ECONOMETRICS REPORT ASSIGNMENT 2

CANDIDATE NUMBER: 702115

BS2280 - ECONOMETRICS 1

1. Introduction

The global number of exceptionally long-lived individuals is increasing due to medical and social advances (Jabr, 2021). This surge in longevity intensifies the exploration of life expectancy factors. However, the complexity of variables and their interactions encourages continued investigation, especially studies that can potentially guide evidence-based policies and interventions related to socioeconomic and public health challenges. As highlighted by Lhachimi et al. (2016), it can benefit society by improving the probability of successful initiatives and policies and optimising the allocation of public and private resources. That is the aim of this report by critically analysing the relationship between socioeconomic, and health factors on average life expectancy in 2011.

2. Methodology

The analysis uses World Health Organization and the United Nations data, featuring primary data for 171 countries in 2011. It covers world health information and socioeconomic quantitative factors like country status, life expectancy, alcohol consumption, BMI, total expenditure, and schooling years. Employing R programming and econometric concepts, four models are evaluated.

The first, a Simple Linear Model, assumes a linear relationship between life expectancy and alcohol consumption:

$$Life\ Expectancy_i = \beta_1 + \beta_2\ Alcohol_i + \varepsilon_i \tag{1}$$

The second, a Multiple Regression Model, incorporates alcohol, schooling, and BMI to measure their combined impact:

$$Life\ Expectancy_i = \beta_1 + \beta_2\ Alcohol_i + \beta_3\ Schooling_i + \beta_4\ BMI_i + \varepsilon_i \tag{2}$$

The third Multiple Regression Model introduces a squared term to explore a potential non-linearity in schooling's impact:

$$Life\ Expectancy_i = \beta_1 + \beta_2\ Alcohol_i + \beta_3\ Schooling_i + \beta_4\ Schooling_i^2 + \beta_5\ BMI_i + \varepsilon_i \tag{3}$$

The last model applies logarithmic transformations to life expectancy and schooling, aiming to capture their potential non-linear relationship:

$$\log(Life\ Expectancy_i) = \beta_1 + \beta_2\ Alcohol_i + \beta_3 \log(Schooling_i) + \beta_4\ BMI_i + \varepsilon_i \tag{4}$$

3. Results and Analysis

In this section, the models are going to be estimated and discussed to understand which better explains life expectancy.

3.1. Simple Linear Model

Table 1: Simple Linear Model Results

	Dependent variable:
	Life Expectancy
Alcohol	0.837
	(0.158)
	t = 5.305
	p = 0.00000
Constant	66.858
	(0.976)
	t = 68.528
	p = 0.000
Observations	171
R^2	0.143
Adjusted R^2	0.138
Residual Std. Error	8.029 (df = 169)
F Statistic	28.145^{***} (df = 1; 169)
Note:	*p<0.1; **p<0.05; ***p<0.01

Based on the results, the estimated regression model is:

$$\widehat{Life\ Expectancy_i} = 66.858 + 0.837\ Alcohol_i + \varepsilon_i \tag{5}$$

The model predicts that without alcohol consumption, life expectancy would average 66.86 years. For every 1-litre increase in alcohol consumption, life expectancy would increase by 0.84 years. In this sense, a test hypothesis to test the significance of intercept and coefficient can be written in the following way:

Intercept $H_0: \beta_1 = 0, \quad H_1: \beta_1 \neq 0$ (6)

Coefficient $H_0: \beta_2 = 0, \quad H_1: \beta_2 \neq 0$ (7)

By calculating t-statistics for intercept and coefficient of alcohol consumption, considering critical t-value at 5% significance level equals 1.97, the following results can be obtained:

Intercept
$$t - value = \frac{\hat{\beta}_1 - B_1^0}{s.e.(\hat{\beta}_1)} = \frac{66.858 - 0}{0.976} \approx 68.53 > t_{\text{crit},5\%} = 1.97 \quad (8)$$

Coefficient
$$t - value = \frac{\hat{\beta}_2 - B_2^0}{s.e.(\hat{\beta}_2)} = \frac{0.837 - 0}{0.158} \approx 5.306 > t_{\text{crit},5\%} = 1.97 \quad (9)$$

As a result, the null hypothesis (H_0) is rejected in both cases, so the intercept and the alcohol coefficient are statistically significant. However, as can be seen in Table 1, the overall fit (R^2) is low, as only 14.3% of the variation in life expectancy can be explained by the variation in alcohol consumption. Also, going back to the interpretation of coefficients, an increase in alcohol consumption leading to an improvement in life expectancy is a counterintuitive statement, as, in practice, this is likely to increase health problems and social harm (Rehm, 2011).

These issues can be explained by the fact that the alcohol coefficient of this model may be subject to omitted variable bias, as it is a very simplistic assumption about what factors explain the evolution of life expectancy by not considering other relevant variables (Dougherty, 2016, p. 263). As shown in Figure 1, the data points on the relationship between both variables indicate a more complex relationship as there is no clear linear pattern and a heterogeneous distribution, being slightly concentrated at the graph's top left side.

Figure 1- Relationship between Life Expectancy and Alcohol with Model 1 Regression Line

3.2. Multiple Regression Model I

Table 2: Multiple Regression Model I Results

	Dependent variable:
	Life Expectancy
Alcohol	-0.142
	(0.118)
	t = -1.208
	p = 0.229
Schooling	2.221
	(0.185)
	t = 12.035
	p = 0.000
BMI	0.086
	(0.023)
	t = 3.833
	p = 0.0002
Constant	39.901
	(1.849)
	t = 21.578
	p = 0.000
Observations	171
R^2	0.671
Adjusted R^2	0.665
Residual Std. Error	5.001 (df = 167)

F Statistic	113.712^{***} (df = 3; 167)
Note:	*p<0.1; **p<0.05; ***p<0.01

According to the results, the estimated regression model is:

$$\widehat{Life\ Expectancy}_i = 39.901 - 0.142\, Alcohol_i + 2.221\, Schooling_i + 0.086\, BMI_i + \varepsilon_i \quad (10)$$

Interpreting the intercept, if all covariates were zero, the average life expectancy would be 39.90 years. In addition, as the p‒value (2e – 16) is lower than 5%, the null hypothesis can be rejected. Therefore, the intercept is highly statistically significant at a 5% significance level, which will also be considered in the following analysis.

Concerning the alcohol coefficient, a 1-litre increase predicts a 0.14-year decrease in life expectancy, but its p-value (0.229) is above 5%, rendering it statistically insignificant. In contrast, the schooling variable exhibits high statistical significance as the p‒value (2e – 16) is lower than 5%, indicating a 1-year increase in schooling corresponds to a 2.22-year rise in life expectancy, *ceteris paribus*.

Interpreting the BMI coefficient, a 1-unit increase leads to a 0.09-year rise in life expectancy, *ceteris paribus*. Moreover, the p-value (0.0002) is below 5%, so the null hypothesis can be rejected, indicating statistical significance.

Analysing the model Goodness-of-Fit (R^2), it can be considered high given that 67.1% of the variation in life expectancy can be explained by the model, which is relatively close to 100%. Despite its high value, an F-test will assess if introducing new variables enhances explanatory power, as seen in the following steps.

Table 3: Simple Linear Regression Model Analysis of Variance

	Df	Sum Sq	Mean Sq	F value	Pr(F)
Alcohol	1	1814.21	1814.21	28.15	0.0000
Residuals	169	10893.55	64.46		

Table 4: Multiple Regression Model I Analysis of Variance

	Df	Sum Sq	Mean Sq	F value	Pr(F)
Alcohol	1	1814.21	1814.21	72.54	0.0000
Schooling	1	6349.63	6349.63	253.90	0.0000
BMI	1	367.48	367.48	14.69	0.0002
Residuals	167	4176.44	25.01		

Stating null and alternative hypotheses:

$$H_0: \beta_3 = \beta_4 = 0, \qquad H_1: \beta_3 \neq 0 \textit{ or } \beta_4 \neq 0 \textit{ or both } \beta_3 \textit{ and } \beta_4 \neq 0 \tag{11}$$

Calculating F-test statistics:

$$F(cost\ of\ df, df\ remaining) = \frac{\frac{\text{reduction in RSS}}{\text{cost in df}}}{\frac{RSS\ remaining}{df\ remaining}} \tag{12}$$

$$= \frac{\frac{(10{,}893.55 - 4{,}176.44)}{2}}{\frac{4{,}176.44}{167}} = 134.2959$$

$$F_{\text{crit,5\%}}(\text{cost in dof, dof remaining}) = F_{\text{crit,5\%}}(2, 167) = 3.05 \tag{13}$$

$$F_{\text{crit,5\%}}(2, 167) = 3.05 < F = 134.2959 \tag{14}$$

Thus, the null hypothesis (H_0) is rejected, so the new set of variables added to the model does have significant explanatory power, improving the model fit in this way.

3.3. Multiple Regression Model II

Table 5: Multiple Regression Model II Results

	Dependent variable:
	Life Expectancy
Alcohol	-0.155
	(0.119)
	t = -1.304
	p = 0.195
Schooling	1.606
	(0.875)
	t = 1.835
	p = 0.069
$Schooling^2$	0.025
	(0.035)
	t = 0.719
	p = 0.474
BMI	0.088
	(0.023)
	t = 3.870
	p = 0.0002
Constant	43.493
	(5.329)
	t = 8.162
	p = 0.000
Observations	171
R^2	0.672
Adjusted R^2	0.664
Residual Std. Error	5.008 (df = 166)
F Statistic	85.166^{***} (df = 4; 166)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

According to the results, the estimated regression model is:

$$\widehat{Life\ Expectancy_i} = 43.493 - 0.155\, Alcohol_i + 1.606\, Schooling_i + 0.025\, Schooling_i^2 + 0.088\, BMI_i + \varepsilon_i \quad (15)$$

Interpreting the intercept if all covariates were zero, the average life expectancy would be 43.49 years. The alcohol coefficient suggests that an additional 1 litre decreases life expectancy by 0.16 years, *ceteris paribus*. Also, the BMI coefficient suggests that a 1-unit rise of it increases life expectancy by approximately 0.09 years, *ceteris paribus*.

Regarding the quadratic schooling coefficient, as it is a quadratic explanatory variable, *ceteris paribus* cannot be applied. Therefore, the marginal effect of schooling on life expectancy can be calculated in the following way:

$$\frac{d\widehat{Life\ Expectancy_i}}{dSchooling} = \beta_3 + 2\beta_4 Schooling = 1.606 + 0.05\ Schooling \qquad (16)$$

Its magnitude will depend on the schooling value. For instance, if Schooling = 0,

$$\frac{d\widehat{Life\ Expectancy_i}}{dSchooling} = 1.606 + 0.05 \times 0 = 1.606 \qquad (17)$$

a 1-year increase in schooling would lead to an increase in average life expectancy by 1.606 years.

If Schooling = 10,

$$\frac{d\widehat{Life\ Expectancy_i}}{dSchooling} = 1.606 + 0.05 \times 10 = 2.106 \qquad (18)$$

a 1-year increase in schooling results in a 2.106-year rise in average life expectancy. However, the square term lacks statistical significance (p-value = 0.719, higher than the 5% significance level), negating a non-linear effect of schooling on life expectancy.

Furthermore, the model Goodness-of-Fit (R2) explains 63.1% of the variability in life expectancy, which suggests a satisfactory model fit when estimating the dependent variable.

It is also important to compare the fit of Multiple Regression Models I and II to evaluate whether the additional squared term improved the model fit or not. So, as can be seen in

Tables 2 and 5, the adjusted R^2 slightly decreased from 0.665 to 0.664 in Model II, suggesting that the squared term did not enhance model fit.

3.4. Logarithmic Model

Table 6: Logarithmic Model Results

	Dependent variable:
	ln(Life Expectancy)
Alcohol	-0.001
	(0.002)
	t = -0.721
	p = 0.472
ln(Schooling)	0.344
	(0.032)
	t = 10.677
	p = 0.000
BMI	0.001
	(0.0004)
	t = 4.212
	p = 0.00005
Constant	3.334
	(0.072)
	t = 46.586
	p = 0.000
Observations	171
R^2	0.631
Adjusted R^2	0.624
Residual Std. Error	0.078 (df = 167)
F Statistic	95.029^{***} (df = 3; 167)
Note:	*p<0.1; **p<0.05; ***p<0.01

Based on the results above, the estimated regression model is:

$$\log(\widehat{Life\ Expectancy_i}) = 3.334 - 0.001\, Alcohol_i + 0.344 \log(Schooling_i) + 0.001\, BMI_i + \varepsilon_i \qquad (19)$$

Interpreting coefficients reveals the alcohol and BMI coefficients as part of a semi-log model. A 1-litre increase in alcohol consumption correlates with a 0.1% life expectancy decrease, while a 1-unit BMI rise corresponds to a 0.1% increase, *ceteris paribus*.

For the schooling coefficient, interpreted in a log-log model, the elasticity of life expectancy concerning schooling was 0.344, knowing that the coefficient is highly significant statistically as the p−value (2e – 16) is lower than 5%. So, a 1% increase in the average years of schooling leads to a 0.344% increase in life expectancy.

Also, the model Goodness-of-Fit (R^2) can be considered high given that it explains 63.1% of the life expectancy variation, which is relatively close to 100%, suggesting a robust model.

3.5. Discussion of Results

Table 7: Comparison of Models Results

	Dependent variable:			
	Life Expectancy	Life Expectancy		ln(Life Expectancy)
	(1)	(2)	(3)	(4)
Alcohol	0.837	-0.142	-0.155	-0.001
	(0.158)	(0.118)	(0.119)	(0.002)
	t = 5.305	t = -1.208	t = -1.304	t = -0.721
	p = 0.00000	p = 0.229	p = 0.195	p = 0.472
Schooling		2.221	1.606	
		(0.185)	(0.875)	
		t = 12.035	t = 1.835	
		p = 0.000	p = 0.069	
$Schooling^2$			0.025	
			(0.035)	
			t = 0.719	
			p = 0.474	
ln(Schooling)				0.344
				(0.032)
				t = 10.677
				p = 0.000
BMI		0.086	0.088	0.001
		(0.023)	(0.023)	(0.0004)
		t = 3.833	t = 3.870	t = 4.212
		p = 0.0002	p = 0.0002	p = 0.00005
Constant	66.858	39.901	43.493	3.334
	(0.976)	(1.849)	(5.329)	(0.072)
	t = 68.528	t = 21.578	t = 8.162	t = 46.586
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Observations	171	171	171	171
R^2	0.143	0.671	0.672	0.631
Adjusted R^2	0.138	0.665	0.664	0.624
Residual Std. Error	8.029 (df = 169)	5.001 (df = 167)	5.008 (df = 166)	0.078 (df = 167)
F Statistic	28.145^{***} (df = 1; 169)	113.712^{***} (df = 3; 167)	85.166^{***} (df = 4; 166)	95.029^{***} (df = 3; 167)
Note:				*p<0.1; **p<0.05; ***p<0.01

As the report's goal is to investigate which one of the estimated models better explains the impact of the chosen factors on life expectancy, one indicator that needs to be prioritised is the Adjusted R^2. Based on the table above, Multiple Regression Models I and II have higher values of $\bar{R}^2$, suggesting a better fit. Between them, the one with a lower Residual Standard Error, indicating a more accurate prediction, and a higher percentage of statistically significant coefficients at a 5% significance level is the Multiple Regression Model I.

Figure 2 - Comparison between the Residuals Histograms of the Estimated Models

Analysing the residuals' histograms above, Model I's graph appears more symmetrical, indicative of approximately normal distribution — a characteristic of well-fitted models (Dougherty, 2016, p. 209). Considering these factors, Model I emerges as the superior choice among the highlighted models.

4. Conclusions and limitations

The exploration of diverse econometric models revealed valuable insights into life expectancy determinants. The Simple Linear Model proved to be ineffective and estimated a counterintuitive relationship between the variables, as it was overly simplistic. The Multiple Regression Model I, by incorporating factors like schooling and BMI, provided a more accurate understanding of life expectancy variations, being chosen further as the fittest model. However, the addition of a squared term lacked statistical significance and made the model prediction capacity less accurate, indicating a limited non-linear relationship with schooling. Finally, the Logarithmic Model, demonstrated a lower Adjusted R2 compared to the multiple regression models, emphasizing a limited non-linear relationship between the variables.

Also, the study's limitations include potential oversimplification, exclusion of other relevant life expectancy determinants and the need for a more detailed exploration of complex relationships among variables. So, future investigations should explore the non-linear dynamics further, addressing the limitations of each model and acknowledging the nuances revealed by more sophisticated models.

References

Dougherty, C. (2016). Introduction to Econometrics (5th ed.). Oxford: *Oxford University Press*. Retrieved from: <https://doi.org/10.1093/he/9780192655783.001.0001>.

Jabr, F. (2021). How Long Can We Live? *The New York Times*. [Online] Available at: <https://www.nytimes.com/2021/04/28/magazine/human-lifespan.html> (Accessed 31 Dec. 2023).

Lhachimi, S.K., Bala, M.M. and Vanagas, G. (2016). Evidence-Based Public Health. *BioMed Research International*. [Online] doi: <https://doi.org/10.1155/2016/5681409> (Accessed 31 Dec. 2023).

Rehm, J. (2011). The Risks Associated With Alcohol Use and Alcoholism. [Online] Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3307043/pdf/arh-34-2-135.pdf>.

Appendix

Appendix A: R commands and outputs

```
> # Preparing workspace
> setwd("C:/Users/maria/OneDrive/Documentos/3. Econometrics
1/BS2280/Coursework2")
> install.packages("xtable")

WARNING: Rtools is required to build R packages but is not currently
installed. Please download and install the appropriate version of
Rtools before proceeding:
https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/maria/AppData/Local/R/win-
library/4.3'
(as 'lib' is unspecified)
trying the URL
'https://cran.rstudio.com/bin/windows/contrib/4.3/xtable_1.8-4.zip'
Content type 'application/zip' length 706178 bytes (689 KB)
downloaded 689 KB
package 'xtable' successfully unpacked and MD5 sums checked
The downloaded binary packages are in
      C:\Users\maria\AppData\Local\Temp\RtmpysCPOT\downloaded_packages

> # Loading libraries
> library(readxl)
> library(stargazer)

Please cite as:
 Hlavac, Marek (2022). stargazer: Well-Formatted Regression and
Summary Statistics Tables.
 R package version 5.2.3. https://CRAN.R-
project.org/package=stargazer

> library(xtable)

Warning message:
package 'xtable' was built under R version 4.3.2

> # Importing the dataset
> data_2011 <- read_excel("2011lifeexpectancy.xls")
> # Model 1. Simple Linear Regression Model
> # Regressing life expectancy on alcohol consumption
> model1 <-
lm(data_2011$Life_Expectancy~data_2011$Alcohol,data=data_2011)
> model1

Call:
lm(formula = data_2011$Life_Expectancy ~ data_2011$Alcohol, data =
data_2011)

Coefficients:
      (Intercept)  data_2011$Alcohol
          66.8582             0.8373

> # Goodness-of-fit of the estimated model
> summary(model1)

Call:
```

```
lm(formula = data_2011$Life_Expectancy ~ data_2011$Alcohol, data =
data_2011)

Residuals:
    Min      1Q  Median      3Q     Max
-22.606  -4.713   1.299   6.259  14.356

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        66.8582     0.9756  68.528  < 2e-16 ***
data_2011$Alcohol   0.8373     0.1578   5.305 3.49e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.029 on 169 degrees of freedom
Multiple R-squared:  0.1428,  Adjusted R-squared:  0.1377
F-statistic: 28.15 on 1 and 169 DF,  p-value: 3.491e-07

> # Formatting the table of results
> stargazer(model1,
+          type = "html",
+          title="Table 1: Simple Linear Model Results",
+          summary = TRUE,
+          align=TRUE,
+          no.space=TRUE,
+          out = "C:/Users/maria/OneDrive/Documentos/3. Econometrics
1/BS2280/Coursework2/Model1.htm",
+          report=("vcstp"))
```

Output:

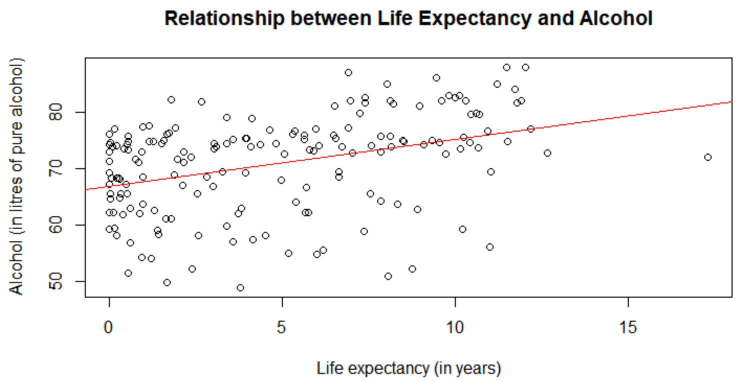
Simple Linear Model Results

	Dependent variable:
	Life_Expectancy
Alcohol	0.837
	(0.158)
	t = 5.305
	p = 0.00000
Constant	66.858
	(0.976)
	t = 68.528
	p = 0.000
Observations	171
R^2	0.143
Adjusted R^2	0.138
Residual Std. Error	8.029 (df = 169)
F Statistic	28.145^{***} (df = 1; 169)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
> # Creating a scatter plot with a regression line
> plot(data_2011$Life_Expectancy~data_2011$Alcohol,
+     main = "Relationship between Life Expectancy and Alcohol",
+     xlab = "Life expectancy (in years)",
+     ylab = "Alcohol (in litres of pure alcohol)")
```

```
> abline(model1, col = "red")
```

Output:




```
> # Obtaining ANOVA table
> anova_table1 <- anova(model1)
> print(anova_table1)

Analysis of Variance Table

Response: data_2011$Life_Expectancy
                   Df   Sum Sq Mean Sq F value    Pr(>F)
data_2011$Alcohol   1   1814.2 1814.21  28.145 3.491e-07
Residuals         169  10893.6   64.46

data_2011$Alcohol ***
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> # Creating a table from the ANOVA table
> xtable_anova1 <- xtable(anova_table1)
> # Saving it to a file
> print(xtable_anova1, type = "latex")
```

Output:

	Df	Sum Sq	Mean Sq	F value	Pr(F)
Alcohol	1	1814.21	1814.21	28.15	0.0000
Residuals	169	10893.55	64.46		

```
> # Model 2.      Multiple linear regression model I
> # Adding Schooling and BMI to the regression model
> model2 <- lm(Life_Expectancy~Alcohol+Schooling+BMI,data=data_2011)
> summary(model2)

Call:
lm(formula = Life_Expectancy ~ Alcohol + Schooling + BMI, data =
data_2011)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5203  -2.7481   0.1784   3.0953  13.7109

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.90147    1.84913  21.578  < 2e-16 ***
Alcohol     -0.14202    0.11752  -1.208 0.228568
Schooling    2.22100    0.18454  12.035  < 2e-16 ***
BMI          0.08639    0.02254   3.833 0.000179 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.001 on 167 degrees of freedom
Multiple R-squared:  0.6713,   Adjusted R-squared:  0.6654
F-statistic: 113.7 on 3 and 167 DF,  p-value: < 2.2e-16

> # Formatting the table of results
> stargazer(model2,
+           type = "html",
+           title="Table 2: Multiple Linear Regression Model I
Results",
+           summary = TRUE,
+           align=TRUE,
+           no.space=TRUE,
+           report=("vcstp"),
+           out = "C:/Users/maria/OneDrive/Documentos/3. Econometrics
1/BS2280/Coursework2/Model2.htm")
```

Output:

Multiple Regression Model I Results

	Dependent variable:
	Life_Expectancy
Alcohol	-0.142
	(0.118)
	t = -1.208
	p = 0.229
Schooling	2.221
	(0.185)
	t = 12.035
	p = 0.000
BMI	0.086
	(0.023)
	t = 3.833
	p = 0.0002
Constant	39.901
	(1.849)
	t = 21.578
	p = 0.000
Observations	171
R^2	0.671
Adjusted R^2	0.665
Residual Std. Error	5.001 (df = 167)
F Statistic	113.712*** (df = 3; 167)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
> # Obtaining ANOVA table
> anova_table2 <- anova(model2)
> print(anova_table2)

Analysis of Variance Table

Response: Life_Expectancy
           Df Sum Sq Mean Sq F value    Pr(>F)
Alcohol     1 1814.2  1814.2  72.543 9.158e-15 ***
Schooling   1 6349.6  6349.6 253.898 < 2.2e-16 ***
BMI         1  367.5   367.5  14.694 0.0001788 ***
Residuals 167 4176.4    25.0
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> # Creating a table from the ANOVA table
> xtable_anova2 <- xtable(anova_table2)
> # Saving it to a file
> print(xtable_anova2, type = "latex")

Output:
```

	Df	Sum Sq	Mean Sq	F value	Pr(F)
Alcohol	1	1814.21	1814.21	72.54	0.0000
Schooling	1	6349.63	6349.63	253.90	0.0000
BMI	1	367.48	367.48	14.69	0.0002
Residuals	167	4176.44	25.01		

```
> # Model 3.      Multiple linear regression model II
> # Including the quadratic term of the Schooling variable
> model3 <- lm(Life_Expectancy ~ Alcohol + Schooling + I(Schooling^2)
+ BMI, data = data_2011)
> summary(model3)

Call:
lm(formula = Life_Expectancy ~ Alcohol + Schooling + I(Schooling^2) +
    BMI, data = data_2011)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2971  -2.8558   0.3025   3.1938  13.9169

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    43.49275    5.32872   8.162 7.84e-14 ***
Alcohol        -0.15535    0.11914  -1.304 0.194084
Schooling       1.60598    0.87542   1.835 0.068366 .
I(Schooling^2)  0.02502    0.03481   0.719 0.473309
BMI             0.08757    0.02263   3.870 0.000156 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.008 on 166 degrees of freedom
Multiple R-squared:  0.6724,  Adjusted R-squared:  0.6645
F-statistic: 85.17 on 4 and 166 DF,  p-value: < 2.2e-16

> # Formatting the table of results
```

```
> stargazer(model3,
+           type = "html",
+           title="Table 3: Multiple Linear Regression Model II
Results",
+           summary = TRUE,
+           align=TRUE,
+           no.space=TRUE,
+           report=("vcstp"),
+           out = "C:/Users/maria/OneDrive/Documentos/3. Econometrics
1/BS2280/Coursework2/Model3.htm")
```

Output:

Multiple Regression Model II Results

	Dependent variable:
	Life_Expectancy
Alcohol	-0.155
	(0.119)
	t = -1.304
	p = 0.195
Schooling	1.606
	(0.875)
	t = 1.835
	p = 0.069
I(Schooling2)	0.025
	(0.035)
	t = 0.719
	p = 0.474
BMI	0.088
	(0.023)
	t = 3.870
	p = 0.0002
Constant	43.493
	(5.329)
	t = 8.162
	p = 0.000
Observations	171
R^2	0.672
Adjusted R^2	0.664
Residual Std. Error	5.008 (df = 166)
F Statistic	85.166^{***} (df = 4; 166)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
> # Model 4. Logarithmic Model
> # Making the log transformation
> data_2011$lnLife_Expectancy <- log(data_2011$Life_Expectancy)
> data_2011$lnSchooling <- log(data_2011$Schooling)
> # Building the new model
> model4 <- lm(lnLife_Expectancy ~ Alcohol + lnSchooling + BMI, data =
data_2011)
> summary(model4)
```

```
Call:
lm(formula = lnLife_Expectancy ~ Alcohol + lnSchooling + BMI,
    data = data_2011)

Residuals:
      Min        1Q    Median        3Q       Max
-0.225370 -0.037052  0.005733  0.043156  0.229383

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.3341212  0.0715694  46.586  < 2e-16 ***
Alcohol     -0.0012928  0.0017932  -0.721    0.472
lnSchooling  0.3444818  0.0322641  10.677  < 2e-16 ***
BMI          0.0014801  0.0003514   4.212 4.13e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07826 on 167 degrees of freedom
Multiple R-squared:  0.6306,  Adjusted R-squared:  0.624
F-statistic: 95.03 on 3 and 167 DF,  p-value: < 2.2e-16

> # Formatting the table of results
> stargazer(model4,
+           type = "html",
+           title="Table 4: Logarithmic Model Results",
+           summary = TRUE,
+           align=TRUE,
+           no.space=TRUE,
+           report=("vcstp"),
+           out = "C:/Users/maria/OneDrive/Documentos/3. Econometrics
1/BS2280/Coursework2/Model4.htm")

Output:
```

Logarithmic Model Results

	Dependent variable:
	lnLife_Expectancy
Alcohol	-0.001
	(0.002)
	t = -0.721
	p = 0.472
lnSchooling	0.344
	(0.032)
	t = 10.677
	p = 0.000
BMI	0.001
	(0.0004)
	t = 4.212
	p = 0.00005
Constant	3.334
	(0.072)
	t = 46.586
	p = 0.000

Observations	171
R^2	0.631
Adjusted R^2	0.624
Residual Std. Error	0.078 (df = 167)
F Statistic	95.029*** (df = 3; 167)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
> # Comparing all models
> stargazer(model1, model2, model3, model4,
+           type = "html",
+           title="Table 5: Comparison of Models Results",
+           summary = TRUE,
+           align=TRUE,
+           no.space=TRUE,
+           report=("vcstp"),
+           out = "C:/Users/maria/OneDrive/Documentos/3. Econometrics
1/BS2280/Coursework2/All_Models.htm")
```

Output:

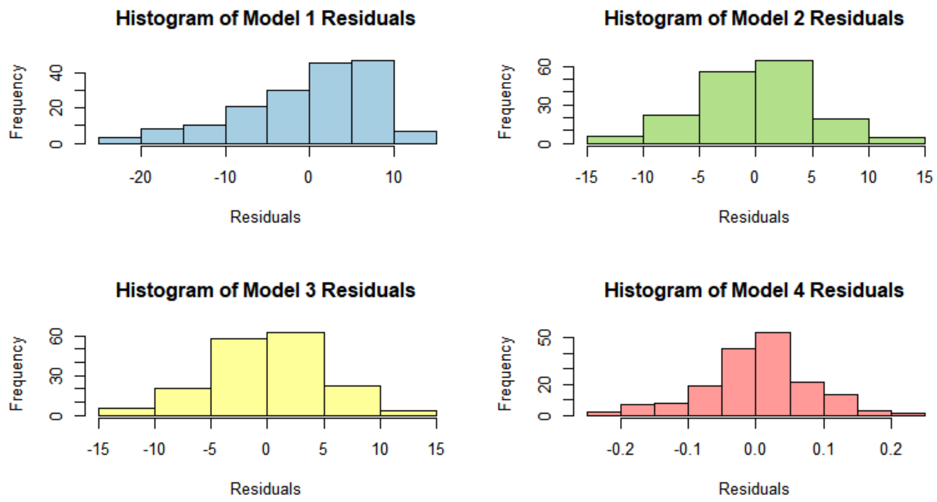
Comparison of Models Results

	Dependent variable:			
	Life_Expectancy	Life_Expectancy		lnLife_Expectancy
	(1)	(2)	(3)	(4)
Alcohol	0.837	-0.142	-0.155	-0.001
	(0.158)	(0.118)	(0.119)	(0.002)
	t = 5.305	t = -1.208	t = -1.304	t = -0.721
	p = 0.00000	p = 0.229	p = 0.195	p = 0.472
Schooling		2.221	1.606	
		(0.185)	(0.875)	
		t = 12.035	t = 1.835	
		p = 0.000	p = 0.069	
I(Schooling2)			0.025	
			(0.035)	
			t = 0.719	
			p = 0.474	
lnSchooling				0.344
				(0.032)
				t = 10.677
				p = 0.000
BMI		0.086	0.088	0.001
		(0.023)	(0.023)	(0.0004)
		t = 3.833	t = 3.870	t = 4.212
		p = 0.0002	p = 0.0002	p = 0.00005
Constant	66.858	39.901	43.493	3.334
	(0.976)	(1.849)	(5.329)	(0.072)
	t = 68.528	t = 21.578	t = 8.162	t = 46.586
	p = 0.000	p = 0.000	p = 0.000	p = 0.000
Observations	171	171	171	171

R^2	0.143	0.671	0.672	0.631
Adjusted R^2	0.138	0.665	0.664	0.624
Residual Std. Error	8.029 (df = 169)	5.001 (df = 167)	5.008 (df = 166)	0.078 (df = 167)
F Statistic	28.145^{***} (df = 1; 169)	113.712^{***} (df = 3; 167)	85.166^{***} (df = 4; 166)	95.029^{***} (df = 3; 167)
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

```
> # Predicting life expectancy data with each model
> data_2011$prediction1 <- predict(model1,
newdata=data.frame(Alcohol=data_2011$Alcohol))
> data_2011$prediction2 <- predict(model2, newdata=data.frame(Alcohol
= data_2011$Alcohol,
+                                                              Schooling
= data_2011$Schooling,
+                                                              BMI =
data_2011$BMI))
> data_2011$prediction3 <- predict(model3, newdata=data.frame(Alcohol
= data_2011$Alcohol,
+                                                              Schooling
= data_2011$Schooling,
+                                                              Schooling
_squared = data_2011$Schooling^2,
+                                                              BMI =
data_2011$BMI))
> data_2011$prediction4 <- predict(model4, newdata=data.frame(Alcohol
= data_2011$Alcohol,
+                                                              lnSchooli
ng = data_2011$lnSchooling,
+                                                              BMI =
data_2011$BMI))
> # Calculating the residuals
> data_2011$residuals1 <- data_2011$Life_Expectancy-
data_2011$prediction1
> data_2011$residuals2 <- data_2011$Life_Expectancy-
data_2011$prediction2
> data_2011$residuals3 <- data_2011$Life_Expectancy-
data_2011$prediction3
> data_2011$residuals4 <- data_2011$lnLife_Expectancy-
data_2011$prediction4
> # Plotting the histograms of residuals
> par(mfrow = c(2, 2))
> hist(data_2011$residuals1, main = "Histogram of Model 1 Residuals",
+     xlab = "Residuals", col = "#A6CEE3")
> hist(data_2011$residuals2, main = "Histogram of Model 2 Residuals",
+     xlab = "Residuals", col = "#B2DF8A")
> hist(data_2011$residuals3, main = "Histogram of Model 3 Residuals",
+     xlab = "Residuals", col = "#FFFF99")
> hist(data_2011$residuals4, main = "Histogram of Model 4 Residuals",
+     xlab = "Residuals", col = "#FF9A98")
> par(mfrow = c(1, 1))
```

Output:




```
> # Finding the t-critical value
> alpha <- 0.05
> df <- 169
> t_critical <- qt(1 - alpha/2, df)
> print(t_critical)
[1] 1.9741
> # Finding f-critical value
> df1 <- 2
> df2 <- 167
> significance_level <- 0.05
> f_critical <- qf(1 - significance_level, df1, df2)
> print(f_critical)
[1] 3.05012
```