3.353)
men good	−13.451***	−17.397***
	(2.247)	(2.167)
men fair	−35.531***	−38.603***
	(4.165)	(4.142)
eth white	−16.135***	
	(3.154)	
eth black	−46.431***	−37.616***
	(3.648)	(2.221)
reg s	−10.508***	
	(1.891)	
Constant	83.068***	48.752***
	(7.159)	(4.960)
R^2	0.201	0.190
Adjusted R^2	0.200	0.190
Residual Std. Error	92.334 (df = 10623)	92.942 (df = 10630)
F Statistic	191.335*** (df = 14; 10623)	356.876*** (df = 7; 10630)

Appendix C: Code

```
## LIBRARIES & DIRECTORY ----

# libraries

library(MASS)

library(carData)

library(car)

#install.packages("stargazer")

#library(stargazer)

# set directory to the location of the R file (only works in RStudio)

install.packages("rstudioapi")

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))



## DATA ----

# make sure the R file and the data file are in the same folder to import data (or set directory manually)

data <- read.table("expenditure.txt")


# dissect the dummy variables into several binary dummy variables and remove the original dummy variable

temp.data <- data.frame(data)


# general health

temp.data$gen_excellent <- ifelse(temp.data$general == 1, 1, 0)

temp.data$gen_vgood <- ifelse(temp.data$general == 2, 1, 0)

temp.data$gen_good <- ifelse(temp.data$general == 3, 1, 0)

temp.data$gen_fair <- ifelse(temp.data$general == 4, 1, 0)

# temp.data$gen_bad <- ifelse(temp.data$general == 5, 1, 0)      # reference

temp.data$general <- NULL


# mental health

temp.data$men_excellent <- ifelse(temp.data$mental == 1, 1, 0)

temp.data$men_vgood <- ifelse(temp.data$mental == 2, 1, 0)

temp.data$men_good <- ifelse(temp.data$mental == 3, 1, 0)

temp.data$men_fair <- ifelse(temp.data$mental == 4, 1, 0)

# temp.data$men_bad <- ifelse(temp.data$mental == 5, 1, 0)       # reference

temp.data$mental <- NULL
```

```
# ethnicity
temp.data$eth_white <- ifelse(temp.data$ethnicity == 1, 1, 0)
temp.data$eth_black <- ifelse(temp.data$ethnicity == 2, 1, 0)
temp.data$eth_na <- ifelse(temp.data$ethnicity == 3, 1, 0)
# temp.data$eth_other <- ifelse(temp.data$ethnicity == 4, 1, 0)  # reference
temp.data$ethnicity <- NULL

# region
temp.data$reg_ne <- ifelse(temp.data$region == 1, 1, 0)
temp.data$reg_mw <- ifelse(temp.data$region == 2, 1, 0)
temp.data$reg_s <- ifelse(temp.data$region == 3, 1, 0)
# temp.data$reg_w <- ifelse(temp.data$region == 4, 1, 0)         # reference
temp.data$region <- NULL

# commit changes to the data
data <- temp.data



## TOTAL MEDICAL DATA ----
par(mfrow = c(2,1))

# mdexpend showing combined expenses on medical visits
data$mdvexpend <- data$dvexpend + data$ndvexpend
hist(data$mdvexpend,
    xlim = c(0, 60000),
    ylim = c(0, 5000),
    breaks = 1000)
#TODO ensure to remove the column before running the regression!
data$mdvexpend <- NULL

# mdvisit showing combined number of medical visits
data$mdvisit <- data$dvisit + data$ndvisit
hist(data$mdvisit,
    xlim = c(0, 60),
    ylim = c(0, 5000),
    breaks = 80)
```

```
#TODO ensure to remove the column before running the regression!
data$mdvisit <- NULL



## INCOME ----
# histogram of the response
par(mfrow = c(1, 1))
hist(data$income,
    ylim = c(0, 700),
    breaks = 100)

# fitting the MLR model and removing colinear variables
lm.fit <- lm(income ~.
           -men_excellent  # VIF = 21.2
           -gen_vgood      # VIF = 10.1
           ,data)
summary(lm.fit)
vif(lm.fit)

# running a mixed selection to only include significant variables
lm.fit.select <- stepAIC(lm.fit, ~.,
                    direction = "both",  # ensure mixed is used
                    trace = FALSE        # less console output
                    )
summary(lm.fit.select)

# Manually removing more variables based on p-values above 1% & 5%
lm.fit.manual <- update(lm.fit.select, ~.
                    -gen_good  # -0.0002, p-value > 5%
                    -ndvisit   # -0.0003, p-value > 1%
                    -eth_na    # -0.0004, p-value > 1%
                    )          # minus value shows the reduction of Adj.R^2 by removing that variable
summary(lm.fit.manual)  # Adj.R^2 = 18.75% --> 18.66%, for removing 3 predictors

# Parsimonious model
lm.fit.parsim <- update(lm.fit.manual, ~.
```

```
                -dvexpend       # -0.0012, approx. 45% zero observations
                -hypertension    # -0.0004, p-value > 1%
                -hyperlipidemia  # -0.0005, greatest p-value
                -reg_mw          # -0.0009, greatest p-value
                -eth_white       # -0.0018, greatest p-value
                -men_good        # -0.0021, greatest p-value
                -men_fair        # -0.0024, greatest p-value
                -reg_s           # -0.0030, greatest p-value
                -gender          # -0.0031, greatest p-value
                )               # minus value shows the reduction of Adj.R^2 by removing that variable
summary(lm.fit.parsim)
#stargazer(lm.fit.manual,lm.fit.parsim, type="text")# Adj.R^2 = 18.66% --> 17.12%, for removing 9 predictors

# comparing two models, manual & parsim, using ANOVA
anova(lm.fit.manual, lm.fit.parsim)

# checking for linearity/ normality of residuals/ outliers/ high leverage points
par(mfrow = c(2, 2))
plot(lm.fit.manual)

# plotting standardized residuals against each predictor for 'manual model' (possible to plot other models)
mod <- lm.fit.manual  # model used for plotting (select / manual / parsim)
par(mfrow = c(4, 4))
fit.names <- names(mod$model)[-1]
lm.fit.resid <- rstandard(mod)
for (i in 1:length(fit.names)) {
  plot(data[[fit.names[i]]], lm.fit.resid,
       xlab = fit.names[i],
       ylab = "Residuals")
  abline(lm(lm.fit.resid ~ data[[fit.names[i]]]),
         col = "red",
         lwd = 1.5)
}

# predicting response and seeing its confidence & predicting intervals
confint(lm.fit.manual)
```

```
fit.predict.values <- data.frame(bmi = 20,
                                 age = 30,
                                 gender = 1,
                                 education = 12,
                                 hypertension = 1,
                                 hyperlipidemia = 1,
                                 dvexpend = 300,
                                 gen_excellent = 0, gen_fair = 1,
                                 men_good = 1, men_fair = 0,
                                 eth_white = 1, eth_black = 0,
                                 reg_mw = 0, reg_s = 1
                                 )
predict(lm.fit.manual, fit.predict.values, se.fit = TRUE, interval = c("confidence"))
predict(lm.fit.manual, fit.predict.values, se.fit = TRUE, interval = c("prediction"))



## sqrt(INCOME) (attempt to fix for non-normal distribution of residuals) ----
# new data set with square for income
sqrt_data <- data.frame(data)
sqrt_data$sqrt_income <- sqrt(sqrt_data$income)
sqrt_data$income <- NULL

# histogram of the response
par(mfrow = c(1, 1))
hist(sqrt_data$sqrt_income,
    ylim = c(0, 350),
    breaks = 100)

# fitting the MLR model and removing colinear variables
sqrt.lm.fit <- lm(sqrt_income ~.
                  -men_excellent  # VIF = 21.2
                  -gen_vgood      # VIF = 10.1
                  ,sqrt_data)
summary(sqrt.lm.fit)
vif(sqrt.lm.fit)
```

```
# running a mixed selection to only include significant variables
sqrt.lm.fit.select <- stepAIC(sqrt.lm.fit, ~.,
                              direction = "both",  # ensure mixed is used
                              trace = FALSE        # less console output
                              )
summary(sqrt.lm.fit.select)

# manually removing more variables based on p-values above 1% & 5%
sqrt.lm.fit.manual <- update(sqrt.lm.fit.select, ~.
                             -reg_ne   # -0.0001, p-value > 5%
                             -reg_mw   # -0.0001, P-value > 5%
                             -eth_na   # -0.0004, p-value > 1%
                             -dvisit   # -0.0004, p-value > 1%
                             -ndvisit  # -0.0003, p-value > 1%
                             )
summary(sqrt.lm.fit.manual) # Adj.R^2 = 20.16% --> 20.03%, for removing 5 predictors

# parsimonious model
sqrt.lm.fit.parsim <- update(sqrt.lm.fit.manual, ~.
                             -dvexpend        # -0.0010, approx. 45% zero observations
                             -hypertension    # -0.0006, greatest p-value
                             -hyperlipidemia  # -0.0003, greatest p-value
                             -bmi             # -0.0015, greatest p-value
                             -eth_white       # -0.0021, greatest p-value
                             -gen_excellent   # -0.0024, greatest p-value
                             -reg_s           # -0.0026, greatest p-value
                             )
summary(sqrt.lm.fit.parsim)  # Adj.R^2 = 20.03% --> 18.98%, for removing 7 predictors

# comparing two models, manual & parsim, using ANOVA
anova(sqrt.lm.fit.manual, sqrt.lm.fit.parsim)

# checking for linearity/ normality of residuals/ outliers/ high leverage points
par(mfrow = c(2, 2))
plot(sqrt.lm.fit.manual)
```

```
# plotting standardized residuals against each predictor for 'manual model' (possible to plot other models)
sqrt.mod <- sqrt.lm.fit.manual  # model used for plotting (select / manual / parsim)
par(mfrow = c(4, 4))
sqrt.fit.names <- names(sqrt.mod$model)[-1]
sqrt.lm.fit.resid <- rstandard(sqrt.mod)
for (i in 1:length(sqrt.fit.names)) {
  plot(data[[sqrt.fit.names[i]]], sqrt.lm.fit.resid,
       xlab = sqrt.fit.names[i],
       ylab = "Residuals")
  abline(lm(sqrt.lm.fit.resid ~ data[[sqrt.fit.names[i]]]),
         col = "red",
         lwd = 1.5)
}

# predicting response and seeing its confidence & predicting intervals for manual model
confint(sqrt.lm.fit.manual)
sqrt.fit.predict.values <- data.frame(bmi = 20,
                                      age = 30,
                                      gender = 1,
                                      education = 12,
                                      hypertension = 1,
                                      hyperlipidemia = 1,
                                      dvexpend = 300,
                                      gen_excellent = 0, gen_fair = 1,
                                      men_good = 1, men_fair = 0,
                                      eth_white = 0, eth_black = 1,
                                      reg_s = 1
                                      )
predict(sqrt.lm.fit.manual, sqrt.fit.predict.values, se.fit = TRUE, interval = c("confidence"))
predict(sqrt.lm.fit.manual, sqrt.fit.predict.values, se.fit = TRUE, interval = c("prediction"))
```