Econometrics

Seminar 3: Classical linear model assumptions

Problem 1. Let *reading_score* denote the average reading comprehension score of students at an elementary school. Suppose we wish to estimate the effect of a free after-school tutoring program on student achievement. If anything, we expect the tutoring program to have a positive ceteris paribus effect on achievement: if a student whose parents work long hours and cannot help with homework gains access to after-school tutoring, his or her reading scores should improve. Let *tutor_prg* denote the percentage of students at the school participating in the free tutoring program.

(i) Suppose you are given funding to run a controlled experiment. Explain how you would structure the experiment in order to estimate the causal effect of *tutor_prg* on *reading_score*.

(ii) Consider the more realistic case where you can only randomly sample *reading_score* and *tutor_prg* from the population. Write the population model as

$$reading\_score = \beta_0 + \beta_1 tutor\_prg + u$$

where, as usual in a model with an intercept, we can assume $E(u) = 0$. List at least two factors contained in u.

(iii) Will a simple regression analysis uncover the ***ceteris paribus effect*** of *tutor_prg* on *reading_score*? Explain.

(iv) In the equation from part (ii), what should be the sign of β_1 if the free tutoring program is effective? What is the interpretation of β_0?

Problem 2. Consider the dataset *StudentsPerformance.csv*. This dataset is a synthetic representation of student performance, designed to mimic real-world scenarios by considering key factors such as study habits, sleep patterns, socioeconomic background, and class attendance. The following variables are present in the data set.

StHours:	Average daily hours spent studying.
SEScore:	A normalized score (0-1) indicating the student's socioeconomic background.
SleepHours	Average daily hours spent sleeping.
Attendance:	The percentage of classes attended by the student.
Grades:	The final performance score of the student.

1. Discuss what data structure do we have. Classify the features. Analyze the variable *Grades*. Compute the corresponding descriptive statistics and interpret the results.

2. Compute Pearson and Spearman correlation coefficients for Grade vs. other features. Comment on the results. What functional form of relation do you expect. Present the corresponding scatterplots and comment on the results.

3. Run linear regression model of *Grades* on *StHours*, *SEScore*, *SleepHours*, *Attendance*. Then run the linear regression of $\log(Grades)$ on *StHours*, *SEScore*, *SleepHours*, *Attendance*. Interpret the estimated values of the coefficients for both models.

4. Then pick one or several independent variables and run the same models as in the previous part but including the logarithms of picked variables to the model. Interpret estimated values of the coefficients for new models.

5. What does R^2 show? What is the difference between R^2 and R^2 adjusted? Can we compare the models considered in the previous part using R^2 and/or R^2 adjusted? Explain.

Problem 3. Consider the dataset *Salary_Data.csv*. It includes five variables: age, experience, job role, and education level and salary

1. Discuss what data structure do we have. Classify the features. Analyze the variable *Salary*. Compute the corresponding descriptive statistics and interpret the results.

2. Compute Pearson and Spearman correlation coefficients for *Salary* vs. *age* and *experience*. Comment on the results. What functional form of relations do you expect. Present the corresponding scatterplots and comment on the results.

3. Run linear regression model of *Salary* on *age* and *experience*. Then run the linear regression of $\log(Salary)$ on *age* and *experience*. Interpret the estimated values of the coefficients for both models.

4. Then run the models from the previous part including the logarithms of *age* and/or *experience* to the model. Interpret estimated values of the coefficients for new models.

5. What does R^2 show? What is the difference between R^2 and R^2 adjusted? Can we compare the models considered in the previous part using R^2 and/or R^2 adjusted? Explain.

Problem 4 (Ch2. P6. Jeffrey M.Wooldridge. Introductory Econometrics)**.** Using data from 1988 for houses sold in Andover, Massachusetts, from Kiel and McClain (1995), the following equation relates housing price (price) to the distance from a recently built garbage incinerator (dist):

$$\widehat{\log price} = 9.40 + 0.312 log(dist)$$

$$n = 135,\ R^2 = 0.162$$

(i) Interpret the coefficient on log(dist). Is the sign of this estimate what you expect it to be?

(ii) Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of price with respect to dist? (Think about the city's decision on where to put the incinerator.)

(iii) What other factors about a house affect its price? Might these be correlated with distance from the incinerator?

Problem 5 (Exploring further 3.4. Jeffrey M.Wooldridge. Introductory Econometrics)**.** Suppose you postulate a model explaining final exam score in terms of class attendance. Thus, the dependent variable is final exam score, and the key explanatory variable is number of classes attended. To control for student abilities and efforts outside the classroom, you include among the explanatory variables cumulative GPA, SAT score, and measures of high school performance. Someone says, "You cannot hope to learn anything from this exercise because cumulative GPA, SAT score, and high school performance are likely to be highly collinear." What should be your response?

Problem 6 (Ch3. Pr5. Jeffrey M.Wooldridge. Introductory Econometrics)**.** In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.

(i) In the model

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u,$$

does it make sense to hold sleep, work, and leisure fixed, while changing study?

(ii) Explain why this model violates Assumption A3 (No perfect collinearity).

(iii) How could you reformulate the model so that its parameters have a useful interpretation and it satisfies Assumption A3?

Problem 7. Suppose you are hired by a fitness center chain to analyze the feasibility of adding evening yoga and pilates classes (7-9 PM) to their schedule. The company conducted a survey asking current members whether they would attend these new evening classes. You have data on 8,000 individuals collected over four months, including: age, sex, employment status, current membership type (basic or premium), frequency of gym visits per week, preferred workout time (morning, afternoon, evening), and whether they would be interested in attending evening yoga/pilates classes (yes/no). When building your model, you observe that 75% of respondents visit the gym primarily in morning hours (6-10 AM). Your colleague argues that your model will be biased and won't represent the preferences of the overall membership base. Is your colleague's criticism valid? Explain your reasoning.

Problem 8. Suppose you are hired by a tech company to study the factors that determine whether candidates who receive job offers actually accept them. You have a large random sample of candidates who were extended offers last year. Your dataset includes: whether each candidate accepted the offer, previous work experience, salary offered, benefits package details, location of the position, candidate's current city, and education level. A colleague comments, "Any analysis of this data will produce biased results because it's not a random sample of all job seekers in the tech industry, but only those who applied and were selected by this specific company." What is your assessment of this criticism?

Problem 9. A researcher wants to know if studying abroad improves language skills. They survey 200 students who studied abroad last year and ask them to self-rate their language improvement on a scale of 1-10.

1. Identify all the problems with this design.
2. Redesign the study to address these issues.
3. Specify data structure, variables, and sampling approach.
4. Explain what causation vs correlation issues remain.