Assignment Cover Sheet

PUBH726 : Applied Biostatistics 2: Regression Methods

Student identification number: 7270243
Date: 3/8/2025

Declaration
I have read and understood the University of Otago's policies on academic integrity, academic misconduct and plagiarism information, as outlined on HSMoodle.
In submitting this assignment I declare that:

- This assignment is entirely my own work, all sources have been referenced appropriately and I have not previously submitted this work or a version of it for assessment in any other paper or programme.
- I have not been party to someone else's plagiarism by collusion, allowing them to copy my work, or helping them plagiarise.

I understand that if I am involved in any form of dishonest practice I will be subject to disciplinary procedures within the University of Otago.

Date due: 3/8/2025

Signature: Johanna Escudero Pino

PUBH 726: Applied Biostatistics 2 – Regression Methods

ASSIGNMENT 2 2025

The assignment has a total of 100 marks, and it contributes 40% of your grade for PUBH 726 Applied Biostatistics 2 – Regression Methods.

In a study on stress among adolescents, the stress levels of a sample of youth were assessed. The researchers are primarily interested in factors that may influence stress levels among adolescents. Among many potential factors that may be associated with stress, age, sex, body weight, amount of exercise, and duration of sleep of these young people have been suggested as being important. The dataset is available on Moodle in an excel file named *stress_2025*. The variables in the dataset are entered in the order described below.

ID	unique subject identifier;
Sex	coded as 0 = female; 1 = male
Age	recorded as years to 1 decimal place
BMI	Body Mass Index (recorded as a continuous index)
Extime	Exercise time recorded as number of hours of exercise per week
Sleep	Duration of sleep on an average weeknight coded as
	0= less than 8 hours, 1 = 8 hours or more
Stress	Stress level measured using a standardised scale and recorded as stress score, higher score indicates higher stress level

Question 1 [75 marks]: The primary research question for this study is to determine the association between exercise time and the outcome stress.

a. [15 marks] Check the data for errors and missingness. Describe the checks you did for each variable and justify your approach to dealing with any identified problems. Include any appropriate graphs or summary statistics (maximum of 1 page including any graphs).

Specific checks on the data were as follows:

Missing Values: No missing values in the data were found.

Duplicate IDs: No duplicate participants were found, each participant had a unique ID

Checked for Encoding Errors for Categorical Variables

Sex: All values were correctly coded as 0 (female) or 1 (male) with no errors identified

Sleep: All values were correctly coded as 0 (<8 hours) or 1 (≥8 hours) with no errors identified.

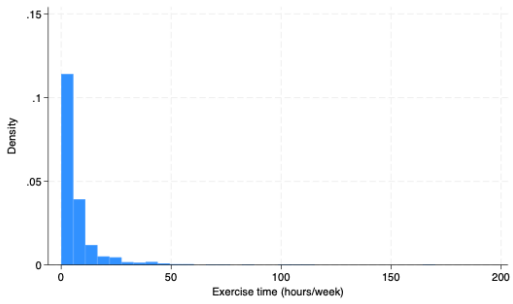
For continuous variables, checks were made to assess whether the distribution was plausible and to identify any extreme or implausible values. ***A histogram was used for exercise time, as further analysis was required due to having a large range (0 – 170 hours/week of exercise and a large standard deviation indicating data was spread out.***

Age: Ranged from 11.7 – 18 years mean=14.6, SD=1.0). The study was for adolescence aged 11 – 18 so values were as expected.

BMI: Ranged from 9.4 – 41.4 (mean=19.1, SD=3.4), values fell within expected ranges for adolescence with no implausible values.

Exercise time hours/week: Ranged from 0 – 170 (mean=6.9, SD=11.1). It had a highly right-skewed distribution as shown by Figure 1, with most participants reporting minimal exercise. There was one implausible value for a participant that reported 170 hours/week, equivalent to over 24 hours/day), which will be inflating the mean.

Figure 1 - Distribution of Exercise Time (hours/week)



Stress Score: Ranged from 0 - 33 (mean=17.4, SD=4.3), range seems appropriate and there did not appear to be implausible values.

Decisions made on data quality were as follows:

Only one extreme value was removed where exercise time exceeded the maximum hours in a week. Additional outliers were not removed as it is not possible to say that they are errors or implausible values as they may represent legitimate extreme values. The final dataset contained 1,332 records after removing the participant that reported 170 hours/week of exercise.

b. [15 marks] Prepare a 'Table 1' that summarises the overall characteristics of the sample.

Table 1 - Characteristics of adolescents participating in stress study (n=1,332)

Characteristic	n	%
Sex		
Female	636	47.8
Male	696	52.2
Sleep duration		
Less than 8 hours	627	47.1
8 hours or more	705	52.9

	Median	Mean (SD)
Age (years)	14.6	14.6 (1.0)
BMI (kg/m²)	18.4	19.1 (3.4)
Exercise time (hours/week)	4	6.8 (10.2)
Stress score	17	17.4 (4.3)

c. [10 marks] Write a paragraph (word limit=300) to accompany your table that describes the overall characteristics of the sample.

The study initially included 1,333 adolescents but one participant was excluded due to reporting 170 hours of exercise per week, which exceeds the maximum hours in a week. This left 1,332 participants with complete data for all variables. The sample had a relatively even gender distribution, with 696 males (52.2%) and 636 females (47.8%). The mean age was 14.6 years (SD = 1.0), with a median age of 14.6 years. This suggests the sample was well-balanced around mid-adolescence (14 to 15 years old), with most children clustered around this age. Sleep duration had a relatively even distribution across participants, with 627 (47.1%) reporting less than 8 hours of sleep on an average weeknight and 705 (52.9%) reporting 8 hours or more. The mean BMI was 19.1 (SD = 3.4) with a median of 18.4, this shows that; BMI values appear relatively normal for this age group, though there is variation (data points are more spread out than for age - seen by a higher standard deviation); a small number of higher BMIs may be present, shown by the mean being larger than the median. Exercise time showed considerable variation, with a mean of 6.8 hours per week (SD = 10.2) and a median of 4 hours per week. The median being lower than the mean indicates that most participants had a relatively sedentary lifestyle with some participants exercising much more than others. The mean stress score was 17.4 (SD = 4.3) with a median of 17, indicating stress levels were similar across most participants although there could be a few high scores seen by a slightly larger mean compared to the median. Overall, these characteristics show a sample of

adolescents with varying exercise habits and stress levels, providing a good population for examining the association between exercise time and stress.

d. [25 marks] Fit a regression model to estimate the effect of exercise time on stress. Consider all other relevant variables, however, consideration of potential interactions is NOT required. Check and report on model diagnostics and potential outliers and any decisions these have informed regarding your final model and analysis. Present a step-by-step outline of the process you have undertaken to identify your final model. Show the results of your final model in an appropriately formatted table.

Table 2 - Association between exercise time and stress levels in adolescents (Males and Females between the ages of 11 and 18 participating in a stress study)

Variable	Coefficient	95% CI	p
Exercise time (hours/week)	0.004	(-0.018, 0.027)	0.709
Age (years)	0.614	(0.376, 0.852)	<0.001
Sex (Male)	-0.768	(-1.227, -0.308)	0.001
Sleep duration (≥8 hours)	-1.348	(-1.815, -0.881)	<0.001
[Constant]	9.5	-	-

N = 1,332; R^2 = 0.061

Outline of Process:

Step 1 Data preparation: Data checks showed that there were 1,332 adolescents with complete data. One participant was excluded due to reporting 170 hours of exercise per week, which was implausible.

Step 2: Linearity The below Scatter plots of stress versus each continuous predictor showed reasonable linear relationships suitable for regression analysis

Figure 2 - Stress Score Versus Age

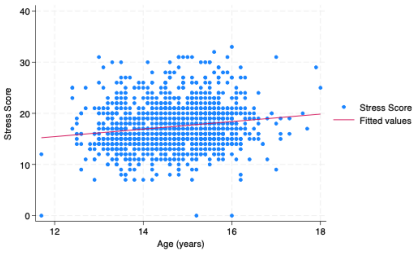


Figure 3 - Stress Score Versus BMI

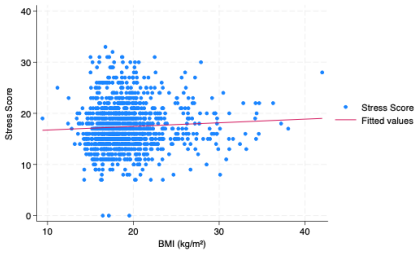
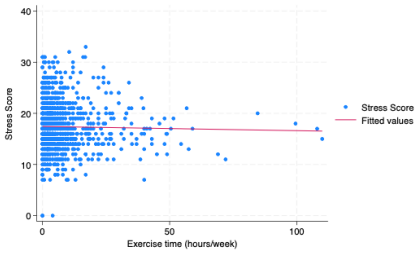


Figure 4 - Stress Score Versus Exercise Time



Step 3: Full model and multicollinearity The full model included exercise time, age, sex, BMI, and sleep duration. All VIF values were less than 1.1, so no multicollinearity concerns.

Step 4: Backwards elimination:

- BMI was removed (p=0.218, highest p-value)
- There was a less than 10% change in the coefficient for exercise time after remove BMI, confirming BMI is not a confounder
    - β0 = coefficient of exercise time with BMI in model = 0.00393329
    - β1 = coefficient of exercise time without BMI in model = 0.00427109
    - Percent change = ((β1 – β0) / β1) × 100 = 8.59%
- Exercise time was kept in the model as it was the exposure variable, despite having the highest p-value (p=0.709)

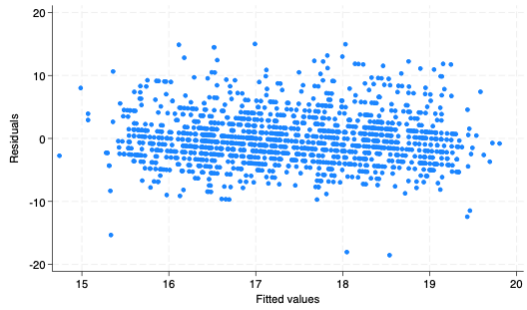
Step 5: Model diagnostics The following regression assumptions were checked:

Linearity: Satisfied based on Figure, 2, 3 and 4 showing linear relationships.

Independence: Satisfied based on the assumption that observations in this study are independent - stress levels reported by one adolescent are not influenced by those of another. IDs are all unique it is assumed that each participant is only represented once in the study.

Homoscedasticity: Satisfied as shown by Figure 5; residuals versus fitted values have an even distribution around zero with no fanning pattern, satisfying the constant variance assumption.

Figure 5 – Scatterplot (Homoscedasticity)



Normality: Satisfied as Histogram and Q-Q plot of residuals showed approximately normal distribution with only minor deviations at extremes. Figure 6 shows normal distribution centred around zero, supporting the normality assumption. Figure 7 shows that Residuals follow the diagonal line closely with minor deviations at extremes, showing that normality assumption is satisfied.

Figure 6 – Histogram (Normality)

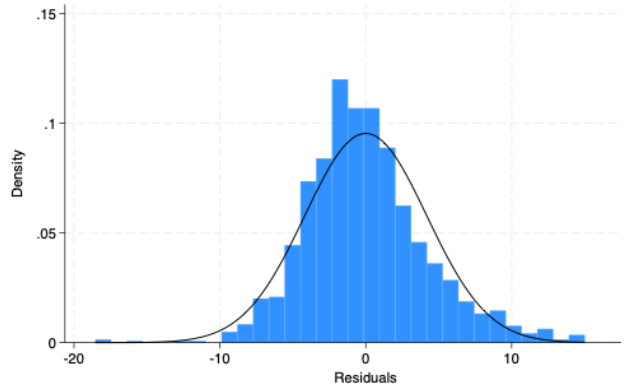
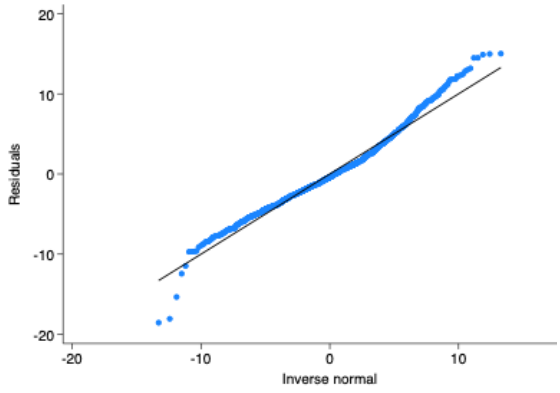
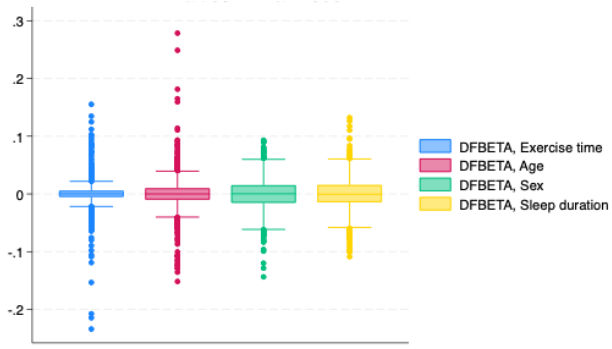


Figure 7 - Q-Q Plot (Normality)



Step 6: Outlier assessment DFBETA analysis showed most values clustered around zero with some outliers, but none appeared extreme and far from everything else so no sensitivity analysis was required. This is demonstrated by Figure 8.

Figure 8 - DFBETA Values Outliers in Final Model



Final model: Stress = Exercise time + Age + Sex + Sleep duration

e. [8 marks] Interpret your final results in a paragraph (word limit=300).

Exercise time in adolescents between 11 and 18 years old was not significantly associated with stress levels (p=0.709). For every one-hour increase in weekly exercise time, stress scores increased by 0.004 points on average, but the 95% confidence interval (-0.018, 0.027) includes zero, indicating no evidence of an association in the population. However, several other factors were significantly associated with stress. Age showed a strong positive association (p<0.001), with stress scores increasing by 0.614 points for every one-year increase in age, and there is 95% confidence that the true increase lies between 0.376 and 0.852 points. Males had significantly lower stress levels than females (p=0.001), with male participants scoring 0.768 points lower on average (95% CI: -1.227, -0.308). Sleep duration was also significantly associated with stress (p<0.001), with adolescents getting 8 or more hours of sleep per night having stress scores 1.348 points lower than those getting less than 8 hours (95% CI: -1.815, -0.881). The model explained 6.1% of the variation in stress levels. All regression assumptions were met, and no influential outliers were identified. Exercise time showed no significant association with stress in this sample, age, sex, and sleep duration were important predictors of stress levels among adolescents.

f. [2 marks] Provide a lay conclusion from the results in no more than two sentences.

In this study of 1,333 adolescents (with one excluded for implausible exercise time), the amount of weekly exercise was not associated with stress levels. However, older adolescents, females, and those getting less than 8 hours of sleep per night experienced significantly higher stress levels.

Question 2 [20 marks]. As it is common for 8 hours of sleep to be recommended, a secondary research question the researchers are interested in is whether there is any association between sleep duration as a categorical variable and the outcome stress.

a. [10 marks] Perform a simple linear regression to estimate the effect of sleep on stress. Present the results in an appropriately formatted table and interpret the output in an accompanying paragraph (Maximum half a page).

Table 3 - Simple linear regression analysis of sleep duration and stress levels in adolescents (Males and Females between the ages of 11 and 18 participating in a stress study)

Variable	Coefficient	95% CI	p
Sleep duration (≥8 hours)	-1.670	(-2.127, -1.214)	<0.001
[Constant]	18.263	(17.931, 18.595)	

N = 1,332; $R^2 = 0.037$

Sleep duration is significantly associated with stress levels in adolescents ($p<0.001$). Adolescents who get 8 or more hours of sleep per night have stress scores that are 1.670 points lower on average compared to those who get less than 8 hours of sleep. We are 95% confident that the true difference in stress levels between the two sleep groups lies between 1.214 and 2.127 points lower for those getting adequate sleep. The constant term of 18.263 represents the mean stress level for adolescents getting less than 8 hours of sleep while those getting adequate sleep have a mean stress level of 16.593 (18.263 - 1.670). This model explains 3.7% of the variation in stress levels. The significant negative association demonstrates that having a recommended sleep duration of 8 or more hours per night is associated with a meaningfully lower stress levels among adolescents. The extent of this difference is a reduction of 1.670 points on the stress scale which represents a meaningful reduction in stress levels, showing the importance of adequate sleep for adolescent wellbeing.

b. [5 marks] Assess whether there is an interaction between age and sleep. For this question, centre the age variable by generating a new variable that subtracts the mean age of the sample from each participant's age. You should restrict your analysis to include only these three variables (i.e., sleep, "centred" age, and the interaction term). Describe what you did to assess if there is an interaction and state your findings in two or three sentences.

To assess the interaction a centred age variable was created by subtracting the sample mean age (14.63 years) from each participant's age, then a fitted a regression model including sleep duration, centred age, and the interaction term was created.

Findings demonstrated that the interaction between sleep duration and age was not statistically significant ($p = 0.533$). This indicates that the association between sleep duration and stress levels does not differ significantly across different ages in this adolescent sample.

c. [5 marks] Write a brief paragraph (word limit=350) presenting your findings on the association between sleep and stress in your final model, providing a lay conclusion.

This study looked at how sleep affects stress levels in 1,333 adolescents aged 11 to 18. One participant was excluded from the analysis because they reported exercising for more hours than there are in a week. It was found that adolescents who usually get 8 or more hours of sleep on weeknights tend to feel significantly less stressed than those who get less sleep. On average, adolescents who met the recommended amount of sleep had stress scores that were lower than those that did not sleep enough. This pattern was clear even when we took other factors into account, like age, gender, and how much exercise they do each week. Even after adjusting for the other factors, the difference was still the same: adolescents who got enough sleep scored about 1.3 points lower on a stress scale, which shows that getting enough sleep appears to make a real difference to how stressed adolescents feel.

The effect of sleep on stress for younger versus older adolescents was also tested. Results demonstrated that the impact that sleep had on stress stayed the same for both younger and older adolescents - the benefit of sleep was the same no matter how old the adolescent was. This means that the positive impact of getting enough sleep applies to all adolescents in this age group.

Overall, the findings strongly show that sleep impacts stress: adolescents who regularly sleep 8 or more hours per night are likely to feel less stressed, and this is true whether they’re 11 or 18 years old. These results highlight the importance of promoting healthy sleep habits to help improve adolescent’s stress and mental health.

Question 3 [5 marks]. Mini Viva: During Week 4, after you have submitted your responses to Questions 1 and 2 of the assignment, we will arrange a time to meet with you for 10 minutes for this part of the assignment. You will be asked a small number of questions based on the following two regression outputs:

```
. regress stress i.sleep age
```

Source	SS	df	MS	Number of obs	=	1,333
				F(2, 1330)	=	37.43
Model	1324.95397	2	662.476985	Prob > F	=	0.0000
Residual	23540.6199	1,330	17.6997142	R-squared	=	0.0533
				Adj R-squared	=	0.0519
Total	24865.5739	1,332	18.6678483	Root MSE	=	4.2071

stress	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
1.sleep	-1.424046	.2373053	-6.00	0.000	-1.88958	-.9585124
age	.5649337	.1209043	4.67	0.000	.3277499	.8021176
_cons	9.873734	1.80558	5.47	0.000	6.33164	13.41583

```
. regress stress i.sleep age sleep_age_int
```

Source	SS	df	MS	Number of obs	=	1,333
				F(3, 1329)	=	25.08
Model	1332.27381	3	444.091269	Prob > F	=	0.0000
Residual	23533.3001	1,329	17.7075245	R-squared	=	0.0536
				Adj R-squared	=	0.0514
Total	24865.5739	1,332	18.6678483	Root MSE	=	4.208

stress	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
1.sleep	-3.71522	3.571474	-1.04	0.298	-10.72156	3.291121
age	.4792086	.1800052	2.66	0.008	.1260833	.832334
sleep_age_int	.1562448	.2430154	0.64	0.520	-.3204908	.6329804
_cons	11.1484	2.681808	4.16	0.000	5.887366	16.40944