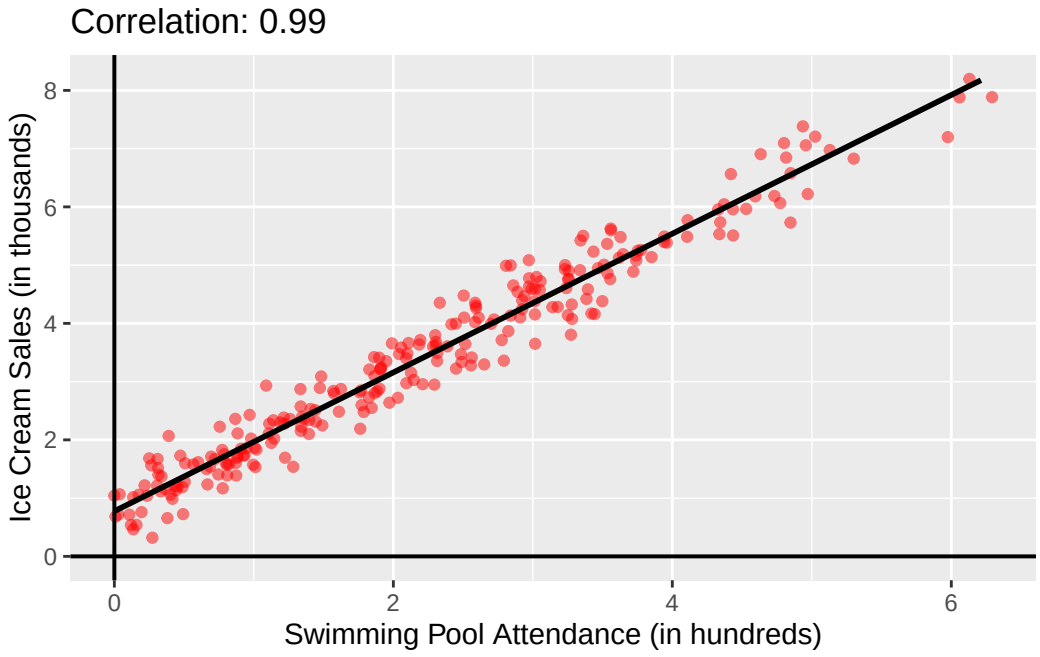
Midterm Assignment

Lajos Galambos

Question 1

Correlation Does Not Imply Causation Simulated data for two variables: ice_cream_sales (in thousands) and swimming_pool_attendance (in hundreds), over 365 days, assuming both increase during summer due to the weather but are not causally related.



I adjusted the graph to display the positive sections of the functions that were provided.

Table 1: Statistics

Statistic	Value
Correlation (Sales and Attendance)	0.9900

Statistic	Value
R-squared (Sales and Attendance)	0.9802
R-squared (Attendance and Temperature)	0.9847
R-squared (Sales and Temperature)	0.9941

1. **What is the correlation coefficient between ice_cream_sales and swimming_pool_attendance?**

The correlation is 0.9900339, which is high and positive. Ice cream sales are highly and positive correlated with swimming pool attendance. If we were to increase swimming pool attendance we could expect increased ice cream sales, on average.

2. **Does this correlation imply that increased ice cream sales cause higher swimming pool attendance or *vice versa*? Explain why or why not. What is the R square for each specification? What do you conclude?**

The increased ice cream sales do not cause higher pool attendance, nor do increased pool attendance cause increased ice cream sales. Higher ice cream sales are associated with higher pool attendance and *vice versa.* The relationship is positive, highly correlated but not causal.

This is not a causal relationship because first, we know that from the description of the task, second, we know nothing about the setup of the experiment (how was the independence of potential outcomes ensured).

The R-squared statistics are used for prediction purposes. High R-squared values imply high predictive power. In this case, in all specifications, the R-squared vales are high, implying that we could consider ice cream sales for the prediction of swimming pool attendance; and we could use swimming pool attendance for predicting ice cream sales, and so on. R squerd wales are identical for both directions: R squared of temperature and ice cream sales are identical to ice cream sales and temperature. Would we use R squared to predict temperature? Well, the temperature has been assumed to be the variable that has an exogenous source of variation. The functional forms of the variables also confirm this. Therefore in my oppinion, using R sruared values to estimate temeprature given ice cream sales or swimming pool attendance is techincally viable but in theoretically it is a bit odd.

3. **Identify the intruder in this scenario and discuss its role.**

Table 2: Statistics

Statistic	Value
Correlation Sales and Temperature	0.9971
Correlation of Attendance and Temperature	0.9923

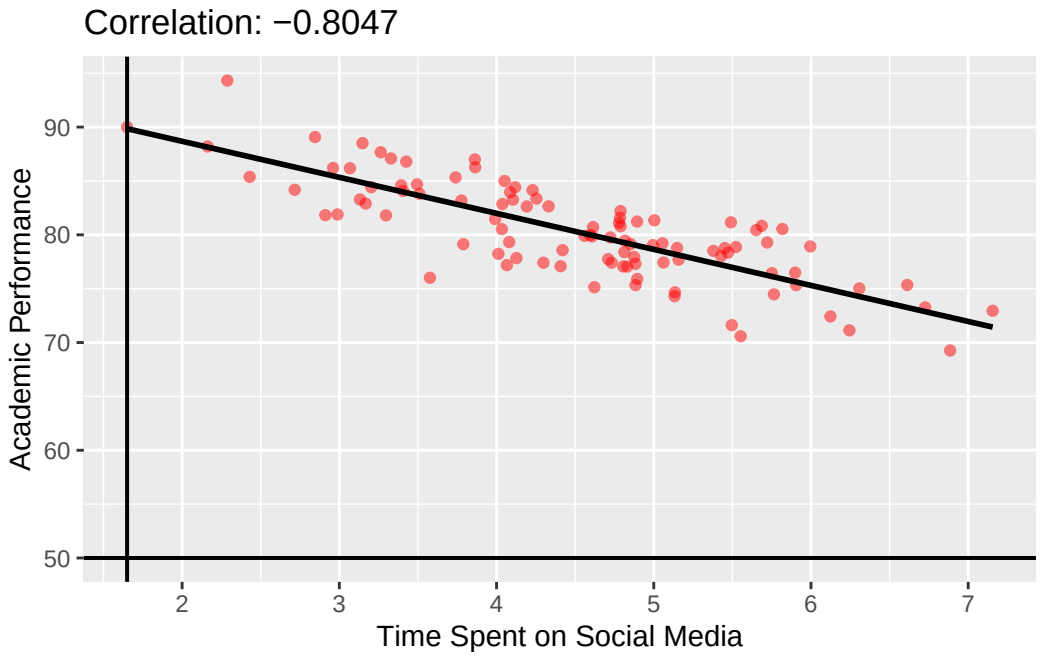
The intruder is the variable that effects other variables, it is a confounder variable. As such, temperature is highly positively associated with both ice cream sales and swimming pool attendance. We can also assume that the temperature is the intruder since other variables, such as swimming pool attendance and ice cream sales have a functional relationship with temperature.

Question 2

Two variables can be correlated due to confounding factors rather than a direct causal relationship.

Simulated data for three variables: time_spent_on_social_media (hours), academic_performance (score), and stress_level (score). Assume stress_level affects both time_spent_on_social_media and academic_performance, creating a spurious correlation between the latter two.

1. **What is the correlation coefficient between time_spent_on_social_media and academic_performance? Plot it.**



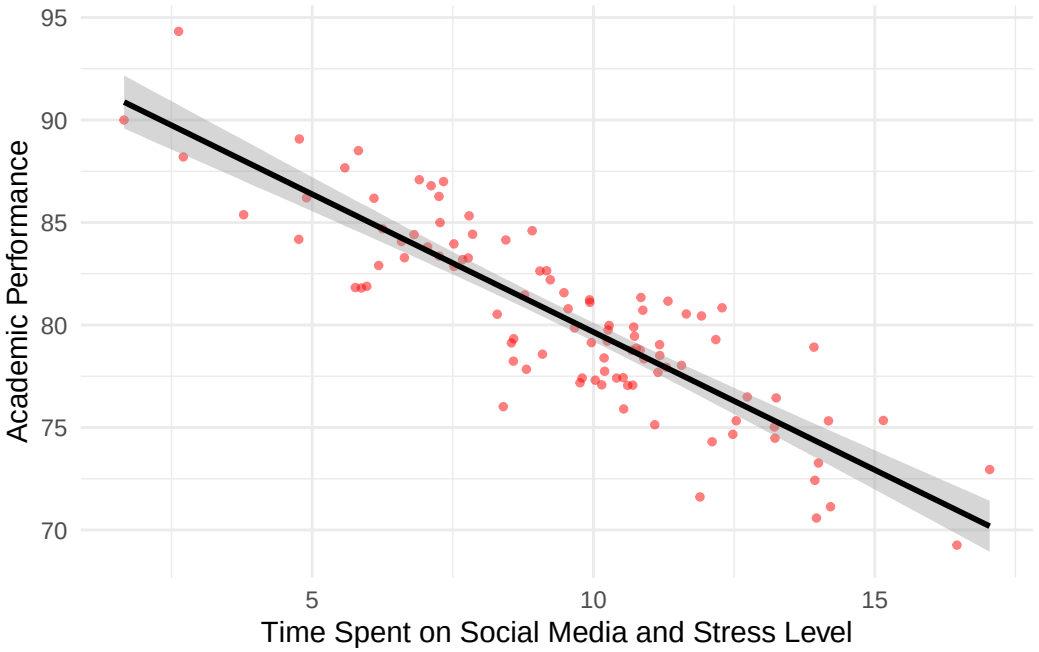
I have adjusted the graph's x axis to start from the lowest observation, and the y axis to start from 50. I did this to see more detail.

The correlation between academic performance and time spent on social media is -0.8046621, which is a strong negative correlation. If we were to increase the time spent on social media we could expect lower academic performance, on average. No causality can be inferred.

Table 3: Summary of Multivariate Regression Model: Academic Performance on Time Spent on Spcial Media and Stress Level

term	estimate	std.error	statistic	p.value
(Intercept)	91.1366	1.1175	81.5521	0.0000
time_spent_on_social_media	-0.1871	0.4903	-0.3817	0.7035
stress_level	-1.9931	0.2815	-7.0794	0.0000

2. **What is the specification you would estimate to understand the effect of time spent on social media on academic performance.**



To give a specification that estimates the effect of time spent of social media on academic performance, we should not forget that the simulated data it was assumed that stress_level affects both time_spent_on_social_media and academic_performance, creating a spurious correlation between the latter two. So I should think about creating a model, where I include the confounding variable: stress level; time spent on social media (to account for the spurious correlation); and academic performance as the outcome variable. Based on the regression outputs, it can be stated that stress level has significant negative effect on academic performance. There is a good reason to think that stress level is the right variable to create causal specification, as both academic performance and time spent on social media are functional outcomes of the stress level.

3. **Explain why this correlation might be misleading in concluding a causal relationship between time spent on social media and academic performance.**

Table 4: Statistics

Statistic	Value
Correlation Time Spent on Soc.Media and Stress Level	0.9098
Correlation of Acadmeic Performance and Stress Level	-0.8759
Correlation Academic Performance and Time Spent on Soc.Media	-0.8047

First, I know from the task description that the time spent on social media and academic performance has a spurious correlation. It is correlation only, and spurious, which means that there is an underlying factor that affect both. Excluding the confounding variable, would be misleading.

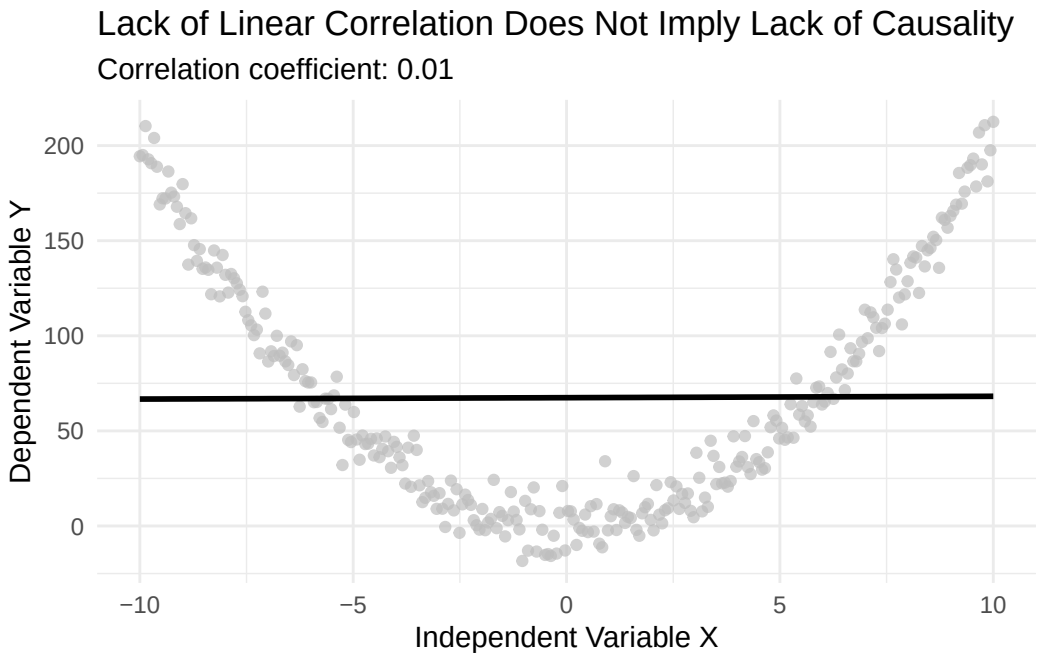
Second, as it has been demonstrated in the previous exercise, relying on the assumptions by the task description, stress level acts as a confounder, and it has causal effect on academic performance.

4. **How could you investigate if stress_level is a confounding factor?**

By looking at the correlations (**Table 4**), I can see that time spent on social media is highly and positively correlated with stress level; and academic performance is highly and negatively correlated with stress level. Furthermore, as demonstrated before, academic performance and time spent on social media is negatively correlated. Furthermore, I know from the functional forms of academic performance and time spent on social media variables that stress level has a functional effect on both.

Question 3

Simulate data to show that lack of correlation does not imply lack of causality.




```
x <- seq(-10, 10, length.out = 300)
y <- 2 * x^2 + rnorm(300, mean = 0, sd = 10)
```

The graph shows simulated data and "y" is quadratic functional form of "x" (y <- 2 * x^2). The implies a correlation that is close to zero, however, still the functional form implies strong relationship between "y" "x". It can still be states that "x" causes "y" regardless of the close to zero correlation. Causality can exist without correlation.

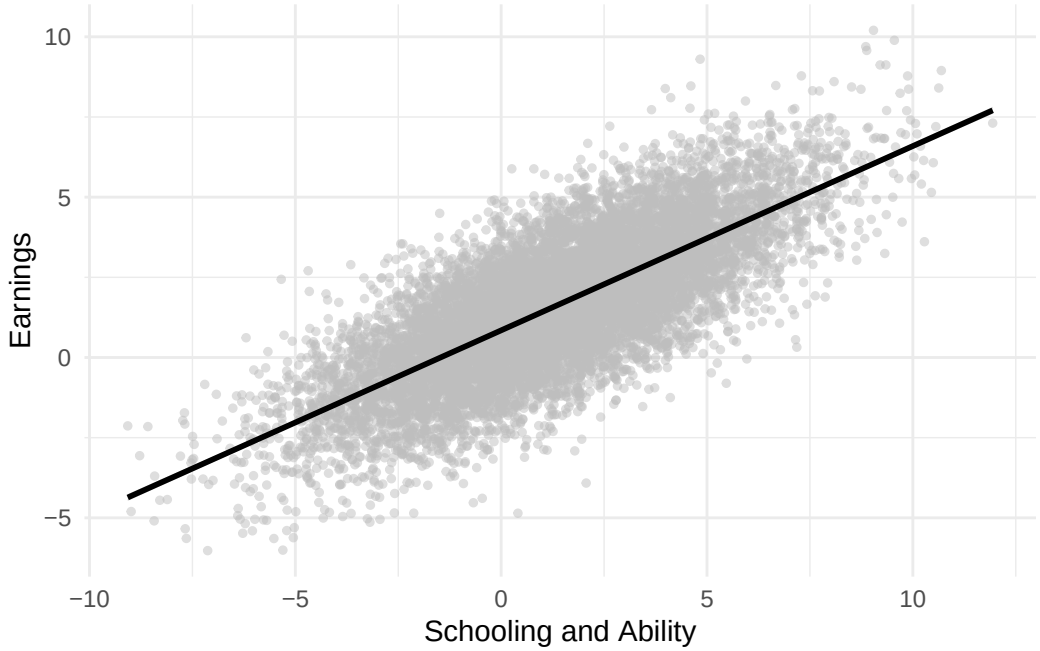
Question 4

What story does the code below tell? Interpret the code and comment on your findings.

Table 5: (Model 1)Summary of Multivariate Regression Model: Earnings on Schooling and Ability

term	estimate	std.error	statistic	p.value
(Intercept)	0.4986	0.0032	153.6527	0
schooling	-0.4007	0.0022	-181.4569	0
ability	1.2013	0.0016	745.2835	0

Model 1 - Regression model of earnings on schooling (plus ability as a confounder):



Model 1, where the outcome variable is earnings and schooling is the dependent variable, a confounder is added. Ability affects both, therefore in the linear model schooling affects earnings, controlling on ability.

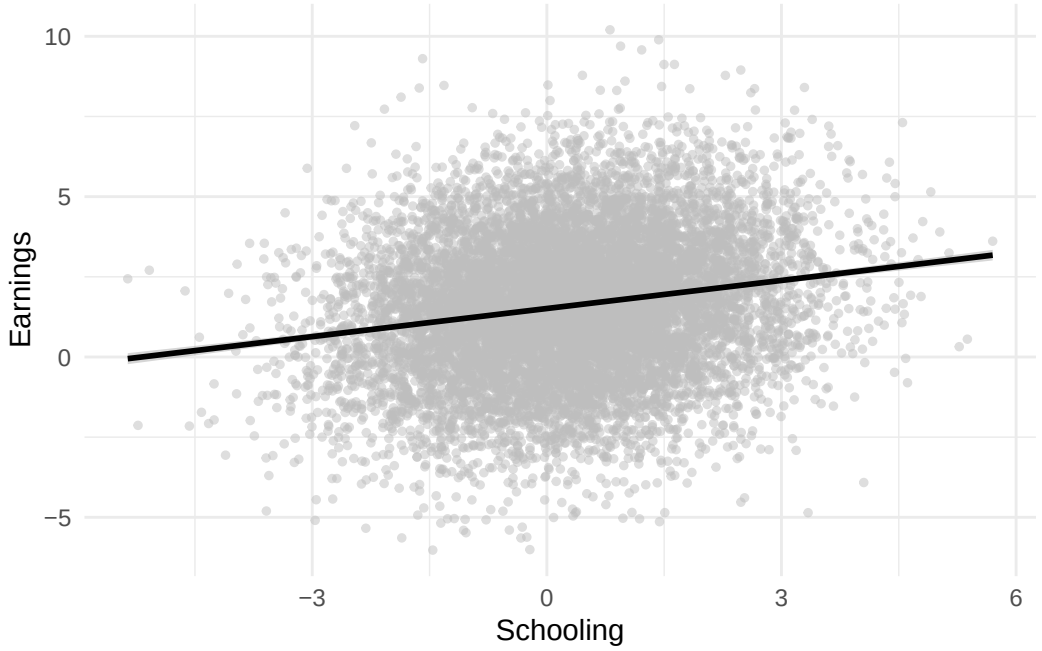
The results from the regression outputs tell us that if we were to hold ability constant, there are negative returns in earnings to schooling; meanwhile, when we hold schooling constant, there is a positive returns in earning to ability. Coefficient estimates are significant.

I also get back the same coefficients from the regressions as in the pre-specified functions for variables. That means that the regression is unbiased, I included all the existing variables.

Table 6: (Model 2)Summary of Regression Model: Earnings on Schooling

term	estimate	std.error	statistic	p.value
(Intercept)	1.5113	0.0222	68.1888	0
schooling	0.2918	0.0151	19.3704	0

Model 2 - Regression model of earnings on schooling (without confounder):



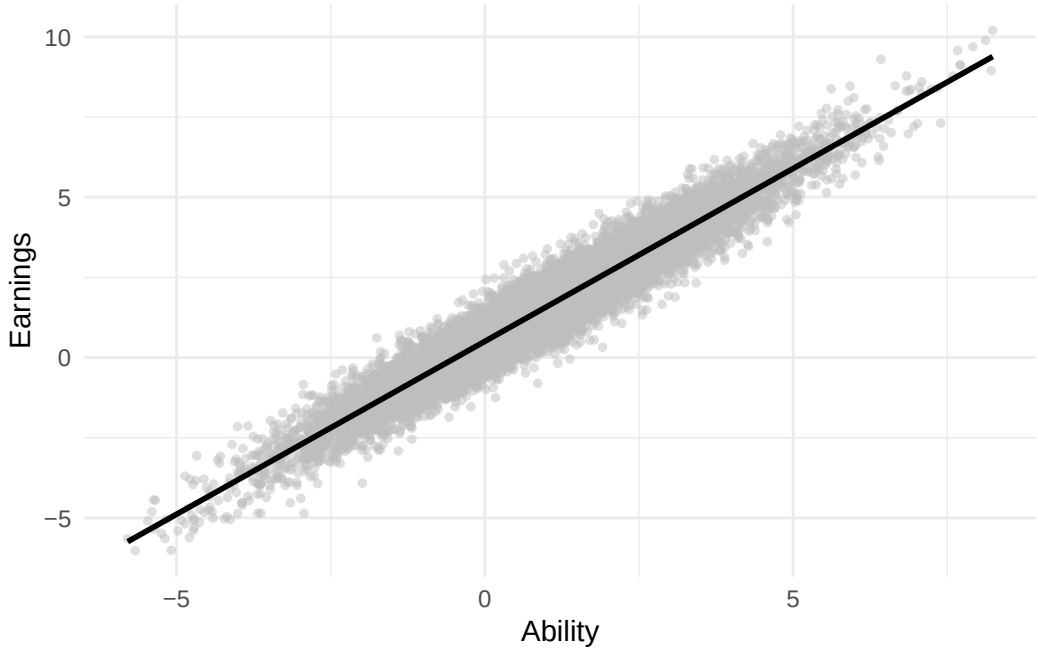
Model 2 shows how earnings are correlated with schooling. It displays the scenario if only schooling was is linked with earnings, what returns do earning have to schooling. It is a positive relationship. For additional years of schooling earnings are expected to increase. The coefficient estimate is weak and significant.

Model 2 has a big omitted variable bias, because as it have been seen above, ability plays a crucial role. Since the coefficient for schooling changes significantly between Model 1 and Model 2, this suggests that ability is an important factor that influences both schooling and earnings. This comparison helps in understanding the potential biases and limitations of simpler models that exclude relevant variables.

Table 7: (Model 3)Summary of Regression Model: Earnings on Ability

term	estimate	std.error	statistic	p.value
(Intercept)	0.5056	0.0067	75.1913	0
ability	1.0782	0.0030	355.8809	0

Model 3 - Regression model of earnings on ability:

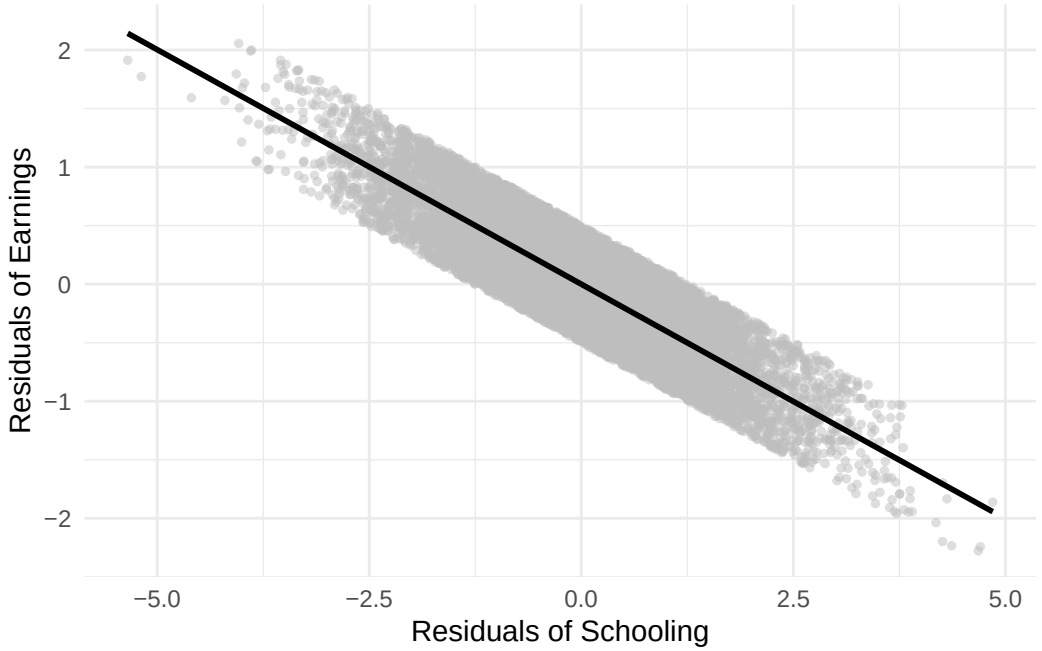


Model 3 is a linear model where the relationship of ability and earnings is being investigated. This model aims to assess the association between an individual's ability and their earnings, independent of their level of schooling. The results from the regression output tell us that given that positive and singificant coefficient, higher ability is associated with higher earnings, all else being equal. One unit increase in ability has a positive return in earnings.

Table 8: (Model 5)Summary of Regression Model: Rsiduals of Earnings on Residuals of Schooling

term	estimate	std.error	statistic	p.value
(Intercept)	0.0000	0.0029	0.000	1
res_schol	-0.4007	0.0022	-181.466	0

Model 4,5 - Regression model of residuals:



Model 4 displays how much of the variation in schooling can be explained by ability alone in a linear model. Residuals of from this model (res_schol) are the residuals from the regression of schooling on ability, representing the part of schooling not explained by ability.

Model 5 gives a linear model for linking the residuals of model 4 (residuals from linear model of schooling on ability - "res_schol") and derives the residual from model 3 ("earnings_res"), the residuals from the regression of earnings on ability, representing the part of earnings not explained by ability.

The coefficient estimates tell that schooling residuals have a significant and negative relationship with earnings residuals. It suggests that, once the effect of ability is accounted for, an increase in the schooling residuals is associated with a decrease in earnings residuals.

Also, the res_school coefficient gives the same estimate as the unbiased regression (with confounder).

Table 9: Summary of Regression Model with Measurement Error in the Main Regressor

term	estimate	std.error	statistic	p.value
(Intercept)	7.5633	0.3335	22.6799	0
X_observed	0.3871	0.0065	59.9719	0

Table 10: Summary of Regression Model with Measurement Error in the Outcome

term	estimate	std.error	statistic	p.value
(Intercept)	2.1049	0.8333	2.5261	0.0117
X_true	0.4966	0.0163	30.4725	0.0000

Question 5

Simulate how well behaved measurement error (in the main regressor or outcome of interest) affects your estimated coefficient of interest. What can you tell about departures from the well behaved instances?

Measurement error in the main regressor (above Table 9):

This is the first scenario for measurement error. Let us assume that we are in a simple linear model with Y as the dependent variable with a constant and an independent variable with Beta1 coefficient plus an error term. In my simulation "model_observed" is being compared to "model_true".

Measurement error in the independent variable (regressor) leads to attenuation bias, where the absolute value of the estimated coefficient is biased toward zero compared to its true value, as the regression output also demonstrates (0,49 vs. 0.38). This bias arises because the variability of the observed regressor is inflated due to the measurement error, weakening the observed association between the regressor and the outcome.

Measurement error in the outcome (above Table 10):

Measurement error in the dependent variable generally does not bias the coefficient estimates but increases the variance of the error term. This results in less precise estimates (higher standard errors), potentially leading to wider confidence intervals and making it harder to detect significant relationships.

Table 11: Summary of Regression Model with Correlated Measurement Error

term	estimate	std.error	statistic	p.value
(Intercept)	6.5726	0.3635	18.0831	0
X_observed_with_correlated_error	0.4107	0.0071	57.8882	0

Table 12: Summary of Regression Model with Heteroscedastic Measurement Error in the Outcome

term	estimate	std.error	statistic	p.value
(Intercept)	1.7289	0.4767	3.6265	3e-04
X_true	0.5057	0.0093	54.2404	0e+00

Non-well behaving scenario:

Model with correlated measurement errors (Table 11):

The correlation between the measurement error and the true value of can lead to biased estimates of the regression coefficients. Unlike classical random measurement error, which primarily affects the precision of the estimates (increasing their standard errors), correlated errors can bias the coefficients in unpredictable directions.

Model with heteroscedastic errors in the outcome (Table 12):

Heteroscedasticity is a phenomenon in regression models where the variance of the errors, or residuals, is not constant across all levels of the independent variables. Heteroscedasticity in the measurement error of the dependent variable violates one of the key assumptions of linear regression, potentially leading to inefficient estimation and inaccurate standard errors. This can affect the reliability of confidence intervals.