Data Analysis with SPSS – multiple linear regression

In this assignment I will examine the effects of cigarette consumption and exercise on health care cost based on the health care cost data that was given for this assignment. In this data analysis, I will be using three data features of the given dataset: total health care costs, average consumption of cigarettes per day and total hours of exercise per week.

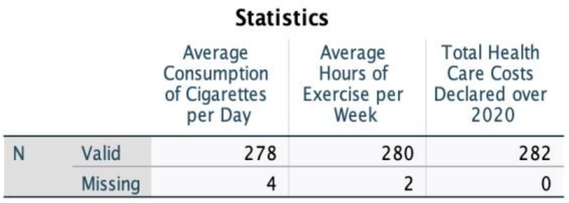
**1. What are the dependent and independent variables in your model? What regression model will be applied and why?**

The data features relevant in this case are total health care costs, average consumption of cigarettes per day and total hours of exercise per week. Out of these data features, the dependent variable is costs, and independent variables are cigarettes and exercise. This is because costs can be affected by the two independent variables, but cigarettes and exercise cannot be affected by the other variables. Costs as the dependent variable is reasonable also because it is the subject of this data analysis.

Multiple linear regression will be applied, as there are more than one independent variables. For predicting total health care costs based on the other variables, multiple linear regression would be the primary model as this approach helps to assess the influence of each predictor on costs while controlling other variables.

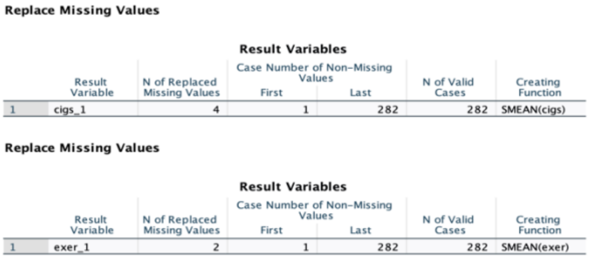
2. **Performing data preprocessing**

First, I checked missing values with descriptive analysis. Results of this analysis are presented in the picture below. Based on this analysis, two variables, Average consumption of cigarettes per day and Average hours of exercise per week, have missing values. These missing values were replaced with the series mean and two new columns with no missing values were created. These new columns are named cigs_1 or SMEAN(cigs) and exer_1 or SMEAN(exer).



Statistics

| | | Average Consumption of Cigarettes per Day | Average Hours of Exercise per Week | Total Health Care Costs Declared over 2020 |
|---|---|---|---|---|
| N | Valid | 278 | 280 | 282 |
| | Missing | 4 | 2 | 0 |



Replace Missing Values

Result Variables

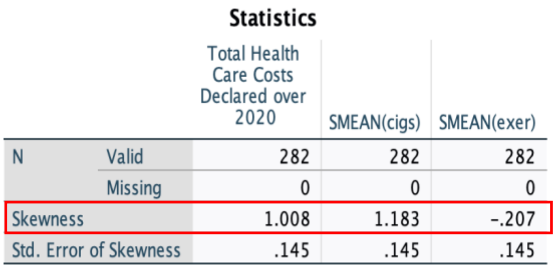
| | Result Variable | N of Replaced Missing Values | Case Number of Non-Missing Values | | N of Valid Cases | Creating Function |
|---|---|---|---|---|---|---|
| | | | First | Last | | |
| 1 | cigs_1 | 4 | 1 | 282 | 282 | SMEAN(cigs) |

Replace Missing Values

Result Variables

| | Result Variable | N of Replaced Missing Values | Case Number of Non-Missing Values | | N of Valid Cases | Creating Function |
|---|---|---|---|---|---|---|
| | | | First | Last | | |
| 1 | exer_1 | 2 | 1 | 282 | 282 | SMEAN(exer) |

Next, I checked the skewness of the variables with descriptive analysis. In this analysis, I checked the skewness of dependent variable costs and the two new variables cigs_1 and exer_1.

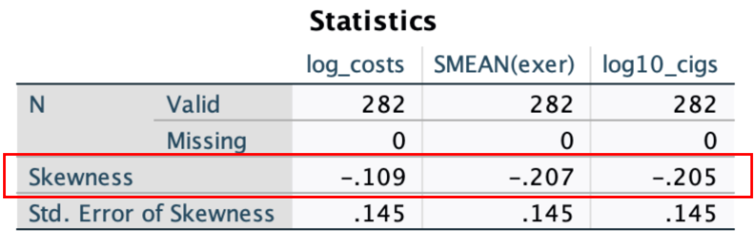


**Statistics**

| | | Total Health Care Costs Declared over 2020 | SMEAN(cigs) | SMEAN(exer) |
|---|---|---|---|---|
| N | Valid | 282 | 282 | 282 |
| | Missing | 0 | 0 | 0 |
| Skewness | | 1.008 | 1.183 | -.207 |
| Std. Error of Skewness | | .145 | .145 | .145 |

Based on this analysis total health care costs and SMEAN(cigs) (aka cigs_1) are skewed, because the skewness value is larger than 1. Variable SMEAN(exer) is not skewed because the value is between 0.5 and -0.5. Variable is highly skewed if the skewness value is larger than 1 or smaller than -1. Variable is considered moderately skewed if the skewness value is larger than 0.5 or smaller than -0.5.

Based on this analysis, I decided to do log transformation for variables costs and SMEAN(cigs). In the log transformation, I will use common logarithm (LG10). For variable SMEAN(cigs), I will use expression LG10(cigs + 1), because the variable has 0 values.

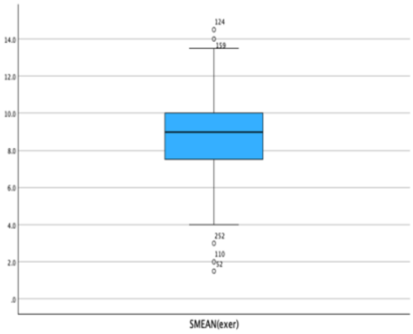


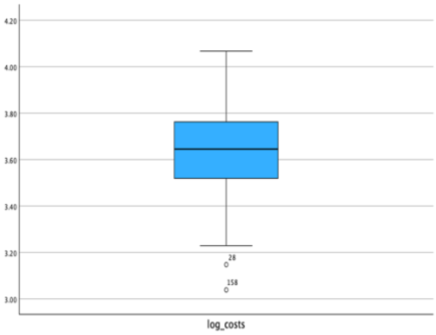
**Statistics**

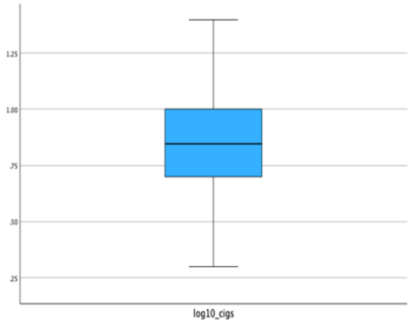
| | | log_costs | SMEAN(exer) | log10_cigs |
|---|---|---|---|---|
| N | Valid | 282 | 282 | 282 |
| | Missing | 0 | 0 | 0 |
| Skewness | | -.109 | -.207 | -.205 |
| Std. Error of Skewness | | .145 | .145 | .145 |

After log transformation, none of the variables are skewed, because skewness is between 0 and -0.5 for each variable.

Next, I identified outliers in my dataset. Outliers are marked with a star character (*) in the analysis. I checked outliers for variables SMEAN(exer), log_costs and log10_cigs, which are the variables with no skewness.

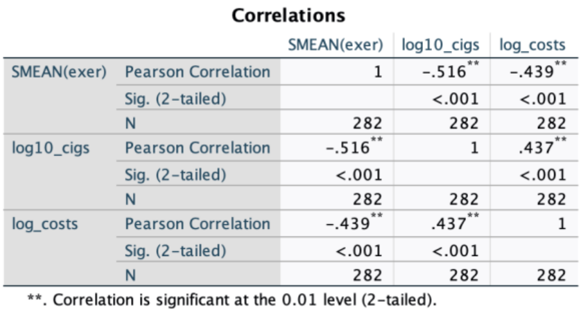






Based on this analysis, there are no outliers in the data.

Next, I conducted correlation analysis. Results of the correlation analysis are presented in the picture below.



**Correlations**

| | | SMEAN(exer) | log10_cigs | log_costs |
|---|---|---|---|---|
| SMEAN(exer) | Pearson Correlation | 1 | -.516$^{**}$ | -.439$^{**}$ |
| | Sig. (2-tailed) | | <.001 | <.001 |
| | N | 282 | 282 | 282 |
| log10_cigs | Pearson Correlation | -.516$^{**}$ | 1 | .437$^{**}$ |
| | Sig. (2-tailed) | <.001 | | <.001 |
| | N | 282 | 282 | 282 |
| log_costs | Pearson Correlation | -.439$^{**}$ | .437$^{**}$ | 1 |
| | Sig. (2-tailed) | <.001 | <.001 | |
| | N | 282 | 282 | 282 |

**. Correlation is significant at the 0.01 level (2-tailed).

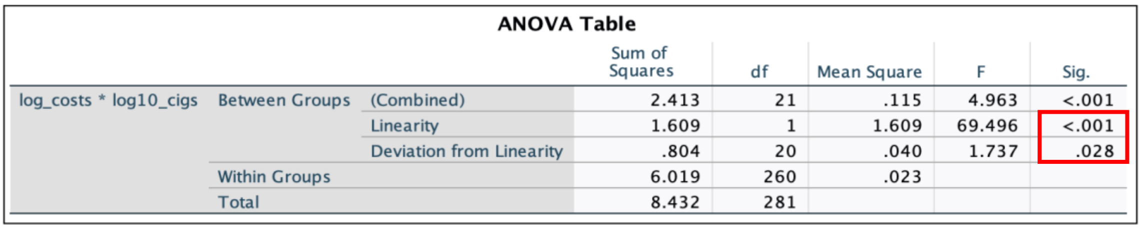
Based on this correlation analysis, variable exercise has significant negative correlation with health care costs and variable cigarettes has significant positive correlation with health care costs. Also, exercise has significant negative correlation with cigarettes.

## 3. Checking the assumptions for multiple linear regression model

I conducted multiple linear regression analysis for the three variables in this data analysis. I placed log_costs as the dependent variable and log10_cigs and SMEAN(exer) as the independent variables. After conducting multiple linear regression analysis, I checked the assumptions of linearity, multicollinearity, normality and homoscedasticity. Next, I will present the methods used to check these assumptions.

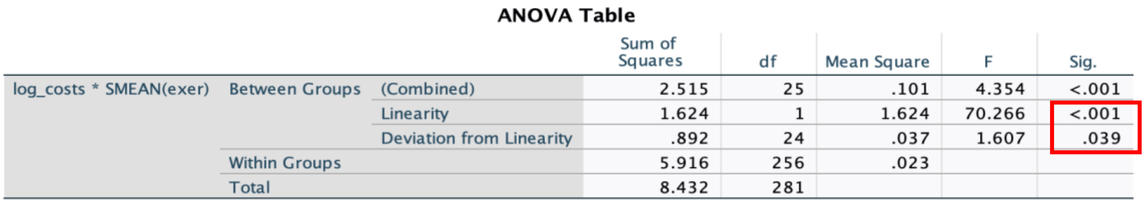
**Linearity:** The relationship between each independent variable and the dependent variable (costs) should be linear. In SPSS I used Analyze → Compare Means → Means to check linearity.

Here are the results of the linearity check:



**ANOVA Table**

| | | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|---|
| log_costs * log10_cigs | Between Groups | (Combined) | 2.413 | 21 | .115 | 4.963 | <.001 |
| | | Linearity | 1.609 | 1 | 1.609 | 69.496 | <.001 |
| | | Deviation from Linearity | .804 | 20 | .040 | 1.737 | .028 |
| | Within Groups | | 6.019 | 260 | .023 | | |
| | Total | | 8.432 | 281 | | | |

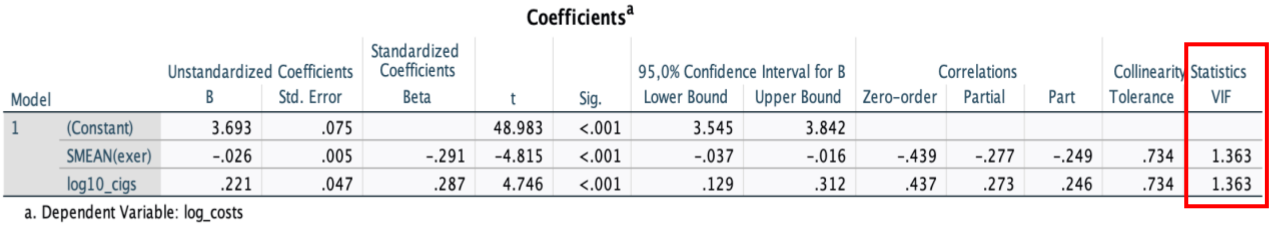


**ANOVA Table**

| | | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|---|
| log_costs * SMEAN(exer) | Between Groups | (Combined) | 2.515 | 25 | .101 | 4.354 | <.001 |
| | | Linearity | 1.624 | 1 | 1.624 | 70.266 | <.001 |
| | | Deviation from Linearity | .892 | 24 | .037 | 1.607 | .039 |
| | Within Groups | | 5.916 | 256 | .023 | | |
| | Total | | 8.432 | 281 | | | |

These results indicate that there is a linear relationship between costs and cigarettes as well as costs and exercise. However, in both cases, the significance value of deviation from linearity is small, which means the deviation is also significant. A p-value less than the typical alpha level (e.g., .05) in deviation from linearity suggests that the relationship between the two variables deviates significantly from a purely linear trend, indicating that a linear model may not fully capture the pattern in the data.

**Multicollinearity**: Independent variables should not be too highly correlated with each other. I checked the variables for multicollinearity by conducting multiple linear regression analysis and checking the VIF-value. The results of multicollinearity check are presented in the picture below:



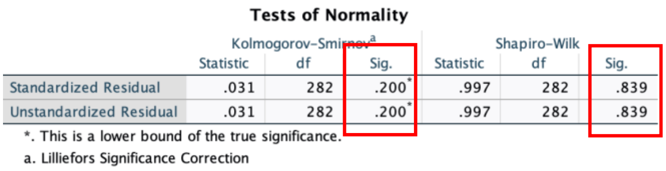
Coefficients[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | 95,0% Confidence Interval for B | | Correlations | | | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Lower Bound | Upper Bound | Zero-order | Partial | Part | Tolerance | VIF |
| 1 | (Constant) | 3.693 | .075 | | 48.983 | <.001 | 3.545 | 3.842 | | | | | |
| | SMEAN(exer) | -.026 | .005 | -.291 | -4.815 | <.001 | -.037 | -.016 | -.439 | -.277 | -.249 | .734 | 1.363 |
| | log10_cigs | .221 | .047 | .287 | 4.746 | <.001 | .129 | .312 | .437 | .273 | .246 | .734 | 1.363 |

a. Dependent Variable: log_costs

Based on the VIF-values, there is no multicollinearity issue between the independent variables, because the VIF-value is not above 5 in any of the rows. Multicollinearity check is passed.

**Normality:** The residuals should be normally distributed. I did the normality check by using descriptive statistics as follows: Analyze → Descriptive statistics → Explore → Add the standardized residual values of dependent variables in the dependent list → Plot

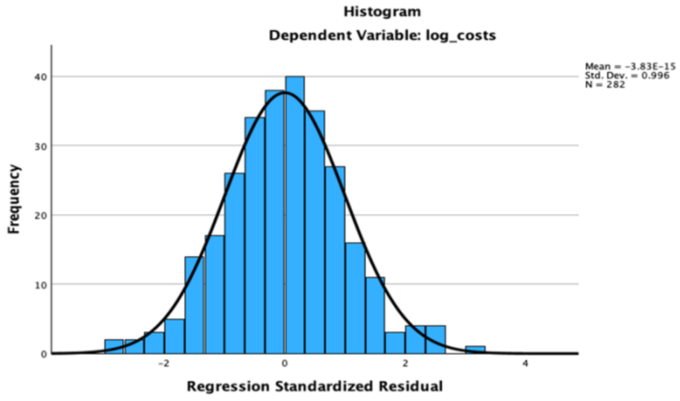


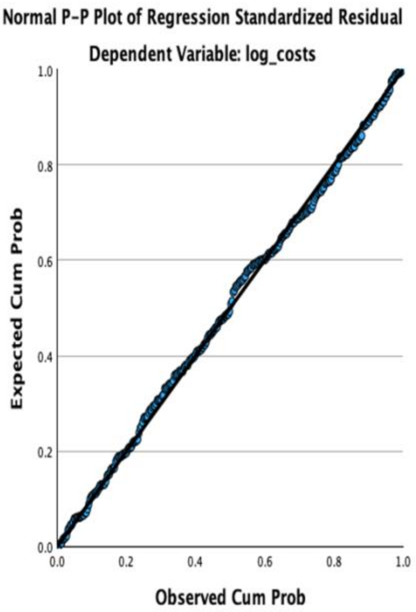
Tests of Normality

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| Standardized Residual | .031 | 282 | .200* | .997 | 282 | .839 |
| Unstandardized Residual | .031 | 282 | .200* | .997 | 282 | .839 |

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

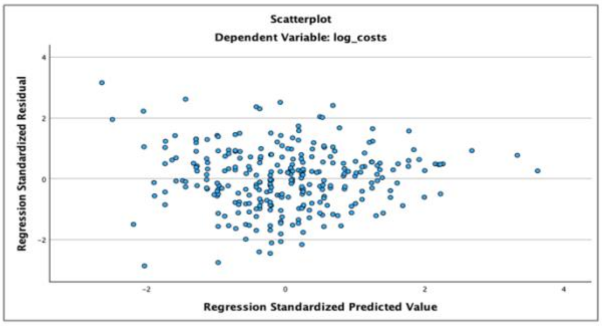




Here are the results of the normality check:

Based on the value of significance in the tests of normality I concluded that normality is supported. This is because the p-value is larger than .05. Also, the histogram and P-P Plot implicate that the standardized residual is normally distributed. In conclusion, normality check is passed.

**Homoscedasticity:** The residuals should have constant variance across all levels of the predicted values. I did the homoscedasticity check with a scatterplot I SPSS.

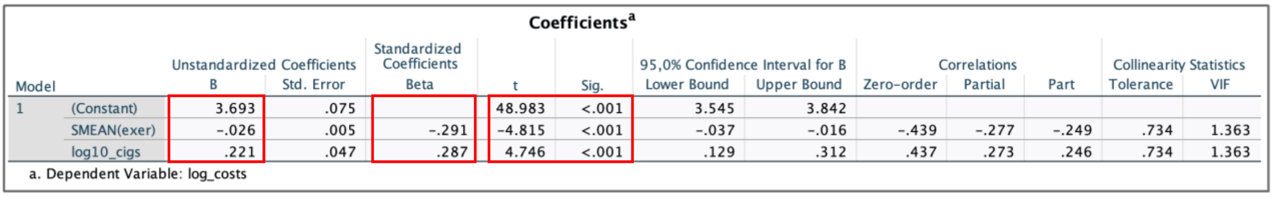


In this scatterplot the data points are evenly distributed in the area. Based on the scatterplot, there is no heteroscedasticity and therefore the homoscedasticity check is passed.

**Conclusions**: The model meets nearly all the key assumptions for multiple linear regression, except the slight deviation from linearity. This suggests that, overall, the model is well-suited for the data with a minor exception.

4. **Interpreting the effects of cigarette consumption and exercise on health care costs**

To interpret the effects of cigarette consumption and exercise on health care costs in a multiple linear regression model, we need to look at the regression coefficients associated with each independent variable. The coefficients are presented in the picture below:



Coefficients[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | 95,0% Confidence Interval for B | | Correlations | | | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Lower Bound | Upper Bound | Zero-order | Partial | Part | Tolerance | VIF |
| 1 | (Constant) | 3.693 | .075 | | 48.983 | <.001 | 3.545 | 3.842 | | | | | |
| | SMEAN(exer) | -.026 | .005 | -.291 | -4.815 | <.001 | -.037 | -.016 | -.439 | -.277 | -.249 | .734 | 1.363 |
| | log10_cigs | .221 | .047 | .287 | 4.746 | <.001 | .129 | .312 | .437 | .273 | .246 | .734 | 1.363 |

a. Dependent Variable: log_costs

Here's what I will be looking at:

- **Unstandardized Coefficient (B) Constant**: This is the intercept, representing the predicted value of the dependent variable when all independent variables are zero.
- **Unstandardized Coefficients (B) for each predictor**: These values represent the change in the dependent variable for a one-unit change in the predictor. Positive values indicate a positive relationship, and negative values indicate a negative relationship.
- **Standardized Coefficients (Beta):** These coefficients show the relative importance of each predictor variable by standardizing them to a common scale. Larger absolute values of Beta indicate stronger effects on the dependent variable.
- **Significance aka the p-value** for each coefficient tests the null hypothesis that the coefficient is zero. A p-value below 0.05 typically indicates that the predictor has a statistically significant effect on the dependent variable.
- **The t-value** tests whether each individual predictor (independent variable) has a statistically significant relationship with the dependent variable, after accounting for the effects of other variables in the model.

Based on the standardized coefficients, increase in exercise lowers the health care costs and increase in cigarette consumption raises the health care costs. This is, because exercise has a negative coefficient and cigarettes a positive coefficient. However, since the health care costs are logarithmic, the effects are multiplicative, not additive.

Based on the unstandardized B-values I have concluded the following formula:

$$\log_{10}(Health\ care\ costs) = 3.693 - 0.026(Hours\ of\ exercise\ per\ week) + 0.221(\log_{10}(Avg\ of\ cigarettes\ per\ day +\ 1).$$

Exercise values can be placed directly in the formula, as they are used in their original scale in the data analysis. The value of average consumption of cigarettes is logarithmic because common logarithm is used in the regression model. When the dependent variable is log_costs, this formula predicts the common logarithm of the health care costs. Therefore, the formula for total health care costs is:

$$Health\ care\ costs = 10^{3.693-0.026(Hours\ of\ exercise\ per\ week)+0.221(\log_{10}(Avg\ of\ cigarettes\ per\ day+\ 1)}.$$

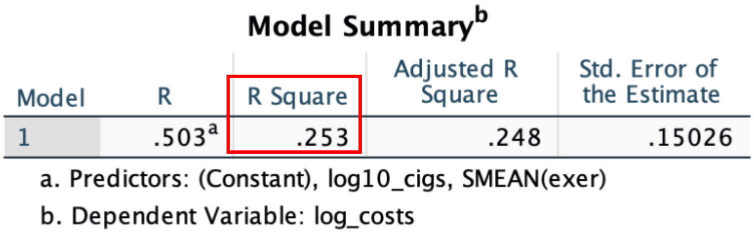
Since all p-values are less than 0.001, it can be concluded that each predictor is statistically significant. This indicates strong evidence that both exercise and cigarette consumption (log-transformed) have meaningful effects on health care costs.

The negative t-value (-4.815) for exercise suggests that exercise has a significant negative effect on health care costs. A larger absolute t-value indicates a stronger effect, and here, a t-value of -4.815 reflects a strong inverse relationship between exercise and health care costs. The positive t-value (4.746) for log10_cigs indicates that cigarette consumption has a significant positive effect on health care costs.

In conclusion, the model suggests a negative relationship between exercise and health care costs and a positive relationship between cigarette consumption and health care costs. All p-values in the model are below 0.001, meaning both predictors are statistically significant contributors to explaining health care costs. Given the multiplicative nature of these effects due to the log transformation, each predictor's impact on costs is proportional rather than absolute.

## 5. **Predictive power of the model based on R Square value**

The predictive power of the model can be interpreted from the R Square value presented in the picture below:



**Model Summary**[b]

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .503[a] | .253 | .248 | .15026 |

a. Predictors: (Constant), log10_cigs, SMEAN(exer)

b. Dependent Variable: log_costs

An R Square value of 0.253 (or 25.3%) is relatively low, indicating moderate predictive power. This means that this model only captures about a quarter of the total variability in health care costs and that 74.7% of the variance in health care costs is explained by other factors not included in this model. These other influencing factors could be other data features in the dataset used in this assignment, or they could be other factors not included in the data.

Models with higher R Square values generally have better predictive power. However, even with a lower R Square value, this model may still provide useful insights into the relationships among variables as it is statistically significant based on the low p-values.

In conclusion, an R Square value of 0.253 indicates that the model provides a moderate level of predictive power, explaining about a quarter of the variability in health care costs. While it does not capture all the factors influencing health care costs, it suggests that exercise and cigarette consumption are statistically significant contributors.

## 6. **Conclusions**

The assignment was to examine the effects of cigarette consumption and exercise on health care cost and write a data analysis project report based on the data analysis. In this report, I have described the phases and steps of my data analysis with SPSS.

I started by identifying the independent and dependent variables, which in this case were consumption of cigarettes and total hours of exercise (independent) and total health care costs (dependent). Then I explained that multiple linear regression will be used as there are more than 1 independent variables.

Then I conducted data preprocessing where I replaced missing values and transformed data to reduce skewness. After this I conducted my regression analysis and checked the basic assumptions of multiple linear regression model.

Lastly, I interpreted that based on this data analysis, exercise has a significant negative effect and consumption of cigarettes has a significant positive effect on health care costs. I also interpreted, that my model has moderate predicting power and other factors, in addition to exercise and cigarette consumption, affect health care costs as well.