Suppose that we are interested in understanding the effect of X_1 on Y. The model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

ⓐ Would you include the variable $X_1 \times X_3$ as a regressor in the model? Explain this in detail.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_1 \times X_3) + u$$

Interaction terms are often added to models to capture potential interaction effects

ⓑ Suppose that in the model in (1) all the assumptions for the unbiasedness of OLS hold, but the errors are heteroskedastic. What would you do? Make sure that you mention the reasons for your decisions.

If errors are heteroskedasticity the ordinary least squares (OLS) estimates remain unbiased, but they are no longer the best linear unbiased estimates (BLUE) because their variance is not minimized.

ⓒ The model is far from complete. Propose 2 other variables which should have been included?

- Interest rates, inflation

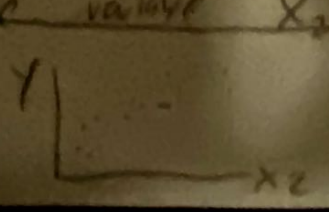
Income / wealth

- time

Geographic region

ⓓ Do you believe it is appropriate for the model to be linear in the variable X_2? Discuss this matter in detail.

plot it



if relation seem straight, it's linear

or check R^2

if not transform log(x)

(5) Suppose that $\hat{\beta}_1 = 0.04$ and $SE(\hat{\beta}_1) = 0.015$. A 5% significance level, can we say that X_1 affects Y? Why?

$H_0: \beta_1 = 0$ (No affect)

$H_1: \beta_1 \neq 0$ (No affect)

$$t = \frac{0.04}{0.015} = 2.67$$

Significant level = 1.96

Since $|t| = 2.67 > 1.96$, we reject H_0. X_1 has significant effect on Y.

(6) Based on the hypothesis test from (5), how does the result impact society? Which recommendations do you have?

Since X_1 has significant effect on Y, suggest educational changes may matter but won't solve societal issues

(5) How would you test the hypothesis that the variable X_2 is superfluous in the model.

$H_0: \beta_2 = 0$ perform F-test

$H_1: \beta_2 \neq 0$

(7) How would you reply to the following statement: "Statement about one or more coefficients" in a scientific manner?

- To test hypothesis test about coefficients
- State Null and alternative hypothesis
- Use test statistics
- Evaluate p-values against significance levels and draw conclusions.

Interpret results with consideration of assumption.

(J) Suppose that the true model is actually

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_3 + u$$

but you run the model in (1) what is the consequence of this?

leads to misspecification bias which leads to omitted variable bias, The estimated coefficients for the remaining variables may be bias and inconsistent b/c the model will not capture the curve between y and x_1

(K) What is the meaning of R^2 of a regression? How do you expect R^2 to change from the model in (2) to the one in (1)?

R2 represents the proportion of variance in Y and X

R^2 will decrease b/c were omitting X_1^2 in model 2 which lead to more variation in Y the model (1).

(L) Assume the estimated coefficients for β_1 and β_2 in (2) are 0.04 and 0.001 respectively and $x_1 = 1$, keeping everything else constant, what is the predicted change in Y when we increase x_1 by 1 unit. What about when $x_1 = 5$

$x_1 = 1$: $0.04(1) + 0.001(1)^2 = 0.04 + 0.001 = 0.041$

$x_1 = 5$: $0.04(5) + 0.001(5)^2 = 0.20 + 0.025 = 0.225$

(M) Suppose that our data where Y is self reported. what problem we may have? What is the consequence of this problem?

Lead to measurement errors and inaccuracies in model, produce biased and inconsistent parameter in regression model.

(n) Suppose that X_1 is missing for 5% of the observations. What problems do you fear? How would you handle this?

we will face selection Bias, to handle this we perform stratified sampling.

(o) X_4 is likely to affect Y. However, we do not have access to data about it. Do you expect the bias of $\hat{\beta}_1$ to be positive or negative?

if X_4 affects Y and X_1 and is omitted, the bias of $\hat{\beta}_1$ depends on the correlation:

- if x_4 is positively correlated with Y and x_1, omitting it leads to upward bias.

(p) Y should depend on x_5. List at least one control variable that we could use as proxy for x_5. Explain your variable of choice.

Proxy - when Y depends on X_5, but X_5 data is unavailable

Ex) X_5 = interest rate, proxy is now GDP

(q) Suppose that your data was collected only on certain type of area. Do you fear some form of sample selection? What would be the implication on the estimated coefficient of x_1? Elaborate on this issue.

data collected in a certain area is not representative of entire population, leads to sample selection bias

Leads to biased estimates for β_1 as it does not reflect actual population.

① Suppose that you have alot of variables which are connected to the topic. It may be related to the topic substantially or tangentially. Explain your decison process about including those variables or not.

- we need to consider how relevant is the variables.
- we need also need to exclude highly collinear variables to avoid inflating standard errors.
- And recognizes that too much variables can lead to overfitting.

Essay Question

Prompt:

with the growth of industrialized food, high level of
suger consumption and large use of unneeded medications,
American policy makes have become more and more concerned
with the health of its population. Poor health
jeopardizes people's welfare and generates high monetary
costs for society.

In economics the link between health and income seems
pretty uncontroversial. After all, healthier people can work longer
and harder. Healthier children are likely to stay in school
longer and learn more, earn more. Even across countries
the relationship seems clear: those with better health are
generally richer, and those that improve their citizen's
health grow faster.

So consider the following question: what is the effect
of physical activity on type 2 diabetes? Your task is
to describe how you might go about answering this
question. Construct a model, and be specific about how
you define your variables. Justify your choices and
describe the assumptions you need. Explain how you might
go about finding the data you need. Comment shortcomings
of your approach, and provide solutions you would implement
when possible.

The goal of this essay is to develop an empirical study in order to understand the effect of physical exercise on the likelihood of developing Type 2 diabetes.

We will need two groups of people. Group 1 is the treatment group that will be assigned to do physical exercise for a certain amount of time per week. Group 2 will be the control, where they will live their regular lives. The allocation of people for this study will be done randomly. It will also be possible to divide the treatment group into sub groups of people based on hours of physical exercise per week / or frequency of exercise per week, and whether more exercise ~~of~~ or frequency of exercise matters with the likelihood of being diagnosed with Type 2 Diabetes.

In our data set needed for this experiment, we will need 1 column of hours or frequency of exercise done by individual, and whether the individual has been diagnosed with Type 2 diabetes.

$$\text{Diabetes}_{it+1} = \beta_1 \text{Exer} + \beta X + \varepsilon$$

Our dependent variable will be called Diabetes, and this variable will take value 1 if the individual is diagnosed with diabetes and 0 otherwise. For our continuous variables, Exer represents the average number of exercises done in a week.

Additionally, we would include controls for individuals personal charactistics - such as family history, race/ethnicity. This will mitigate the omitted vaible bias to be correlated with exercise an affect the probability of developing type 2 diabetes.

β_1 can be interpreted as the change in the probability of being diagnosed with type 2 diabetes by increasing 1 hour of average weekly exercise.

There an requires assumption to interpret the coefficient of β_1 as the causal effect of exercise in the probability of being diagnosed with type 2 diabetes, is this conditional with the other variables, is true nothing else that is correlated with the amount of exercise carried out ~~in~~ t that could also affect the probability of being diagnosed with type 2 diabetes.

One major shortcomings of this approach is the potential existence of variables that change the likelihood of being diagnosed with type 2 diabetes and are also correlated with exercising. For example, alchool consumption or sugar consumption. If they were present in our dataset, we would include them as controls. An additional problem would be the existence of relatively large measurement error in Exer. This measurement error would bias down our estimates of the effect of exercising on the probabity of being diagnosed with diagnosed with type 2 diabetes.