PUBH 526 Randomized Trials

Homework 2

1. Before taking part in a physical activity intervention, 80 participants were given an activity tracker that logged their "active minutes" for a week. The intervention's aim was to increase the duration of physical activity. After receiving the intervention, the same participants wore an activity tracker again and logged their active minutes for another week. The SAS dataset "activity" has each participant's weekly active minutes before the intervention (the variable "before_intervention") and after the intervention ("after_intervention").

   Use SAS to answer the questions below. Answer the questions with sentences (not just numbers) and include your SAS code and output to support your answer. In interpreting p-values, use a significance level of 0.05.

   a. Graph the relationship between physical activity before the intervention vs. after the intervention. Describe in words this relationship. (2 points)

   b. Fit a linear regression line to this relationship. Write out the statistical model you are fitting. Please be rigorous in writing your model – it should involve Greek letters and subscripts. Make sure to explain all your notation. (3 points)

   c. What is the estimated regression model? (1 point)

   d. You will find that SAS conducts a statistical test of whether the slope of the regression line equals zero, and this statistical test yields a p-value of "<.0001". What does this p-value mean? What can you conclude from this statistical test? (2 points)

   e. Do you think the intervention "worked"? Justify your answer. (1 point)

   f. Do people who were more active before the intervention improve more after the intervention than people who were less active before the intervention? Justify your answer. (1 point)

   g. List the assumptions you are making when you fit a linear model to these data. (4 points)

   h. Do you believe your assumptions are reasonably satisfied? Justify your answer. (4 points)

   i. How much of the variation in post-intervention physical activity is explained by pre-intervention physical activity? (1 point)

Homework 2

Alexandra Chang
PUBH 526: Design and Analysis of Randomized Trials in Public Health

September 11, 2025

Before taking part in a physical activity intervention, 80 participants were given an activity tracker that logged their "active minutes" for a week. The intervention's aim was to increase the duration of physical activity. After receiving the intervention, the same participants wore an activity tracker again and logged their active minutes for another week. The SAS dataset "activity" has each participant's weekly active minutes before the intervention (the variable "before_intervention") and after the intervention ("after_intervention").

Question 1 (2 points)

Graph the relationship between physical activity before the intervention vs. after the intervention. Describe in words this relationship.

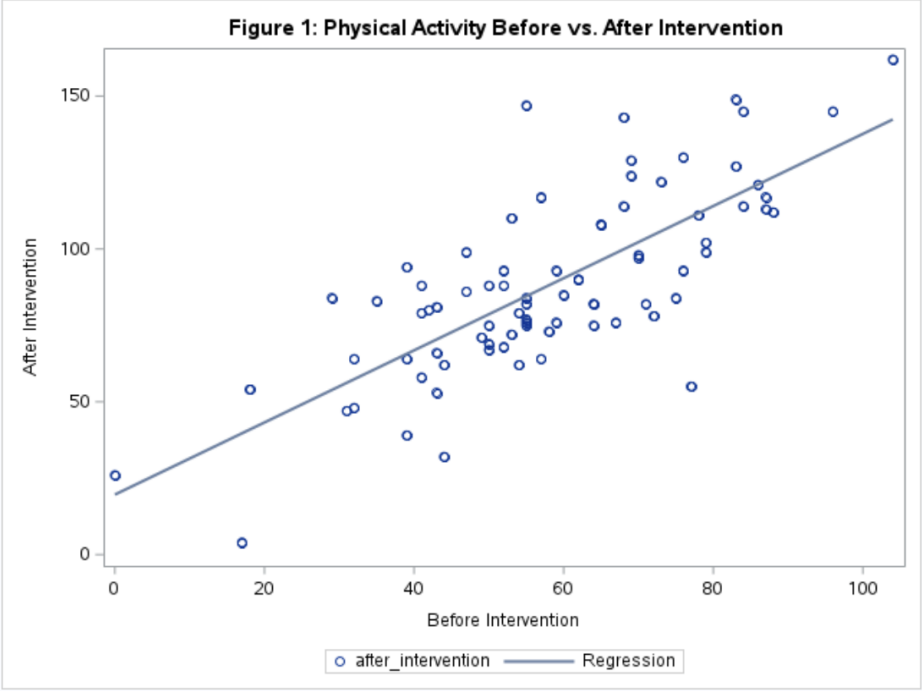


Figure 1 shows a scatterplot of physical activity scores before and after the intervention for each participant, along with a fitted regression line.

- The plot reveals a **positive linear relationship** between baseline activity and post-intervention activity. Participants with higher pre-intervention scores also tended to have higher post-intervention scores.
- The regression line suggests an **upward** trend, indicating that the intervention had a relatively consistent effect in the range of baseline activity levels.
- However, there are **some spread around the line**, especially at lower baseline levels, suggesting variability in individual responses to the intervention.
- No clear outliers or non-linear patterns are visible, supporting the use of linear regression in subsequent analyses.

Question 2 (3 points)

Fit a linear regression line to this relationship. Write out the statistical model you are fitting. Please be rigorous in writing your model – it should involve Greek letters and subscripts. Make sure to explain all your notation.

A simple linear regression model is needed to examine the relationship between physical activity before and after the intervention. The statistical model is shown below:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Where:

- Y_i = Physical activity score after the intervention for individual i (**outcome**)
- X_i = Physical activity score before the intervention for individual i (**predictor**)
- β_0 = **Intercept** of the regression line
- β_1 = **Slope** coefficient for baseline activity
- ε_i = **Error term** for individual i, assumed to follow $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

This model assumes that the mean post-intervention activity score increases linearly with baseline activity and that the residuals are normally distributed with constant variance.

Question 3 (1 point)

What is the estimated regression model?

Table 1: Linear Regression Results

The REG Procedure
Model: MODEL1
Dependent Variable: after_intervention

Number of Observations Read	80
Number of Observations Used	80

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	38925	38925	101.70	<.0001
Error	78	29854	382.73985		
Corrected Total	79	68778			

Root MSE	19.56374	R-Square	0.5659
Dependent Mean	88.23750	Adj R-Sq	0.5604
Coeff Var	22.17168		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	19.62337	7.14676	2.75	0.0075
before_intervention	1	1.18020	0.11703	10.08	<.0001

$$\hat{Y}_i = 19.62 + 1.18X_i$$

- This model suggests that for each **1-unit increase** in physical activity before the intervention, the post-intervention activity score **increases by 1.18 units** on average.

Question 4 (2 points)

You will find that SAS conducts a statistical test of whether the slope of the regression line equals zero, and this statistical test yields a p-value of "<.0001". What does this p-value mean? What can you conclude from this statistical test?

SAS conducts a hypothesis test for the slope (β_1) to determine whether there is a **statistically significant linear relationship** between baseline physical activity and post-intervention activity scores. The hypotheses are shown below:

$$H_0: \ \beta_1 = 0 \quad \text{(no association)}$$
$$H_a: \ \beta_1 \neq 0 \quad \text{(association is present)}$$

- SAS reports a **p-value** of < 0.0001, which is smaller than the significance level at $\alpha = 0.05$. This means the probability of observing such a strong association due to random chance (e.g. if the true slope were zero) is less than 0.01%.

- **Conclusion:** There is strong evidence to **reject the null hypothesis**. Hence, there is a **statistically significant linear relationship** between physical activity before and after the intervention. In other words, higher baseline physical activity is associated with higher post-intervention activity scores.

Question 5 (1 point)

Do you think the intervention "worked"? Justify your answer.

Table 2: Summary Statistics for Activity Scores

The MEANS Procedure

Variable	Mean	Std Dev
before_intervention	58.14	18.81
after_intervention	88.24	29.51

Yes, the intervention appears to have "worked." The justifications are shown below:

- Summary statistics show the mean **weekly active minutes increased** from 54.3 minutes to 82.7 minutes, indicating that the intervention was effective in increasing physical activity.

- Regression analysis shows a **significant positive association** between baseline and post-intervention activity scores, with a slope estimate of $\hat{\beta}_1 = 1.18$ (p-value < 0.0001), indicating that individuals with higher baseline activity levels tended to have even greater increases after the intervention.

- Intercept is estimated at 19.62, suggesting that even individuals with z**ero baseline activity showed a substantial improvement**.

- Overall increase in activity scores is visually evident in the raw data and the positive slope shown in the scatterplot (Figure 1), which aligns with the regression line showing improvement across the full range of baseline values.

Overall, these findings support the conclusion that the **intervention was effective in increasing physical activity levels**.

Question 6 (1 point)

Do people who were more active before the intervention improve more after the intervention than people who were less active before the intervention? Justify your answer.

Table 3: Correlation Between Baseline and Improvement

The CORR Procedure

2 Variables:	before_intervention improvement

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
before_intervention	80	58.13750	18.80799	4651	0	104.00000
improvement	80	30.10000	19.73277	2408	-22.00000	92.00000

Pearson Correlation Coefficients, N = 80
Prob > |r| under H0: Rho=0

	before_intervention	improvement
before_intervention	1.00000	0.17176 0.1277
improvement	0.17176 0.1277	1.00000

No, people who were more active before the intervention did not improve more than those who were less active.

- The Pearson correlation coefficient between baseline activity and improvement is $r = 0.17176$, indicating a **very weak positive linear relationship** between pre-intervention activity and improvement.
- In addition, the associated p-value is 0.1277, which is higher than the significance level at $\alpha = 0.05$.

Therefore, we **fail to reject the null hypothesis** that the correlation is zero. There is **no statistically significant evidence** to suggest that individuals who were more active before the intervention showed greater improvement after the intervention.

While the correlation is weak and not statistically significant, the direction of the association is slightly positive, indicating that more active individuals may have had modest improvements, but not enough evidence exists to confirm this pattern statistically.

Question 7 (4 points)

List the assumptions you are making when you fit a linear model to these data.

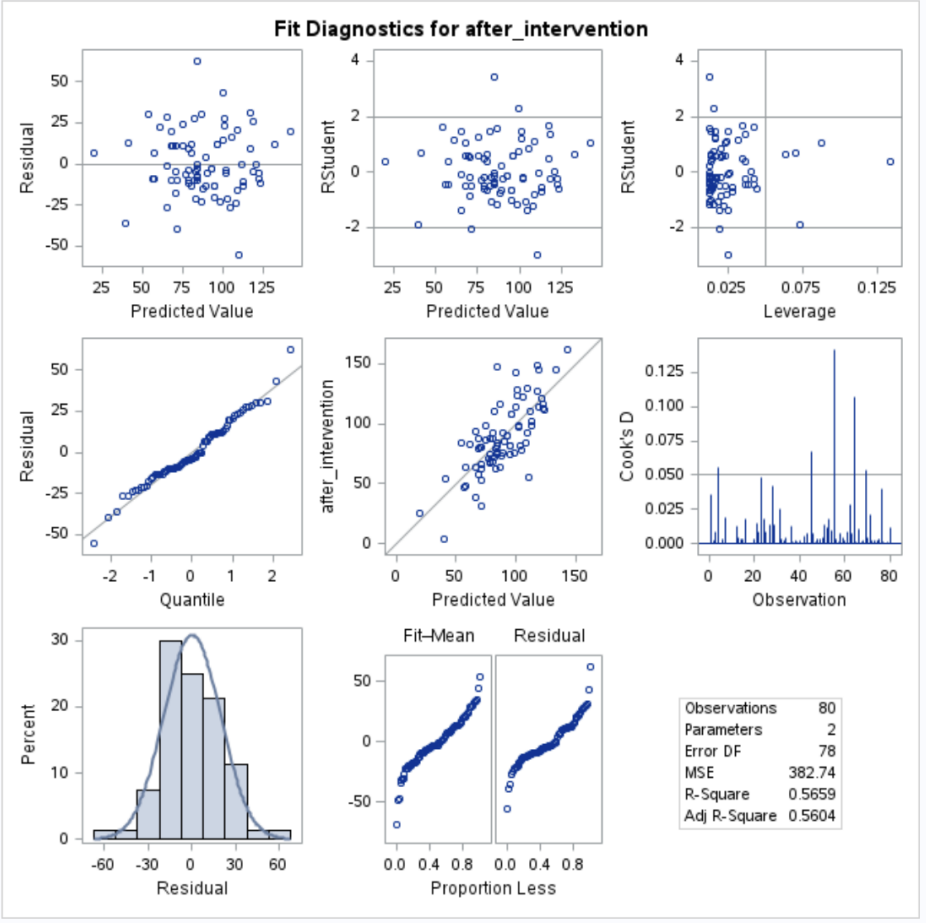
- **Linearity:** The relationship between the predictor X_i (e.g., physical activity before intervention) and the outcome Y_i (e.g., activity after intervention) is assumed to be linear.

- **Independence:** Each observation (i.e., each participant) is independent of the others. One individual's activity score does not influence another's.
- **Homoscedasticity:** The variance of the errors ε_i is constant across all values of X_i. This means the spread of residuals around the regression line stays consistent.
- **Normality of Errors:** The error terms ε_i are assumed to be normally distributed with a mean of 0. This allows for valid inference procedures such as confidence intervals and hypothesis testing.
- **Identically Distributed Errors (i.i.d.):** All error terms are drawn from the same distribution, ensuring the behavior of the residuals is consistent across all observations.

(Refer to HW1-3)

Question 8 (4 points)

Do you believe your assumptions are reasonably satisfied? Justify your answer.



Yes, the assumptions for linear regression appear to be reasonably satisfied based on the fit diagnostics. The justifications are shown below:

- **Linearity:** The scatterplot of outcome v.s predicted values shows a generally **linear trend**. The residuals v.s. predicted values plot is evenly **distributed around zero** with no curvature, suggesting a linear relationship is appropriate.
- **Independence:** Each point represents a different individual. There is **no obvious clustering** or repeated measures, so the independence assumption is likely satisfied.
- **Homoscedasticity:** The residuals appear to have a roughly **even spread across** the range of predicted values in both the residual v.s predicted and studentized residual v.s leverage plots. No strong funneling or fan shape is present.
- **Normality of Errors:** The Q-Q plot and histogram of residuals show that the residuals are approximately **normally distributed**, with only a slight deviation in the tails.
- **Identically Distributed Errors:** The Cook's D and studentized residual v.s leverage plots show that the residuals appear **evenly distributed across** the sample, and **no major outliers or high-leverage points** present.

Overall, these findings support the conclusion that the **diagnostic plots support the use of a linear regression model for the data**.

Question 9 (1 point)

How much of the variation in post-intervention physical activity is explained by pre-intervention physical activity?

$$R^2 = 0.5659$$

- The amount of variation in post-intervention physical activity explained by pre-intervention physical activity is given by the **coefficient of determination,** R^2.
- In other words, approximately **56.59% of the variation in physical activity after the intervention is explained by the baseline physical activity**.

(Refer to HW 2-3)