MATH 254: Introduction to Statistics **Mintaek Lee**

Lesson 6: Simple Linear Regression with Inference for Slope

Corresponding Workbook Module: 6

Example 6.1 *(Testing whether the relationship between x and y is significant)*
Suppose Mintaek wanted to know if the average number of hours students spend on social media per day is linearly related to their first year GPA or not. He randomly selects 30 students from a university and measures how many hours they spend on social media per day. Use the appropriate statistical methods and set up the hypotheses that best match Mintaek's interests.

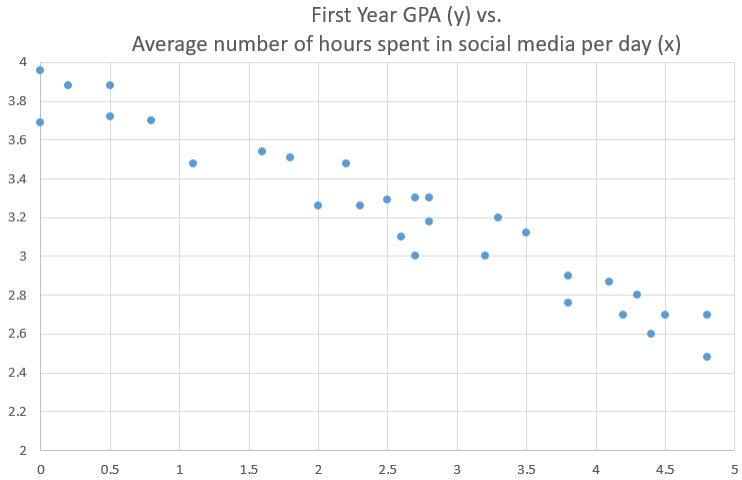
We first need to recognize that **Social** (the average number of hours students spend on social media per day) is the explanatory/independent (x) variable and **GPA** (the first year GPA) is the response/dependent (y) variable.

Step 1: Null and alternative hypotheses would then be:

- H_0: $\beta_1 = 0$, the average number of hours students spend on social media per day and the first year GPA are NOT linearly related.
- H_A: $\beta_1 \neq 0$, the average number of hours students spend on social media per day and the first year GPA are linearly related.

Step 2: We now need to check conditions. For regression problems, we use the "**LINE**".

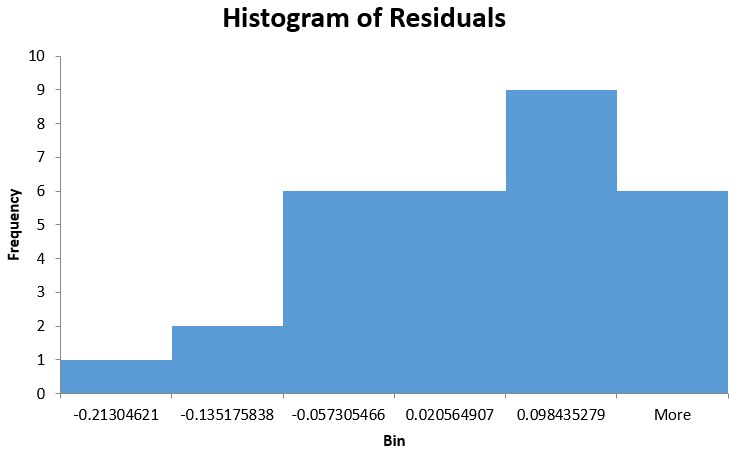
- **L**inear: We need to make sure that the pattern looks linear. We can check this by observing the scatterplot.



Looking at the **scatterplot**, the relationship between x and y appear to be linear.

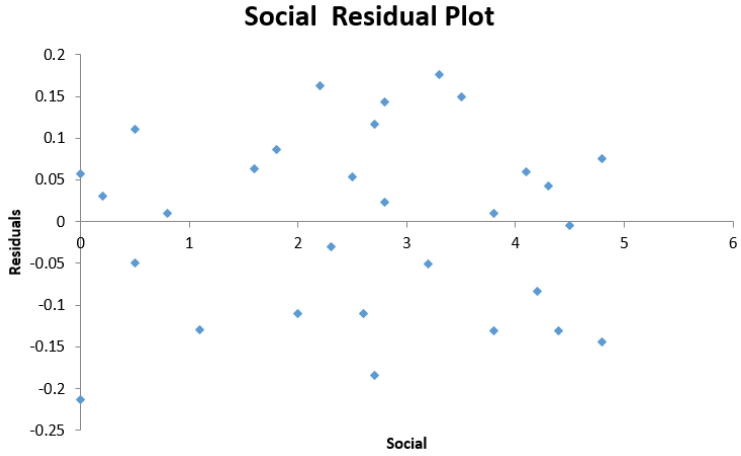
- **I**ndependent observations: We need to make sure that observations were collected independently (randomly). Since Mintaek selected 30 students randomly, we can say that this condition is met.

- **N**ormally distributed residuals: We need to make sure that residuals appear to have come from a normal distribution. We can check this by looking at the histogram of residuals.



Looking at the **histogram of residuals**, we see that the distribution of residuals appear to be fairly symmetric without any noticeable skewness. We can assume that this condition is met.

- **E**qual variation: We need to make sure that variation of residuals is consistent throughout the explanatory or independent (x) variables. We can check this by looking at the residual plot.



Looking at the **residual plot**, we can see that variation of residuals is consistent throughout different x values.

Step 3: Since we verified all the necessary conditions, we can now fit the least squares regression line and find t-test statistic and P-value. You do not need to calculate any of them by hand. You will either be required to use software or appropriate outputs will be provided to you.

The software output for the regression analysis is shown below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.96448735
R Square	0.93023585
Adjusted R Square	0.92774428
Standard Error	0.11094408
Observations	30

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.90304621	0.041101817	94.96042996	1.11E-36	3.818852955	3.987239466	3.818852955	3.987239466
Social	-0.26647026	0.013790796	-19.32232617	9.999E-18	-0.294719426	-0.238221096	-0.294719426	-0.238221096

From the output above, we find the t-test statistic value of $t = -19.3223$ and corresponding P-value of 9.999×10^{-18} (approximately 0) for the slope term.

Step 4:

Since P-value was 9.999×10^{-18} (approximately 0), thus smaller than the standard significance level $\alpha = 0.05$, we would reject our null hypothesis that the average number of hours students spend on social media per day and the first year GPA are NOT linearly related. Therefore, we conclude that there is significant evidence to suggest that the average number of hours students spend on social media per day and the first year GPA are linearly related.

Example 6.2 *Since Mintaek verified that the average number of hours students spend on social media per day is linearly related to the first year GPA, he wants to make some inferences. Below are some short questions you can answer based on the findings above. Relevant software outputs are also shown below.*

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.96448735
R Square	0.93023585
Adjusted R Square	0.92774428
Standard Error	0.11094408
Observations	30

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.90304621	0.041101817	94.96042996	1.11E-36	3.818852955	3.987239466	3.818852955	3.987239466
Social	-0.26647026	0.013790796	-19.32232617	9.999E-18	-0.294719426	-0.238221096	-0.294719426	-0.238221096

Technically, you still would need to check all conditions (LINE) for the simple linear regression. Since they were checked in Example 6.1, it will be omitted here.

Part A. *Using the software output given above, find the least squares regression line for the given data.*

We look at the *Coefficients* column on the bottom table. We see that the number corresponding to Intercept is 3.903, which is b_0 and the number corresponding to Parties is -0.266, which is b_1.

We therefore find $\hat{y} = b_1 + b_0\, x = 3.903 - 0.266x$ as the least squares regression line for the given data.

Part B. *Use the regression line for this data to predict the first year GPA of a student who spent 4 hours on social media per day on average.*

We substitute 4 into x since x was the explanatory variable (the average number of hours students spend on social media per day). We find $\hat{y} = 3.903 - 0.266 \times 4 = 2.84$.

A student who spent 4 hours on social media per day on average is predicted to have a first year GPA of 2.84.

Part C. *Use the regression line for this data to predict the first year GPA of a student spend 15 hours on social media per day on average.*

This time, we find $\hat{y} = 3.903 - 0.266 \times 15 = -0.09$.

Note that 15 hours on social media per day is outside the given data range for average number of hours students who spent on social media per day. So, we don't necessarily know if the trend would continue to be linear for $x = 15$. This is called an extrapolation which is a poor statistical practice. Furthermore, one cannot simply obtain a negative GPA. So we do not make a prediction for this student.