DJ Holmer, Zoe Absalonson
MATH 342
Dr. Amy Anderson
Final Project

Abstract

Sleep is a fundamental aspect of human life, critical for maintaining health and functionality. This study investigates the factors influencing sleep efficiency, defined as the proportion of time spent in bed actually sleeping. Using data from the Sleep Efficiency Dataset by Melike Dilekci, which includes responses from 386 individuals, we examined the impact of seven predictors: percentage of time in REM sleep, percentage of time in deep sleep, caffeine consumption, alcohol consumption, physical activity, age, and number of awakenings. Our regression analysis of the model we ran, $Y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \beta_3 x_{i_3} + \beta_4 x_{i_4} + \beta_5 x_{i_5} + \beta_6 x_{i_6} + \beta_7 x_{i_7}$, with x_{i_k} representing different predictor variables mentioned above, revealed an R-squared value of 0.7786, indicating that 77.86% of the variability in sleep efficiency is explained by these predictors. The ANOVA F-test yielded a p-value less than 0.05, demonstrating that the model is statistically significant and that at least one predictor has a non-zero coefficient. The residual standard error we found to be 0.06425 also suggests a good fit of the model. Our findings show that most predictors, except for caffeine consumption, are significant in predicting one's sleep efficiency. This research provides valuable insights into how various lifestyle and biological factors contribute to sleep quality, offering potential pathways for improving overall sleep health.

Introduction

Along with food and water, sleep is one of the most important elements of human life. With sleep being so vital to human functionality, the scientific question that we are trying to answer is if certain factors can help predict a person's "sleep efficiency" (the proportion of the time they spend in bed actually sleeping). The factors we analyzed as predictors for sleep efficiency in individual people were the percentage of REM and deep sleep during the night, caffeine and alcohol consumption, physical activity level, age, and number of times the individual woke up during the night. By choosing to look further into sleep habits, this study may provide insight as to how a person optimizes their overall quality of sleep.

Data

We collected our data from a public dataset, Sleep Efficiency Dataset (Dilekci, 2023), from Melike Dilekci. This dataset was compiled via a survey answered by 452 people (a person being an

"individual" for our purposes). From this, we looked only into the individuals who fully answered every survey prompt (some people opted out), which came out to be 386 total individuals, so $n = 386$. This was simply due to the fact that we needed full data for each individual. Our dataset consisted of seven predictor (independent) variables and one response (dependent).

The response variable was sleep efficiency (y), which was measured by the proportion of time in bed spent asleep. The predictor variables were percentage (%) of total sleep time spent in REM sleep (x_1), percentage (%) of total sleep time spent in deep sleep (x_2), caffeine consumed (milligrams) in the 24 hours prior to bedtime (x_3), alcohol consumed (ounces) in the 24 hours prior to bedtime (x_4), level of physical activity (x_5), measured by the number of times the individual exercised per week (1-5), age in years of the individual (x_6), and the number of times the individual woke up during the night (x_7).

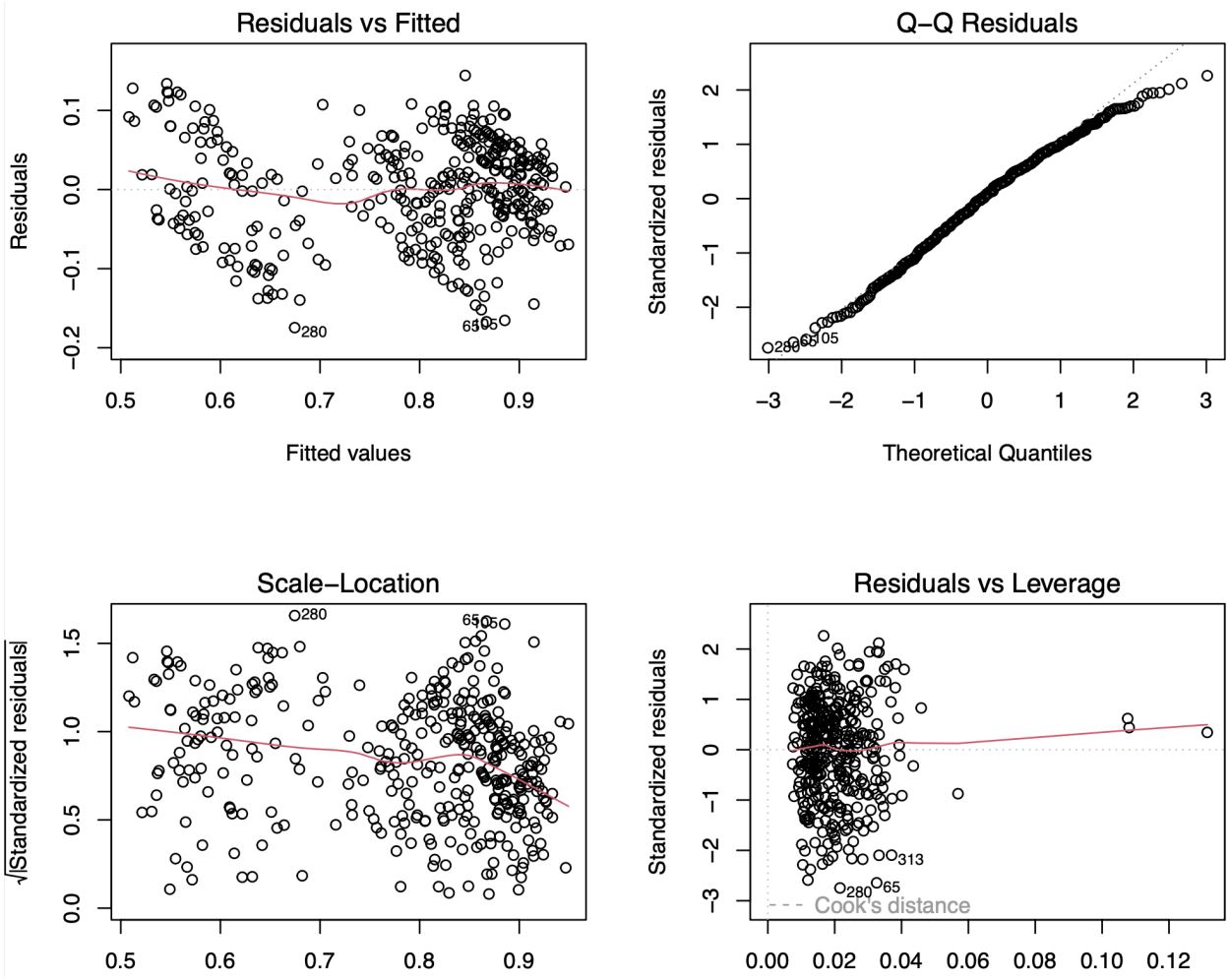
Model/methods

We used a standard multiple regression model to fit a "line" of the form:

$$Y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \beta_3 x_{i_3} + \beta_4 x_{i_4} + \beta_5 x_{i_5} + \beta_6 x_{i_6} + \beta_7 x_{i_7},$$

where Y_i is the predicted sleep efficiency for for individual number i (i = 1, 2, …, 386) and x_{i_j} is the value of the predictor variable x_k (k = 1, 2, …, 7) for individual i (i = 1, 2, …, 386).

Diagnostics

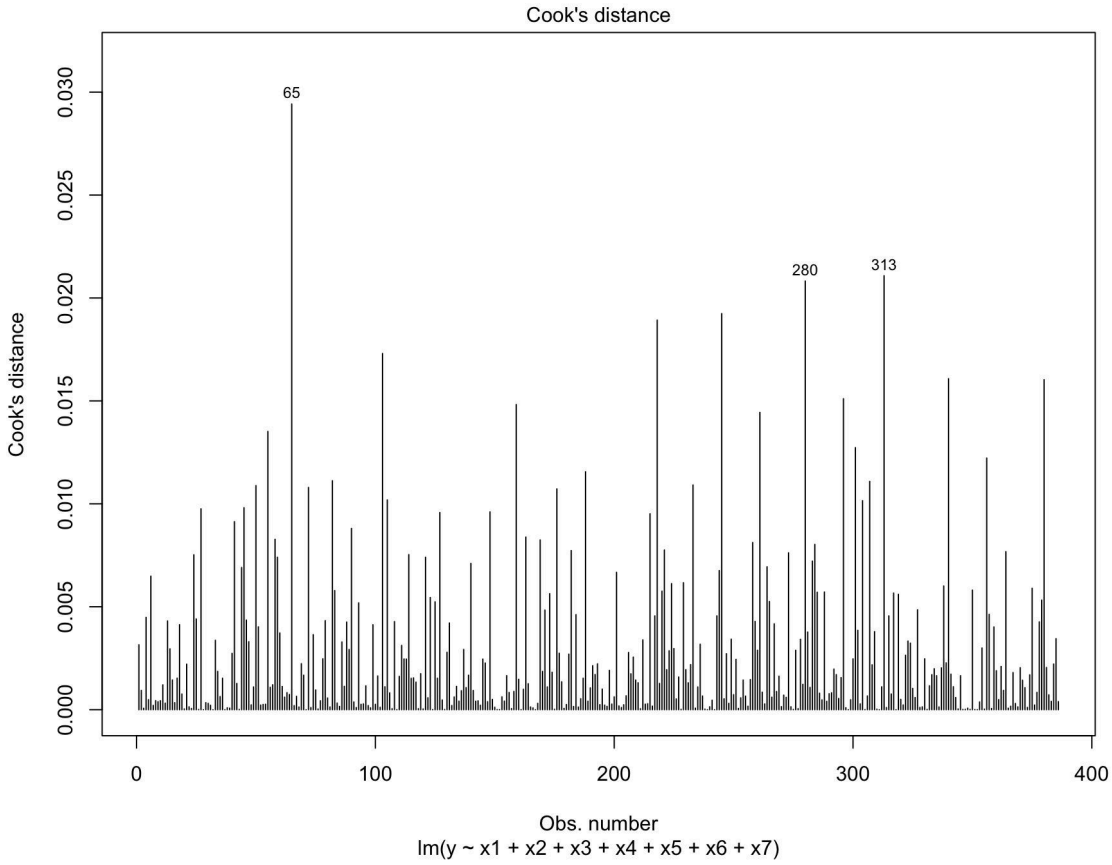


In the context of the fit of our model, as conveyed by our Residuals vs. Fitted plot, there does not seem to be an issue with linearity as there is a clear mean about zero. The Scale-Location plot also reassures us of homoscedasticity as there appears to be a general slope of zero across the plot with no difference in variance of residuals for different fitted values. Focusing in on error term distribution, though there is a slight tail on the ends of the plot, the Q-Q plot indicates that the residuals came from a normal distribution with a mean of 0 and a variance of 1.

We used the Residuals vs. Leverage plot in our analysis of outliers/influential points. Given that $k = 7$ (the number of predictor variables) and $n = 386$ (sample size), the cutoff leverage value for an influential individual was $\frac{2(k+1)}{n} = \frac{2(7+1)}{386} = \frac{16}{386} \approx 0.0415$. Using R, we identified the points with leverage values above this cutoff. Those individuals were 69 (leverage value = 0.1314548), 83 (0.1076442), 99 (0.04588334), 137 (0.1080849), 220 (0.05689806), and 326 (0.0435159). We are not worried about these individuals though, as they have Cook's distances less than 0.5, a common cutoff for

Cook's distance, meaning they have the power to influence the "best-fit line" but happen to approximately fall on the same "line" as the rest of the points in our regression. Here is what we observed about the influential individuals that we believe were the main factors affecting their leverage value: Individuals 69, 83, and 137 had higher caffeine intake compared to others (200mg), individual 99 had more awakenings than most others (4 awakenings), individual 326 had higher alcohol intake than the rest (5oz), and individual 220 had low percentages of REM sleep (15%) and deep sleep (22%) compared to others.

The diagnostic plots label individuals 65, 280, and 313 as outliers. After closer inspection, we figured this was due to the fact that those individuals have the highest Cook's distance values compared to the others (still lower than 0.5 though). The Cook's distance diagnostic plot is shown below to highlight this. We also looked closer at the specific data in these individuals (x_1, x_2, etc.) and noticed that they should have gotten better sleep (higher sleep efficiency) based on the other predictor values in comparison to other individuals with similar predictor values that we observed in the study.



Results

ANOVA Table

Source	Degrees of Freedom (d.f.)	Sum of Squares	Mean Square	Test Statistic (F)	P-value
Regression	7	SSR = 5.4866	MSR = 0.7838	$\frac{MSR}{MSE} = 189.8599$	1-Fcdf(0, 189.8599, 7, 378) ≈ 0
Residual Error	378	SSE = 1.5605	MSE = 0.0041		
Total	385	SST = 7.0471			

Our resulting regression output gave us an r^2 value of 0.7786, the proportion of variability in sleep efficiency explained by the regression. Looking at the ANOVA table, the F-test gives us a p-value of approximately 0, which is less than a significance level of 0.05 so there is sufficient evidence to conclude that one or more of our predictors have non-zero coefficients, so our model is statistically significant. Lastly, the remaining standard deviation of the error terms was 0.06425, which is relatively small in proportion to the scale of our model, indicating a better fit of the model to the data.

Table of Coefficients

	Estimate (β_i)	Standard Error	Test statistic (T)	P-value
Intercept	0.3721699	0.0310651	11.98	$< 2 \times 10^{-16}$
x_1	0.0049637	0.0008826	5.624	3.63×10^{-8}
x_2	0.0057975	0.0002507	23.121	$< 2 \times 10^{-16}$
x_3	0.0001638	0.0001186	1.381	0.168135
x_4	-0.0067295	0.0022683	-2.961	0.003202
x_5	0.0071110	0.0023567	3.017	0.002722
x_6	0.0009338	0.0002511	3.718	0.000231
x_7	-0.0308622	0.0026410	-11.686	$< 2 \times 10^{-16}$

In our table of coefficients, all of the p-values, except for x_3 (caffeine intake), are less than 0.05, meaning we have sufficient evidence to conclude that the coefficients for the predictors x_1, x_2, x_4, x_5, x_6, and x_7 are statistically significant with nonzero coefficients, and x_3 is insignificant in our model since we do not have sufficient evidence to conclude that its coefficient is nonzero.

In more closely analyzing the relationships between the response and predictor variables, we found the following:

- Sleep efficiency and percentage of REM sleep - As the percentage of REM sleep increases, sleep efficiency increases, so greater values of x_1 correspond to greater values of y.
    - Coefficient (β_1): 0.0049637 (> 0)
- Sleep efficiency and percentage of deep sleep - As the percentage of deep sleep increases, sleep efficiency increases, so greater values of x_2 correspond to greater values of y.
    - Coefficient (β_2): 0.0057975 (> 0)
- Sleep efficiency and alcohol consumption - As an individual increases their alcohol consumption, their sleep efficiency decreases, so greater values of x_4 correspond to lower values of y.
    - Coefficient (β_4): -0.0067295 (< 0)
- Sleep efficiency and physical activity - As physical activity increases, sleep efficiency increases, so greater values of x_5 correspond to greater values of y.
    - Coefficient (β_5): 0.0071110 (> 0)
- Sleep efficiency and age - As an individual's age increases, their sleep efficiency increases, so greater values of x_6 correspond to greater values of y.
    - Coefficient (β_6): 0.0009338 (> 0)
- Sleep efficiency and awakenings - The more times an individual wakes up during the night, the lower their sleep efficiency is, so greater values of x_7 correspond to lower values of y.
    - Coefficient (β_7): -0.0308622 (< 0)

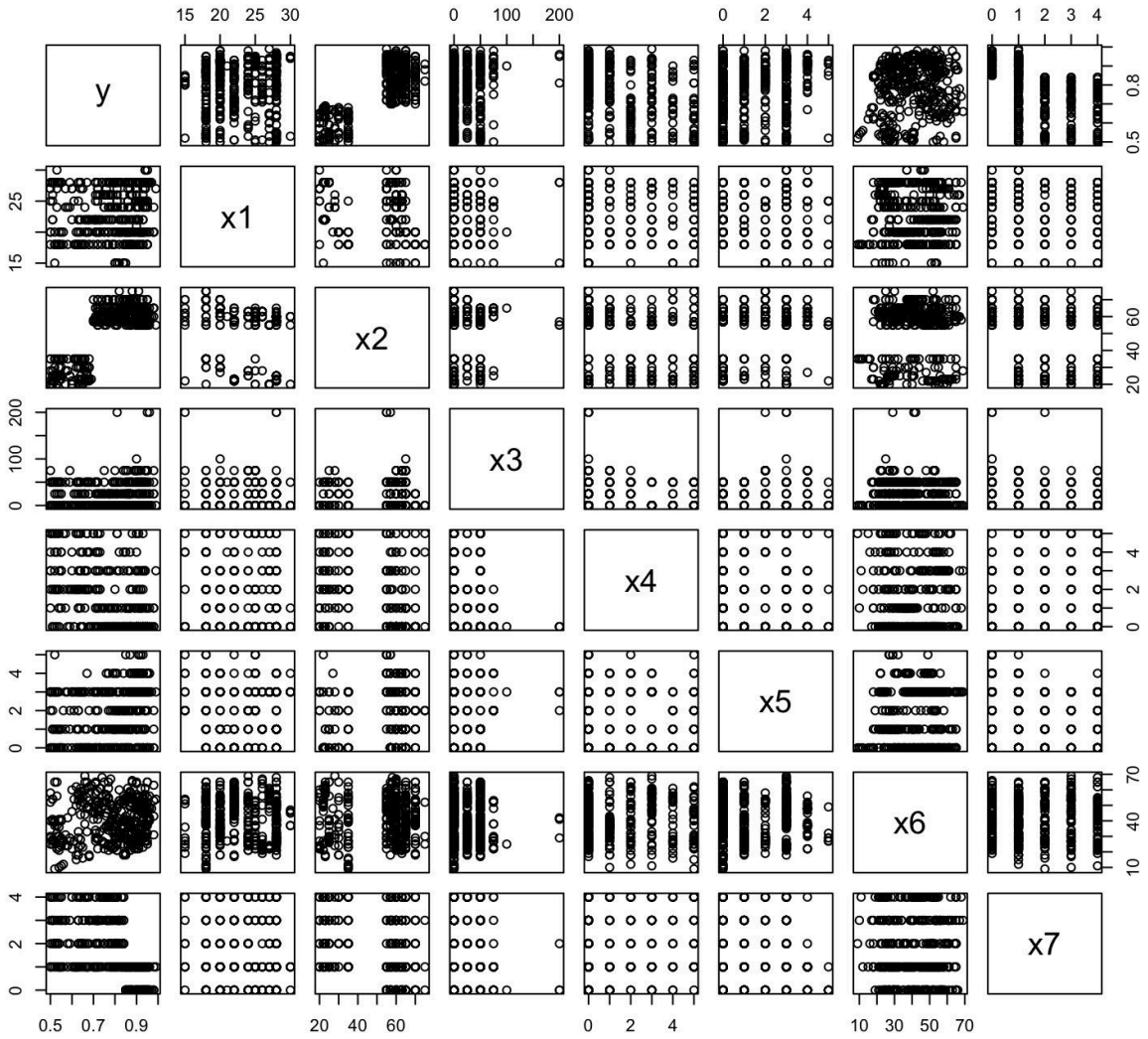
Conclusion

This study aimed to identify key factors influencing sleep efficiency, the proportion of time spent asleep while in bed, through the analysis of a standard multiple regression model, fitting a "line" of the form: $Y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + \beta_3 x_{i_3} + \beta_4 x_{i_4} + \beta_5 x_{i_5} + \beta_6 x_{i_6} + \beta_7 x_{i_7}$. Our analysis included seven

predictors: the percentage of time in REM sleep, the percentage of time in deep sleep, caffeine consumption, alcohol consumption, physical activity, age, and the number of awakenings. The regression model showed that 77.86% of the variability in sleep efficiency could be explained by these predictors, indicating a strong relationship between the predictors and the response variable, which was verified in the diagnostic plots, having no issues with linearity, heteroscedasticity, distribution of error terms, or influential points (substantially affecting the regression). The ANOVA F-test, having a p-value less than the set significance level of 0.05, revealed the predictors were significant in predicting sleep efficiency. Individual hypothesis testing on the β_i coefficients uncovered that x_3, the predictor variable for caffeine consumption, was insignificant, with a p-value greater than 0.05. Notably, minimizing alcohol consumption and increasing physical activity during the day were identified as effective ways to improve sleep quality. It must be highlighted that our study includes limitations, such as potential biases in survey responses and the absence of information on underlying sleep conditions. For instance, the lack of a significant relationship between caffeine consumption and sleep efficiency might be explained by individual differences in response to caffeine, such as those seen in people with ADHD, who might experience an opposite effect. Overall, our findings provide valuable insights into the determinants of sleep quality and suggest potential solutions to improve sleep health.

Supplementary Items

Matrix Plot



Checklist

Abstract

- ☐ Does your project have an abstract?
- ☐ Does your abstract make sense as a stand-alone document that is a one-paragraph version of your paper for a person that will never see the main paper?
- ☐ Does your paper make sense as a stand-alone document for a person who never read the abstract?

Introduction

- ☐ Does your project have an introduction?
- ☐ Does the introduction introduce the topic of your paper? Your paper needs to be about the scientific question you are addressing, not about the statistical methodology.
- ☐ Does the introduction motivate the reader to care about your topic?
- ☐ Does the introduction contain any necessary background information a reader would need to understand the context or motivation of your paper?

Data

- ☐ Does your project have a section that introduces your dataset?
- ☐ Is there a reference for where you obtained your data?
- ☐ Does your paper clearly indicate what an "individual" is in your dataset? Your variables are something that is measured for each "individual" in the dataset.
- ☐ Does your paper go over every variable (including both the predictor variables and the response variable) you use, telling what the variable measures and giving its units?
- ☐ Do you define every variable or abbreviation used in your paper? For example, it is a problem to say your predictor variables are height and weight, but then, suddenly x1 and x2 show up somewhere. If you are going to call height x1, you need to say that.

Model/Methods

- ☐ Does your paper state the sample size (number of individuals in the dataset)? • Model/methods
- ☐ Is your model stated explicitly? I want to see an equation that contains coefficients for your variables.
- ☐ When you look at your paper wholistically, is it about the scientific questions you are asking of the data (which is good), or mostly about your journey in coming up with a model (which is not good—nobody cares about all 20 different transformations you attempted).
- ☐ If you mentioned using a scatterplot of two of the variables to choose your model, did you include a figure showing that plot? This could happen, for example, if you used the matrix plot to decide to add a higher power of one of your predictors, or if you were thinking about multicollinearity.
- ☐ If you used variance inflation factors to decide to drop some variables does the paper give the variance inflation factors? This could be in a paragraph or in a table. Note: Most student papers don't end up going in this direction.

If you did a model selection procedure (usually a bad idea)

- ☐ Did you explain exactly what you did? Would a reader be able to fully reproduce your analysis based on the description in your paper?
- ☐ What proportion of the data was used to select the model? What proportion was used in the final analysis?
- ☐ What variables were included in the maximal model?
- ☐ Did you use forward, backward, or forward-backward stepwise? If you used forward-backward stepwise, did you start with the maximal model or the minimal model?
- ☐ What measure (e.g. AIC) did you use to decide on what model was best?
- ☐ Are the table of coefficients and other results that you present based on running your model on data OTHER than the data used to choose the model? You will not get credit for a paper where you use the same data to choose the model and use that model to draw conclusions about the coefficients in the model.

Diagnostics

- ☐ Does your paper include the set of four standard diagnostic plots? – Does your paper refer to your diagnostic plots?
- ☐ You will not get credit for a paper that doesn't show a nice residual plot unless you have talked to me about it.
- ☐ Does your paper mention any outliers that are numbered in the diagnostic plots?
- ☐ Does your paper mention any influential points? If the influential points have low values for Cook's distance, it is enough to mention the points (and explain why you aren't worried about them). If you have points with high values for Cook's distance using your final model, you should show results both with and without those points in the results section.
- ☐ For both outliers and influential points: Does the paper get into these points? What about them makes them weird? You need to dive into the data for this and really be able to tell the reader what is happening.
- ☐ For both outliers and influential points: Does your paper generally refer to these as "points" as if you are completely out of touch with the meaning of the data? That's bad. Remember, these "points" are "people" or "countries" or something in the real world. Saying "point 45 is influential" doesn't mean much, but saying "Paraguay is was influential in this regression because its mean level of education was much lower than that of other countries of a similar size" is much more interesting.

Results

- [ ] First things: Does the paper give any indication as to how well the model explains tha variability in the response variable? The paper should address this in a few ways. Some ideas: What was r2? What does the ANOVA F-test say about whether there is evidence that any of your predictors have non-zero coefficients? After your regression is done, what is the remaining standard deviation of the error terms?

ANOVA table

- [ ] Does the paper include the ANOVA table? Make sure it is the ANOVA table that has one line for the regression, one line for the residuals, and, optionally, one line for the totals. R, unlike most programs, doesn't like to give you the overall ANOVA table but you can recreate it from the sequential sum of squares table.
- [ ] Did you double-check that you included the ANOVA table and not the sequential sum of squares table. Again: The ANOVA table has just one line for the regression, which should have a degree of freedom equal to the total number of predictor variables. If you are looking at a table with a line for each variable where the df for each of these variables is 1, then that is the sequential sum of squares table, not the ANOVA table. * Does the paper refer to the ANOVA table?

Table of Coefficients

- [ ] Does your paper include a table of coefficients?
- [ ] Does your paper refer to the table of coefficients?
- [ ] Did you put your table of coefficients into a nice form (no raw R output, no slangy notation (e.g. for the p-value column, the heading R gives is not formal notation).
- [ ] Does your paper go through what the table says about which predictor variables are significant?
- [ ] For the significant variables, does the paper go through what the table of coefficients says about the direction of the relationship between that variable and the predictor? For example, is it bigger or smaller values of x1 that correspond to higher values of Y , on average? (Note: if your model has interaction terms or higher powers of the predictor, there might be some work that needs to be done to figure this part out.)

Supplementary items

- ☐ Matrix plot. Did you include a matrix plot that shows a scatterplot of each variable vs the other variables? Put this in as a separate page at the end of the paper.
- ☐ A printout of this checklist with checkmarks for each question.

Miscellaneous

- ☐ ∗ If the word “significant” appears in the paper, is it accompanied with a description of the test that was run and a p-value that is less than 0.05? “Significant” is a technical term.
- ☐ ∗ If the word “correlated” or “correlation” appears in the paper, does it refer to the relationship between exactly two variables? Was the correlation coefficient between those variables actually computed?