Introductory Econometrics
Home Assignment 1
Suggested Solution

by Hieu Nguyen

Fall 2024

Solution of the assignment is to be delivered electronically by **DATE** 23:59:59 the latest. Late submissions will not be accepted, resulting in zero points.

Form teams of two people, please. Only one team member is supposed to submit the solution with both team members' names and email addresses on the first page of the document. Teams are required to work independently, and any form of plagiarism will be treated accordingly. Please understand that the main advantage of teamwork is the synergy from solving the problems together and the possibility to share and discuss your econometric knowledge with your teammate. It is not about a pure division of tasks. So, please, do cooperate and make sure you both understand all solutions completely.

The text itself can be written in any software of your choice (MS Word, LaTeX, Pages etc.), but the .pdf format [5 MB max, .xls(x) can be attached in .zip] of the final document is required.

Please, name the file **Surname1_Surname2_HA01.pdf.**

In your report, please, be clear and reasonably concise, but do explain all essential steps (e.g., important matrices) of your solution/reasoning. Keep in mind that not only the correctness of your answers and interpretations is assessed, but also the text-editing quality is an integral part of your output.

Fingers crossed!
Hieu Nguyen

Problem 1: Test scores

(2 points, 0.5pt each)
A nationwide test score has a mean of 63 points and a variance of 121.

1. Convert the following raw scores to standardized Z values: 52, 91.
2. What raw scores correspond to standardized values $Z = 2$ and $Z = -1.5$?
3. Assuming that the test score is normally distributed, what is the probability that a randomly selected individual who participated in the test has obtained a test score higher or equal to 41?
4. Assuming normality again, what is the probability that a randomly selected individual has obtained a test score between 55 and 75?

Solution:

We employ formulas from the lecture #1 slides for the standardization of a random variable, for the probability computational rule, and the definition of the CDF:

$$X \sim N(\mu, \sigma^2) \implies Z = \frac{X - \mu}{\sigma} \sim N(0, 1),$$
$$P(X > x) = 1 - P(X \leq x).$$

We then compute:

1. $$Z_{52} = \frac{52 - \mu}{\sigma} = \frac{52 - 63}{\sqrt{121}} = \frac{-11}{11} = -1$$
   (note that the variance $\sigma^2 = 121$ so the standard deviation $\sigma = 11$); similarly:
   $$Z_{91} = \frac{91 - 63}{11} \approx 2.55.$$
2. $$Z = 2 = \frac{X - \mu}{\sigma} = \frac{X - 63}{11} \implies X = 85;$$
   similarly:
   $$Z = -1.5 \implies X = 46.5.$$
3. $$P(55 \leq X \leq 75) = P(X \leq 75) - P(X \leq 55) =$$
   $$P\left(\frac{X - 63}{11} \leq \frac{75 - 63}{11}\right) - P\left(\frac{X - 63}{11} \leq \frac{55 - 63}{11}\right)$$
   $$= P(Z_{75} \leq 1.09) - P(Z_{55} \leq -0.73) =$$
   $$F_Z(1.09) - F_Z(-0.73) \approx 0.8621 - 0.2327 = 62.9\%$$

Problem 2: Modeling demand for beer consumption

(8 points)

Let us consider a simple regression model to explain the demand for beer. From the theory of consumer choice in microeconomics, we know that the demand for goods also depends on income. We will thus focus on this trivialized linear relationship. The data file `HomeAssignment_01_Problem2data.xlsx` contains a data sample of 30 observations of annual beer consumption (in liters) and annual income (in USD thousands) collected from randomly selected households.

Answer the following questions. Make sure you show all the matrices you construct and compute in your solution.

1. (1 pt) Formulate the econometric model. Using the OLS formula, estimate the intercept and slope parameters. Show the estimated model equation.
2. (1 pt) Interpret the meaning of the estimated coefficient of income. Does the direction of the income effect follow your economic intuition? In case it does not, provide a possible explanation(s).
3. (1 pt) Does the estimated intercept make sense in this situation? If yes, provide your economic interpretation. If not, explain why
4. (1 pt) Find and list the model's estimated/fitted values and residuals. Do the sum of the residuals up to zero?
5. (0.5 pt) Predict the beer consumption for households with an annual income of USD 60,000 and with USD 30,000.
6. (1pt) Consider carefully the Classical Assumptions step-by-step. Which of them are likely to be violated? Explain your reasoning properly.
7. (1 pt) What are the consequences in case some specific Classical Assumptions are violated? Think mainly about OLS properties (unbiasedness, consistency, efficiency).
8. (1 pt) Comment on the overall results of your analysis. Does the model suggest a realistic relationship between beer consumption and households' income?

Attached: `HomeAssignment_01_Problem2data.xlsx`

Solution:

1. We are asked to analyze the influence of income on beer consumption. Since beer can be considered the normal good the demand for which reflects a direct relationship with a consumer's income, we might expect a

positive relation:

$$\text{beer consumption} = f(\text{income}).$$

This relationship can be presented in a simple linear regression model form:

$$\text{beer consumption} = \beta_0 + \beta_1 \text{income} + \epsilon,$$

where we expect $\beta_1 > 0$.

Let us rewrite the model and the data in matrix form:

$$y = X\beta + \epsilon,$$

where:

$$y = \begin{pmatrix} 81.7 \\ 56.9 \\ 49.9 \\ 64.1 \\ 65.4 \\ 51.7 \\ 64.1 \\ 58.1 \\ 46.3 \\ 61.7 \\ 65.3 \\ 57.8 \\ 63.5 \\ 50 \\ 65.9 \\ 46.8 \\ 48.3 \\ 55.6 \\ 53.8 \\ 47.9 \\ 57 \\ 51.6 \\ 51.6 \\ 54.2 \\ 57.7 \\ 51.7 \\ 44.3 \\ 55.9 \\ 52.1 \\ 52.5 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 35.1 \\ 1 & 36.6 \\ 1 & 51.6 \\ 1 & 35.5 \\ 1 & 37.2 \\ 1 & 48.4 \\ 1 & 37.2 \\ 1 & 37.6 \\ 1 & 48.4 \\ 1 & 38.2 \\ 1 & 39.4 \\ 1 & 38.7 \\ 1 & 40.0 \\ 1 & 46.7 \\ 1 & 40.5 \\ 1 & 48.8 \\ 1 & 40.4 \\ 1 & 41.1 \\ 1 & 48.1 \\ 1 & 41.1 \\ 1 & 42.5 \\ 1 & 46.7 \\ 1 & 42.4 \\ 1 & 43.4 \\ 1 & 47.3 \\ 1 & 43.9 \\ 1 & 45.9 \\ 1 & 44.5 \\ 1 & 46.0 \\ 1 & 44.8 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

To find the OLS estimates, we compute the following matrices:

$$X'X = \begin{pmatrix} 30 & 1278 \\ 1278 & 55040.06 \end{pmatrix},$$

$$(X'X)^{-1} = \begin{pmatrix} 3.0718 & -0.0713 \\ -0.0713 & 0.0017 \end{pmatrix},$$

$$X'y = \begin{pmatrix} 1683.40 \\ 70973.82 \end{pmatrix},$$

which gives us the following:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 108.82 \\ -1.24 \end{pmatrix},$$

and thus the estimated model is:

$$\widehat{\text{beer consumption}} = 108.82 - 1.24\text{income}.$$

Please note that the different rounding might lead to slightly different numerical results in the following questions (which is not considered a mistake, of course).

2. Interpretation: one more USD thousand of annual income is associated with a decrease in beer consumption by 1.24 liter. The negative direction of the effect surprisingly does not follow our basic economic intuition outlined above. The are two main possible explanations. First, maybe our first impression above was not entirely theoretically correct, and beer is instead an example of an inferior good: with increasing income, people might reduce beer consumption (maybe for health reasons and awareness of the negative aspects of alcoholic beverages or because, on the other hand, they switch to higher quality and more expensive alcoholic drinks). Second, since our model is highly trivialized, the estimated effect can be considerably biased because of incorrect model specifications: other important right-hand side (RHS) variables likely correlated with income are missing/omitted in the equation (price, price of substitutes). If the bias is strong, it can even switch the direction of the estimated effect. Also, the underlying consumption function may be nonlinear in variables, while we estimate a purely linear consumption function. All these effects end up in the stochastic error term and might (very likely in this case) lead to a violation of the Classical Assumption of zero conditional mean.

3. Rather not. One can simply see that the cloud of data is very far away from zero and even the smallest annual incomes exceed 35 thousands. Zero income is thus rather a theoretical extreme far form reality (a potential idea of homeless households tending to consume large volumes of beer makes some sense but these data are likely unavailable due to the nonresponse bias). Potential additional impacts captured by the intercept hinder its interpretation oven more: it also absorbs a possible nonzero mean of the error term (see random sampling assumption) and a potential constant impact of any specification errors (e.g., omitted explanatory variables, will learn later)

4. The fitted values and residuals (e_i = beer consumption$_i$−beer $\widehat{\text{consumption}}_i$) are:

$$\text{beer } \widehat{\text{consumption}} = \begin{pmatrix} 65.4 \\ 63.5 \\ 45.0 \\ 64.9 \\ 62.8 \\ 48.9 \\ 62.8 \\ 62.3 \\ 48.9 \\ 61.6 \\ 60.1 \\ 60.9 \\ 59.3 \\ 51.0 \\ 58.7 \\ 48.4 \\ 58.8 \\ 58.0 \\ 49.3 \\ 58.0 \\ 56.2 \\ 51.0 \\ 56.4 \\ 55.1 \\ 50.3 \\ 54.5 \\ 52.0 \\ 53.8 \\ 51.9 \\ 53.4 \end{pmatrix}, \quad e = \begin{pmatrix} 16.3 \\ -6.6 \\ 4.9 \\ -0.8 \\ 2.6 \\ 2.8 \\ 1.3 \\ -4.2 \\ -2.6 \\ 0.1 \\ 5.2 \\ -3.1 \\ 4.2 \\ -1.0 \\ 7.2 \\ -1.6 \\ -10.5 \\ -2.4 \\ 4.5 \\ -10.1 \\ 0.8 \\ 0.6 \\ -4.8 \\ -0.9 \\ 7.4 \\ -2.8 \\ -7.7 \\ 2.1 \\ 0.2 \\ -0.9 \end{pmatrix}$$

The residuals indeed sum up 0.00.

5. Predictions:

$$\text{beer } \widehat{\text{consumption}}_{\text{income=30}} = 108.82 - 1.24 \cdot 30 \approx 71.6,$$

$$\text{beer } \widehat{\text{consumption}}_{\text{income=60}} = 108.82 - 1.24 \cdot 60 \approx 34.4.$$

6. • **Linearity:** From the theory of consumer choice in microeconomics, we also know that the demand for goods does not only depend on income but also on price and prices of other goods in the economy—particularly substitutes (wine, liquors, etc.) and possibly also

complements (water served in restaurants, non-alcoholic beverages, etc.). being most likely violated because the model is too simple, i.e., not correctly specified in terms of RHS explanatory variables. Also, the functional form linear in variables might not be completely correct (but this will be discussed in other lectures later).

- **Random sampling:** Not violated because the sample is collected from randomly selected households.
- **Zero conditional mean:** Most likely violated as essential RHS variables omitted from the equation (price, price of substitutes) are most likely correlated to the overall income level of the population. All these effects end up in the stochastic error term, which leads to a correlation between the error and the income variable. Intuitively, the OLS estimator then incorrectly assigns to the included explanatory variables parts of the effects of the omitted variables (to the extent of how strongly they are correlated).
- **Homoskedasticity:** Perhaps violated since the differences in behaviors of high-income and low-income consumers.
- **No perfect collinearity:** Cannot be violated because we only have one explanatory variable.
- **Normality of the error term:** The error term can be considered a cumulation of many additional influences that are not captured in the model, including only the income variable. Thus, this assumption is likely not violated.

7. Based on the last part, the violation of zero conditional mean leads to a biased and inconsistent OLS estimator. On the other hand, as we do not observe an indication of a violation of homoskedasticity, efficiency of the OLS estimator does not seem affected.

8. The analysis suggests a negative impact of increasing annual income on the beer consumption among households and can be expressed by the estimated model equation:

$$\widehat{\text{beer consumption}} = 108.82 - 1.24\text{income}.$$

This result goes against our basic economic intuition. Still, it can be explained by the potential inferiority of beer or as a result of the biased and inconsistent OLS estimator of this relationship because our model is highly trivialized. Especially other important RHS variables are missing from the equation, such as the price or price of substitutes, that are most likely correlated with the income variable. This most likely leads to a violation of zero conditional mean.