Mikael Kankaanpää

ASSIGNMENT 3: DATA ANALYSIS PROJECT REPORT

Effects of Cigarette Consumption and Exercise on Health Care Cost

TIJO.220 Data Analytics in Decision-Making
H291883
November 18, 2024

TABLE OF CONTENTS

1. INTRODUCTION, VARIABLES & MODEL SELECTION

This study examines the effect of cigarette consumption and exercise on health care costs, with a particular focus on whether these lifestyle factors contribute to variations in health care expenses. Using data from a sample of individuals surveyed in 2020, the aim is to understand if higher cigarette consumption and different levels of exercise are associated with increased health care costs. A multiple linear regression model will be employed to analyze these relationships.

The dependent variable in this analysis is *costs*, representing the total health care costs declared over 2020. This variable captures the primary outcome of interest and will serve as the variable we aim to predict.

The independent variables used in the analysis are the following:

- *cigs* (average consumption of cigarettes per day): This variable measures cigarette consumption and is included to evaluate its effect on health care costs.
- *exer* (average hours of exercise per week): This variable measures exercise levels and is included to examine how physical activity may impact health care expenditures.

Table 1 summarizes the selection of variables and describes them in more detail.

Table 1: Selected variables and their details

Variable Type	Name	Data Type	Description
Dependent Variable	costs	Continuous	Total health care costs declared over 2020
Independent Variable	cigs	Discrete	Average consumption of cigarettes per day (pcs.)
	exer	Discrete	Average hours of exercise per week

A multiple linear regression model was chosen for this analysis due to its suitability in evaluating the linear relationship between a continuous dependent variable (costs) and multiple independent variables (cigs, exer). Multiple linear regression helps in isolating the effects of each independent variable, making predictions, and understanding the magnitude and direction of their impacts. Additionally, it provides statistical measures to assess the significance and fit of the model

2. DATA PREPROCESSING

This section describes the data preprocessing actions taken to ensure data integrity and suitability for regression analysis.

2.1 Missing Values

First, the dataset was examined for missing values across all variables included in the analysis. Using SPSS' descriptive statistics, it was found that a small proportion of missing data was present in both independent variables (2 instances respectively, see Table 3). As the number of missing values for each variable represents only 0,7 % of all cases (n = 282), the cases with missing values were excluded via listwise deletion.

2.2 Outliers

Outliers were examined using boxplots (Figure 1) and z-scores, where extreme values are typically defined as z-scores greater or lesser than ±3[1]. Table 2 presents the highest and lowest values and their z-scores for each variable.

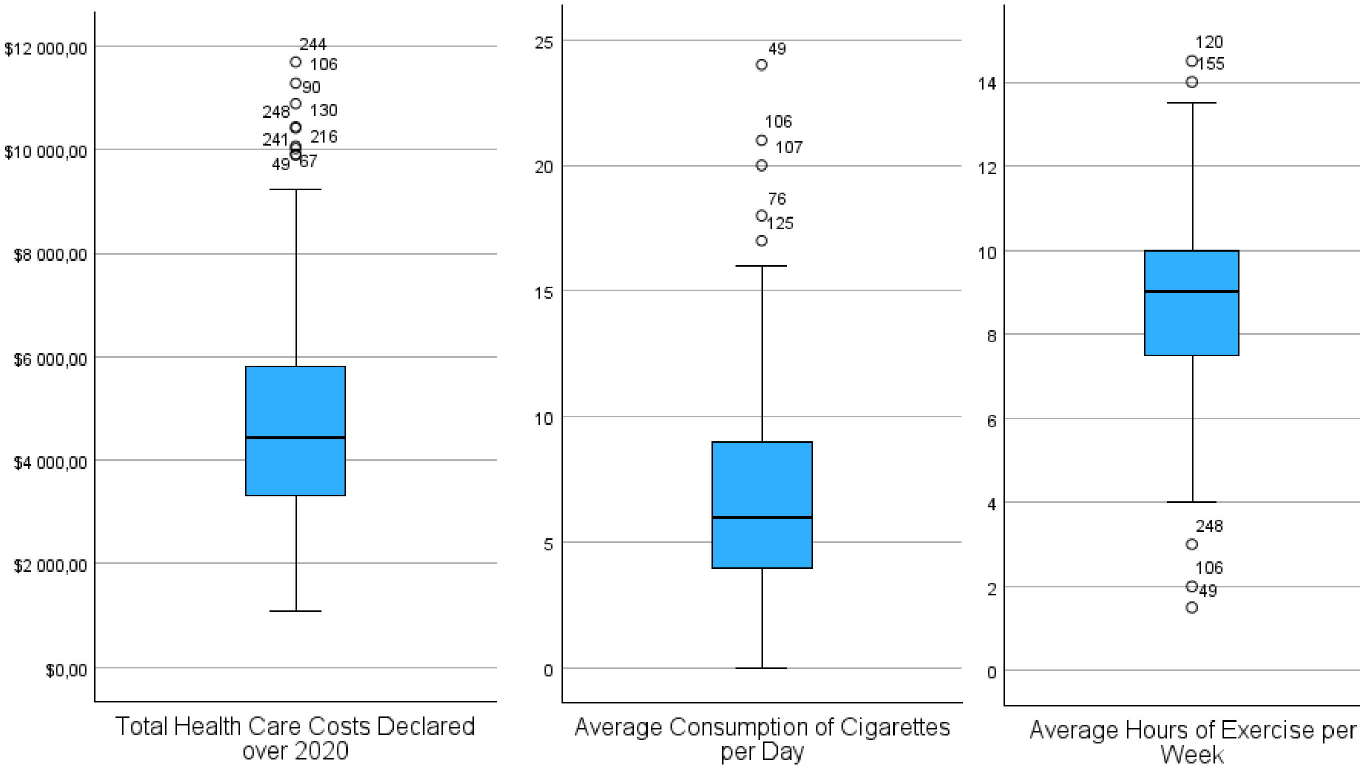


Figure 1: SPSS Outlier Boxplots

[1] According to: Fearn, T. (2008). Interpreting z- and T-Statistics. *NIR News*, 19, 15-16. https://doi.org/10.1255/nirn.1080.

Table 2: Extreme values and z-scores for each variable

			Case No.	Value	Z-Score
Total Health Care Costs Declared over 2020	Highest	1	244	$11 698,02	3,539
		2	106	$11 288,69	3,329
		3	90	$10 892,69	3,126
	Lowest	1	154	$1 095,77	-1,900
		2	28	$1 408,17	-1,740
		3	232	$1 702,88	-1,588
Average Consumption of Cigarettes per Day	Highest	1	49	24	4,369
		2	106	21	3,617
		3	107	20	3,366
	Lowest	1	175	0	-1,650
		2	40	0	-1,650
		3	252	1	-1,399
Average Hours of Exercise per Week	Highest	1	120	15	3,029
		2	155	14	2,769
		3	154	14	2,509
	Lowest	1	49	2	-3,726
		2	106	2	-3,466
		3	248	3	-2,946

The highest z-scores for health care costs (*costs)* exceed the ±3 threshold, indicating the presence of extreme values. SPSS' outlier boxplots indicated further extreme values (i.e. values not included in Table 2) for the variable, although these values all had z-scores less than 3. The lowest *costs* z-scores do not suggest extreme outliers, a result that is validated by the boxplot analysis as well. To handle the extremely high values, a log transformation is applied to *costs* to reduce their impact. These outliers in *costs* will be retained post-transformation, as the values are still plausible and there is no indication that they are a result of errors in data collection, measurement, or entry.

The highest z-scores for cigarette consumption (*cigs)* exceed the ±3 threshold, with two additional potential outliers identified by the boxplot analysis. Conversely, the lowest z-scores remain within an acceptable range, and no outliers are noted among the low values in the boxplot analysis either. Similar to *costs*, these outliers should not be excluded from the analysis, as there is no indication of data entry errors or unrealistic values. Elevated cigarette consumption may represent legitimate behavior, offering valuable insights into its impact on health care costs.

Regarding average hours of weekly exercise (*exer)*, both the highest and lowest z-scores exceed the ±3 threshold, with additional potential outliers noted in the boxplot analysis. These extreme values may disproportionately influence the regression model, yet they can be argued to still be plausible. To assess their effect, the analysis will be conducted both with and without these outliers, allowing for a sensitivity analysis to help evaluate if they unduly influence the model.

2.3 Variable Skewness

The skewness of each variable was assessed using descriptive statistics in SPSS (Table 3), with the following results:

- **Health care costs (*costs*) exhibited a skewness of 1,008**, indicating moderate positive skewness. Since this value is above 1, it further justifies the log transformation[2] proposed in Section 2.2 by reducing skewness, stabilizing variance, and improving model fit for regression analysis.
- **Cigarette consumption *(cigs)* showed a skewness of 1,147**, indicating moderate positive skewness. Given this level of skewness, a log transformation[2] was applied to the variable, with constant c = 1 added to all the values to account for 0 values.
- **Average hours of weekly exercise *(exer)* had only a slight negative skewness of -0,206**, suggesting that it is approximately normally distributed. Therefore, no transformation was necessary for this variable.

Table 3: Results of SPSS' Descriptive Statistics analysis

		Total Health Care Costs Declared over 2020 (*costs*)	Average Consumption of Cigarettes per Day (*cigs*)	Average Hours of Exercise per Week (*exer*)
N	Valid	282	280	280
	Missing	0	2	2
Mean		$4,796.9013	6,56	8,67
Std. Deviation		$1,940.89838	3,980	1,918
Skewness		1,008	1,147	-,206
Std. Error of Skewness		,145	,146	,146
Minimum		$1,095.77	0	2
Maximum		$11,698.02	24	15

2.4 Correlations

Correlation analysis was conducted via SPSS to examine the relationships between the log-transformed variables *log_cigs, log_costs*, and *exer*. The results revealed the following:

- **Correlation between *log_cigs* and *log_costs***: A moderate positive correlation (r = 0,45, p < 0,01) was observed, indicating that higher cigarette consumption is associated with higher health care costs. This supports the hypothesis that smoking contributes to increased health-related expenses.
- **Correlation between *log_cigs* and *exer***: A moderate negative correlation (r = -0,505, p < 0,01) was found, suggesting that individuals who smoke more tend to exercise less. This relationship underscores the inverse association between these two lifestyle behaviors.

[2] Natural logarithm (ln) transformation was used, as it is widely utilized in statistical analysis and makes interpretation of regression model coefficients straightforward and intuitive with percentage changes.

- **Correlation between *log_costs* and *exer***: A moderate negative correlation (r = -0,44, p < 0,01) was identified, indicating that greater exercise levels are associated with lower health care costs. This aligns with the idea that physical activity may mitigate health care expenditures.

The results of the correlation analysis are depicted in Table 4 below.

Table 4: Correlation Analysis results

		Total Health Care Costs Declared over 2020 (log_*costs*)	Average Consumption of Cigarettes per Day (log_*cigs*)	Average Hours of Exercise per Week (*exer*)
Total Health Care Costs Declared over 2020 (log_*costs*)	Pearson Correlation	1	,450**	-,440**
	Sig. (2-tailed)		<,001	<,001
	N	278	278	278
Average Consumption of Cigarettes per Day (log_*cigs*)	Pearson Correlation	,450**	1	-,505**
	Sig. (2-tailed)	<,001		<,001
	N	278	278	278
Average Hours of Exercise per Week (*exer*)	Pearson Correlation	-,440**	-,505**	1
	Sig. (2-tailed)	<,001	<,001	
	N	278	278	278

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

These correlations provide preliminary evidence for significant relationships among the key variables. Although the correlation between *log_cigs* and *exer* is relatively strong (r = -0,505), it does not indicate multicollinearity concerns, as it is below the threshold of 0,7. Therefore, both variables can be included in the regression model without compromising its interpretability.

3. ASSUMPTION CHECKS FOR THE MODEL

This section of the report describes the actions taken to ensure that the key assumptions underlying the analysis are satisfied. Each assumption is examined using appropriate statistical tests and visual diagnostics to ensure the robustness of the model and the validity of the conclusions.

3.1 Linearity

Linearity tests were conducted in SPSS to analyze relationships between both predictors and the dependent variable. The tests indicated linearity between the predictors and the dependent variable at p < 0,001 significance. The results of these tests are depicted in Table 5 and Table 6 below.

Table 5: SPSS Linearity test results between health care costs and cigarette consumption

			Sum of Squares	df	Mean Square	F	Sig.
Log_costs * log_cigs	Between Groups	(Combined)	13,056	21	,622	5,068	<,001
		Linearity	8,994	1	8,994	73,321	<,001
		Deviation from Linearity	4,062	20	,203	1,656	,041
	Within Groups		31,401	256	,123		
	Total		44,456	277			

Table 6: SPSS Linearity test results between health care costs and cigarette consumption

			Sum of Squares	df	Mean Square	F	Sig.
Log_costs * exer	Between Groups	(Combined)	13,238	24	,552	4,470	<,001
		Linearity	8,623	1	8,623	69,879	<,001
		Deviation from Linearity	4,615	23	,201	1,626	,038
	Within Groups		31,219	253	,123		
	Total		44,456	277			

Visual representations of the linearity between the predictors and dependent variable are depicted in the following Figure 2.

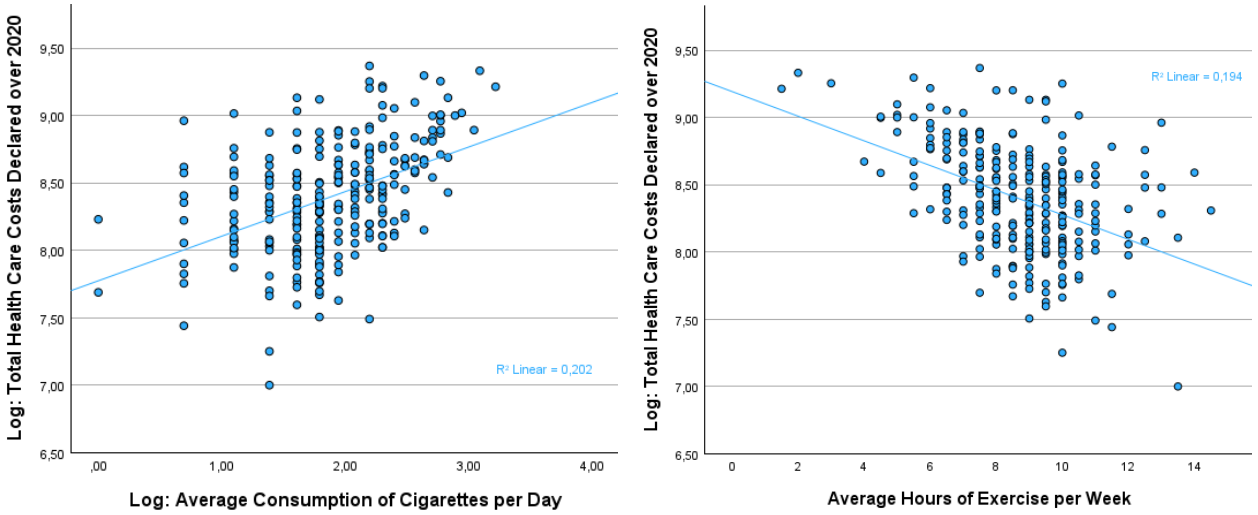


Figure 2: Linearity between the predictors and dependent variable

3.2 Homoscedasticity

A plot of standardized residuals against standardized predicted values was generated to test for homoscedasticity (Figure 3). Visual inspection indicated a random distribution of residuals, supporting the assumption of homoscedasticity. This suggests that the variance in errors remains consistent across levels of the predicted variable.

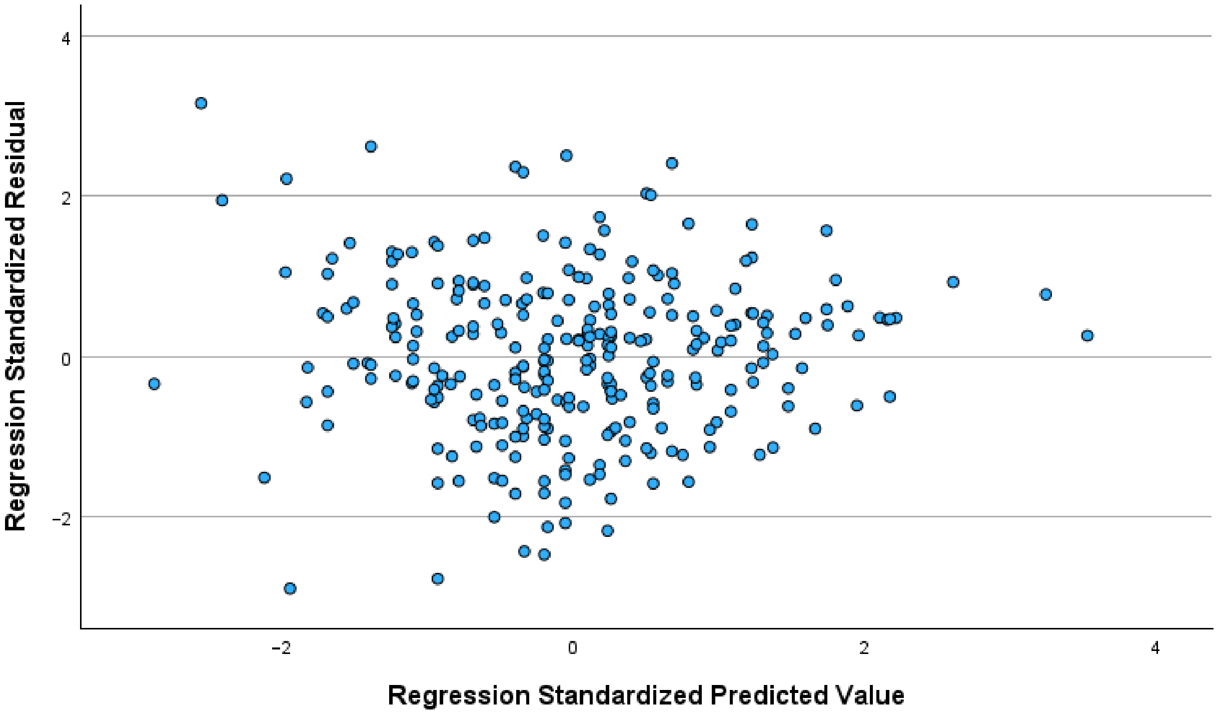


Figure 3: Scatter plot of residuals against predicted values

3.3 Normality of Residuals

The normality of residuals was visually assessed using a P-P plot and a histogram of standardized residuals (Figure 4).

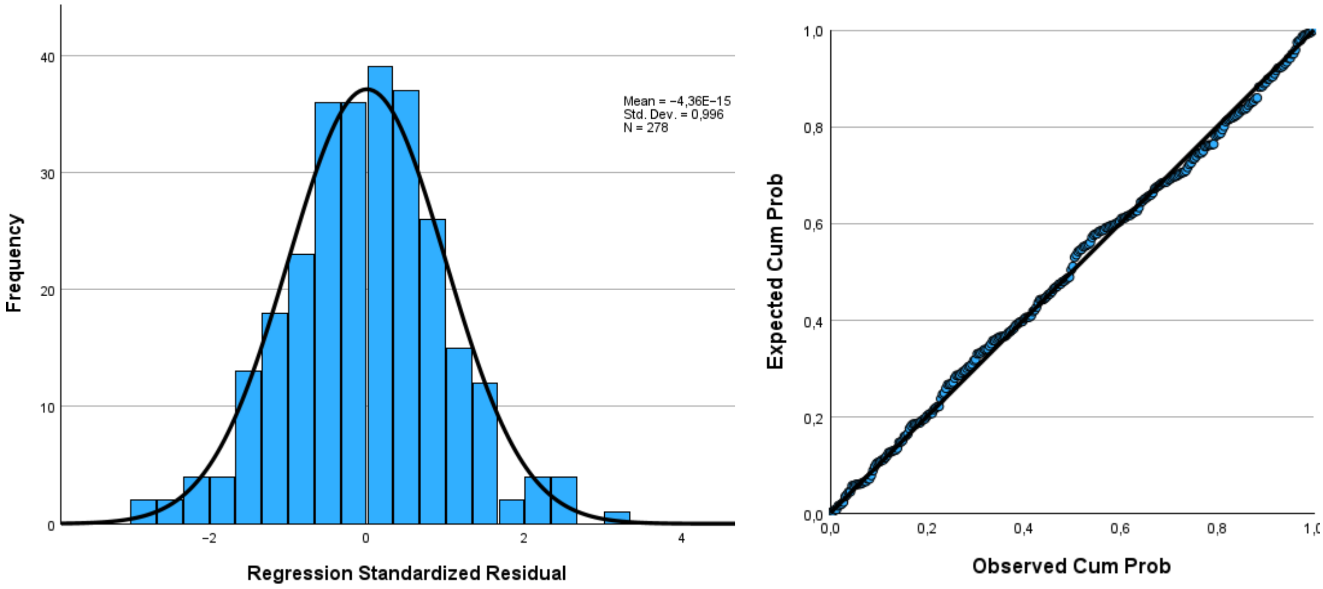
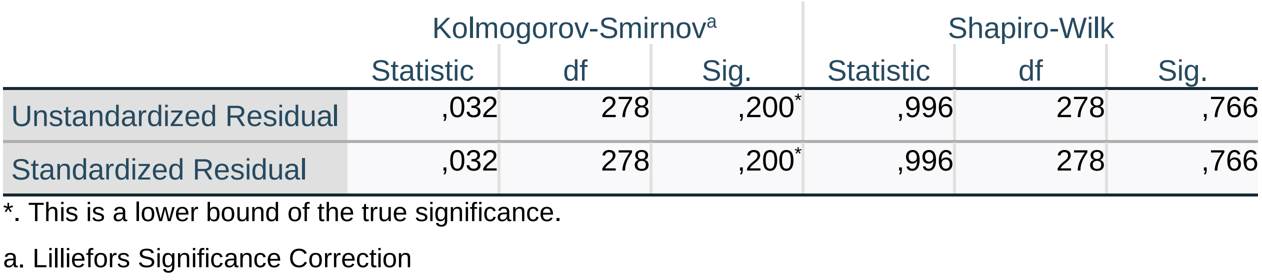


Figure 4: Histogram and P-P plot used in normality assessment

In addition, Kolmogorov-Smirnov (K-S) and Shapiro-Wilk (S-W) were run in SPSS (Table 7), indicating that the residuals of the regression model do not significantly deviate from normality, which allows interpreting the regression coefficients reliably.

Table 7: Results from K-S and S-W tests of normality



	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,032	278	,200*	,996	278	,766
Standardized Residual	,032	278	,200*	,996	278	,766

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

a. Lilliefors Significance Correction

3.4 Multicollinearity

Variance Inflation Factors (VIFs) were checked for both predictors. The VIF values were 1,343 < 10 for both as shown in Table 8, indicating no issues with multicollinearity. This means the independent variables are not excessively correlated and thus appropriate for inclusion in the model.

4. MODEL INTERPRETATION

A multiple linear regression analysis was conducted to examine the effects of cigarette consumption (*log_cigs*) and exercise (*exer*) on health care costs (*log_costs*). The results indicated that both predictors were statistically significant contributors to the model, which provides insights into their individual effects on health care costs. Table 8 below contains the regression coefficients for the model.

Table 8: Regression coefficients

		(constant)	log_cigs	exer
Unstandardized Coefficients	B	8,491	0,224	-0,060
	Std. Error	0,168	0,044	0,012
Standardized Coefficients	Beta		0,305	-0,286
	t-value	50,577	5,087	-4,771
	Significance	<,001	<,001	<,001
95,0% CI for B	Lower Bound	8,160	0,137	-0,084
	Upper Bound	8,821	0,311	-0,035
Correlations	Zero-order		0,450	-0,440
	Partial		0,293	-0,276
	Part		0,263	-0,247
Collinearity Statistics	Tolerance		0,745	0,745
	VIF		1,343	1,343

The model coefficients are interpreted in more detail in the following subsections.

4.1 Intercept (*constant*)

The intercept (8,491) represents the expected value of the dependent variable (*log_costs*) when all independent variables are zero. As the dependent variable is log-transformed, the actual amount of health care costs is $e^{8,491}$ ≈ 4 891,94 $, when *log_cigs* and *exer* are 0. In other words, this value provides a baseline level of health care costs in the absence of cigarette consumption and exercise.

4.2 Cigarette Consumption (*log_cigs*)

The coefficient for *log_cigs* (0,224) indicates the expected increase in *log_costs* for a one-unit increase in *log_cigs* in a logarithmic scale, when holding exercise constant. In practical terms, a 1%

increase in cigarette consumption is associated with a 0,224% increase in health care costs, 10% increase is associated with 0,244 · ln(1,10) ≈ 0,021 or 2,1% increase in health care costs, etc. This positive and statistically significant relationship (p < 0,001) suggests that higher cigarette consumption is associated with higher health care costs. The statistical significance (p < 0,001) and t-value (t = 5,087 > 3,29) both indicate that this relationship is very unlikely to be due to chance, whereas the standardized coefficient (Beta = 0,305) shows the relative importance of cigarette consumption in predicting health care costs.

4.3 Exercise (*exer*)

The coefficient for *exer* (-0.060) represents the expected change in *log_costs* for a one-unit increase in *exer*, holding cigarette consumption constant. A negative coefficient suggests that higher levels of exercise are associated with lower health care costs. In practice, the decrease in health care costs is 1 - $e^{-0,060}$ ≈ 0,059 or 5,9% per hour of exercise. The statistical significance (p < 0,001) and t-value (t = -4,771 < -3,29) both support the reliability of this relationship. The standardized coefficient (Beta = -0,286) indicates the relative importance of exercise in predicting health care costs.

4.4 Conclusions

The results indicate that cigarette consumption increases health care costs, while exercise reduces them. Both predictors are statistically significant and contribute meaningfully to the model. These findings emphasize the financial burden of smoking and the potential cost-saving benefits of physical activity in reducing health care expenses.

5. PREDICTIVE POWER OF THE MODEL

The model's R-squared value was 0,263 (see Table 9), indicating that approximately 26,3% of the variance in health care costs can be explained by cigarette consumption and exercise. This suggests that while the **model has medium predictive power**, other factors likely contribute to health care costs that are not included in this analysis. The adjusted R-squared value, which accounts for the number of predictors, was 25,8%. This value closely aligns with the R-squared, affirming that our predictors (log_*cigs* and *exer*) meaningfully contribute to explaining health care costs.

The F-value is F(2, 275) = 49,136 with a significance level of p < 0,001 as indicated by the regression ANOVA results (Table 10). This confirms that the model as a whole is statistically significant, meaning that cigarette consumption and exercise together significantly predict health care costs.

Table 9: SPSS' Regression Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,513	,263	,258	,34511

Table 10: SPSS' Regression ANOVA Results

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11,704	2	5,852	49,136	<,001
	Residual	32,752	275	,119		
	Total	44,456	277			