NIFB22001U Introduction to Econometrics Exam 2022/2023

Abdullah Faqih Al Mubarok

August 25, 2023

Part 1: The Returns to Education

1. Consider the following model:

$$wage = \alpha + \beta educ + u \tag{1}$$

Which assumptions are required in order for β to be a consistent and unbiased estimator of the causal effect of education on hourly wages? Explain each assumption in detail.

Answer:

To make sure unbiasedness ($E(\hat{\beta}) = \beta$) and consistency, we have to make sure that model (1) obey SLR.1 - SLR.4 as stated on Wooldridge (2018) which are described as follow:

(a) SLR.1: The parameter (β) should be linear which is satisfied by model (1).

(b) SLR.2: Every individual on the wage data was randomly sampled from the population. This means that we assume the data collection method was not affected by sampling problems.

(c) SLR.3: The value of the year of education ($educ$) should be vary within samples. Although it is subject to sampling variation, we can calculate that the $var(educ) = 3.36$ which satisfies this assumption.

(d) SLR.4: The expected value of u is zero given any value of $educ$ ($E(u|educ) = 0$). This means that $educ$ is assumed to be not correlated with other unobserved factors, u.

2. Estimate model (1) and interpret your results. Do the assumptions you presented in the previous question seem reasonable? Discuss.

Answer:

From the table 1, we can see that the estimation result $\hat{\beta}$ is statistically significant even at 1% level. This means that we can **confidently reject**, even at 1% significant level, the null hypothesis ($H_0 : \beta = 0$). In addition, one year additional of

Table 1: Estimation result of model (1)

	Dependent variable:
	wage
education	1.068***
	(0.029)
Constant	1.734***
	(0.357)
Observations	1,000
R^2	0.579
Adjusted R^2	0.578
Residual Std. Error	1.672 (df = 998)
F Statistic	1,370.163*** (df = 1; 998)
Note:	*p<0.1; **p<0.05; ***p<0.01

education corresponds to \$1.068 increase in hourly wage. It is expected that more years of education leads to higher wage since education can give more opportunities to higher paying jobs which require specialization e.g. engineer, and doctor. In addition, 57.9% variation of *wage* could be explained by *educ*, which is pretty good.

However, not all of the assumptions that we have on point (1) might be reasonable. First of all, since we do not have any details about how the data was collected, it might be that the observations were not randomly selected (SLR.2 violation), i.e. it suffers from self-selection problem where the high earners tend to not report their wage due to tax avoidance. In addition, there might be several other unobserved factors which correlated to *educ* (SLR.4 violation).

3. Why might model (1) suffer from endogeneity? Provide a few examples.

   Answer:

   It is important to consider the presence of unobserved factors that correlated with *educ* such as IQ and *parents_income*. Individuals with high IQ exhibit a propensity for attaining advanced levels of education, either due to their inherent capability to do so or their personal desire to pursue such educational achievements. Furthermore, high *parents_income* can facilitate individuals in their pursuit of higher education by alleviating concerns related to financial constraints. In addition, *parents_income* can also affect how well the nutrition that individuals got which is also correlated with the IQ development (Nyaradi et al. (2013)). Those two unobserved factors can be the reason of violation of SLR.4, which in turns make model(1) suffer from endogeneity.

4. What makes a good instrumental variable? Explain how the number of public

libraries within a 5-km radius of the individual's childhood residence, library, could be a valid instrument for educ.

Answer:

To be a good IV, an instrumental variable z must satisfies two assumptions:

$$cov(z, x) \neq 0 \quad \text{and} \quad cov(z, u) = 0$$

For the prior assumption, the number of library seems to be correlated to the *education* since the higher number of libraries means easier accessibility to books, hence giving more motivation for children to pursue higher education. It can also be checked, with subject to sampling variation, that the $cov(library, educ) = 3.89$. In addition, the *library* could be exogenic if it does not depend on other factors, u, such as the pupolation density (more people, more libraries), and social-economic status of the neighborhood (well-funded area, more libraries).

5. Implement a two-stage least squares (2SLS) regression to estimate the supposed causal effect of years of education on hourly wages. Interpret your results in comparison to model (1).

Answer:

Table 2: Estimation result of model(1) by OLS and IV estimators

	Dependent variable:	
	wage	
	OLS	*instrumental variable*
	(1)	(2)
education	1.068***	0.773***
	(0.029)	(0.040)
Constant	1.734***	5.339***
	(0.357)	(0.492)
Observations	1,000	1,000
R^2	0.579	0.535
Adjusted R^2	0.578	0.534
Residual Std. Error (df = 998)	1.672	1.758
F Statistic	1,370.163*** (df = 1; 998)	
Note:		*p<0.1; **p<0.05; ***p<0.01

from table 2, we can see that now the estimated coefficient $\hat{\beta}$ using IV estimator is 0.733 which is lesser than the estimation we have by OLS estimator. This means that previous estimation by OLS seems to be overestimated due to exclusion of *libraries* bias.

6. Discuss whether the assumptions needed for library to be a valid instrument for wage seem reasonable.

   Answer:

   As stated on question 5, we assume that the number of libraries is not correlated with several other factors such as population density, and social-economic status of the children's neighborhood. This may imply that the libraries were randomly built, or they were built irrespective to those two factors. In fact, it is most likely not. Hence, the assumptions that we imposed earlier seem not reasonable in reality.

Part 2: Global Food Consumption Patterns

2.1 Data description and preliminary analysis

1. Describe the supplied data. This description should include a table of descriptive statistics for all variables in the data set.

   Answer:

Table 3: Summary of descriptive statistic of examdata2023 02.RData

Statistic	N	Mean	St. Dev.	Min	Median	Max
AgLandShare	159	39.51	20.95	2.63	41.00	85.64
AgGdpShare	159	11.27	10.71	0.10	7.70	59.49
ArableLandShare	159	15.62	14.08	0.24	11.24	61.46
GdpPerCapita	159	21,339.78	21,386.52	750.31	14,033.98	118,961.50
GiniCoef	19	39.30	8.71	24.40	40.20	54.20
HCI	159	0.56	0.14	0.29	0.57	0.81
LifeExp	159	72.05	7.53	52.78	72.77	85.39
PopYoung	159	27.18	10.65	11.92	25.64	48.95
PopMiddle	159	63.34	6.18	48.62	64.59	83.41
PopOld	159	9.49	6.79	1.26	7.01	29.58
PopFemale	159	49.96	3.25	27.35	50.38	54.89
PopMale	159	50.04	3.25	45.11	49.62	72.65
PopRural	159	41.16	22.91	0.00	41.52	86.66
PopUrban	159	58.84	22.91	13.35	58.48	100.00
Year	159	2,018.00	0.00	2,018	2,018	2,018
CalTotal	140	2,918.32	450.91	1,776	2,913.5	3,871
CalVegetal	140	2,340.12	276.73	1,652	2,336	3,089
CalAnimal	140	578.16	356.63	38	532.5	1,638
PropCalVegetal	140	0.81	0.10	0.55	0.81	0.98
PropCalAnimal	140	0.19	0.10	0.02	0.19	0.45

   The provided dataset comprises cross-sectional data encompassing 159 countries during the year 2018. Moreover, upon examination of table 3, it becomes evident

that a significant proportion of countries lack Gini coefficients, with only 19 countries possessing this measure. Furthermore, the variables related to consumption, namely CalTotal, CalVegetal, CalAnimal, PropCalVegetal, and PropCalAnimal, also exhibit missing values, albeit with only 19 instances of missing data.

In addition, when examining the summary of *GdpPerCapita*, it becomes evident that there exists a significant disparity among countries, with a standard deviation of $21,386.5. While G*GdpPerCapita* is not a direct indicator of income per capita, it can serve as a useful proxy that potentially reflects the purchasing power of individuals in a country. Consequently, it may have an impact on consumption patterns. Moreover, when examining the relative consumption of vegetable and animal products, a discernible disparity becomes apparent. The median proportion of animal products in total consumption is lower than that of vegetables.

2. The supplied data contains missing data points. Does this affect the representativeness of the sample?

   Answer:

   No, it does not. The data remains representative. The distribution of countries with missing *PropCalAnimal* is observed to be widespread across various continents and socio-economic contexts. There exist nations characterized by varying levels of income disparity, with examples including Qatar and Congo representing high-income and low-income countries respectively. Similarly, there are countries exhibiting divergent levels of human capital index (HCI), with Hong Kong and Tanzania serving as high HCI and low HCI nations respectively. Moreover, it should be noted that approximately 88% of countries lack the Gini coefficient values. Hence, it is necessary to not use this variable since it limits our analysis.

3. How may a nonrandom sample affect any of our OLS estimates? Discuss based on the missing observations you identify. Finally, you should justify whether the missing observations are missing completely at random, missing at random, or missing not at random.

   Answer:

   The affect of nonrandom sample into the OLS estimation result depends whether it is missing completely at random (MCAR), missing at random (MAR) or missing not at random. The latter could lead to bias and inconscistent estimator due to correlation between the sample selection with unobserved factor, u. In the case of our data, the missing data points of *PropCalAnimal* seem to be not correlated with other factors and completely random (MCAR) as explained on (2). Thus, we can conclude that we do not have sampling issue due to missing data points.

Regardless of your answer to the previous two questions, you should proceed with removing observations with missing values in the variable *PropCalAnimal* from the data set.

4. Assume that the consumption of vegetal and animal products are normal goods. Furthermore, suppose that the consumption of animal products is a luxury good.

Oppositely, let us assume that the consumption of vegetal products is a necessity. How would you expect the proportion of calories from animal and vegetal products to be related to the GDP per capita? Explain your reasoning.

Answer:

Firstly, lets assume that GDP per capita can serve as an accurate measure of a country's purchasing power. In this context, if vegetable consumption is a necessity, individuals in countries with higher GDP per capita may indeed have the financial ability to afford greater quantities of vegetables. However, it is unlikely that the proportion of vegetable consumption will experience a substantial increase. In nations with low GDP per capita, there is a tendency of higher consumption of plant-based products since they are cheaper compared to animal-based products. Conversely, in countries characterized by low GDP per capita, the consumption of animal products tends to be limited due to it is expensive. In contrast, in higher GDP per capita countries, there is a relatively greater proportion of animal product consumption, as individuals in these countries has greater purchasing power.

5. Investigate the relationship between the proportion of calories from animal and vegetal products against the GDP per capita graphically. What do you find? Explain whether people seem to be substituting vegetal products with animal products as their income increases.

Answer:

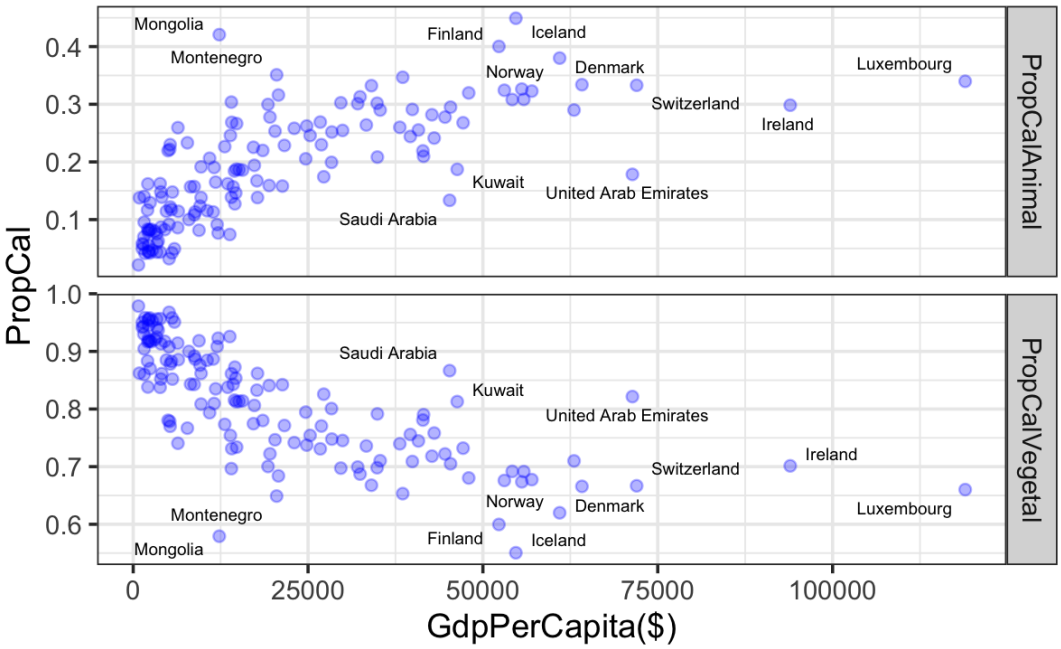

Figure 1: Proportion of Animal and Vegetable Products respective to GDP Per Capita

As shown on figure 1, we could see that as *GdpPerCapita* increases, the proportion of animal consumption tends to be larger. on the contrary, the proportion of vegetable consumption has tendency to decrease as the *GdpPerCapita* go up. Those two proportion seems to substitute each other and there is a clear pattern: that

countries with higher income levels tend to replace their vegetable-based products with animal-based products.

6. Another hypothesis is that due to pro-environmental and pro-climate preferences, people in countries with high human capital at a certain point decide to consume fewer animal products in favor of vegetal products. Given the available data, how would you examine this hypothesis in further detail?

   Answer:

   Our interest is to know the ceteris paribus effect of HCI on $PropCalAnimal$. At first, we should build an economic model of it. Given our hypothesis that the consumption behavior undergoes a change "at a certain point", it is plausible that the relationship between $PropCalAnimal$ and HCI is in a quadratic fashion:

   $$PropCalAnimal = \beta_0 + \beta_1 HCI + \beta_2 HCI^2 + u$$

   In addition, since at some point, the increment in HCI results in fewer animal products, we should expect that the quadratic function is concave down ($\beta_2 < 0$) and the inflection point of HCI, in the context of the quadratic function is symmetrical point, should be positive ($\beta_1 > 0$).

   if we use OLS as the estimator, to make unbiased and consistent estimate result, we need to obey several assumptions. First, Every parameters β_k, should be linear, which is, so far, already satisfied by the above quadratic model. Second, every nation on the data should be randomly sampled. It means that there is no sampling bias which wee already checked on (3). Third, there is no perfect linear relationship between explanatory variables, which is so far satisfied. Lastly, all of the expalanatory variable(s) should not be correlated with unobserved factors, u. In the above population model, HCI still have association with u since the development of an individual is highly correlated with the wealth that they have.

   For this reason, $GdpPerCapita$ might be positively correlated with HCI. In addition, there might be also positive correlation between $PopUrban$ and $GdpPercapita$. It is due to the increment of $GdpPercapita$ might be caused by higher paying jobs availability. Those jobs are mostly in the urban area, thus increasing the $PopUrban$. Moreover, the $GdpPerCapita$ might have negative correlation with the $AgLandShare$ and $AgGdpShare$ since higher GDP per capital countries tends to be more industrialized, hence reducing $AgLandShare$ and $AgGdpShare$. Finally, we add four variables into our previous population model:

   $$\begin{aligned} PropCalAnimal = \beta_0 + \beta_1 HCI + \beta_2 HCI^2 + \beta_3 GdpPercapita \\ + \beta_4 PopUrban + \beta_5 AgLandShare + \beta_6 AgGdpShare + u \end{aligned}$$

   Assuming finite variance of u, then we can inference our hypothesis by one-sided t-statistic with the following hypotheses:

   For β_1:

   $$\begin{aligned} H_0 &: \ \beta_1 = 0 \\ H_A &: \ \beta_1 > 0 \end{aligned}$$

For β_2:

$$H_0 : \beta_2 = 0$$
$$H_A : \beta_2 < 0$$

2.2 Regression analysis

1. Compare the distributions of PropCalAnimal and log(PropCalAnimal). Which of the two variables would you prefer to use in a regression analysis? Explain your reasoning.

   Answer:

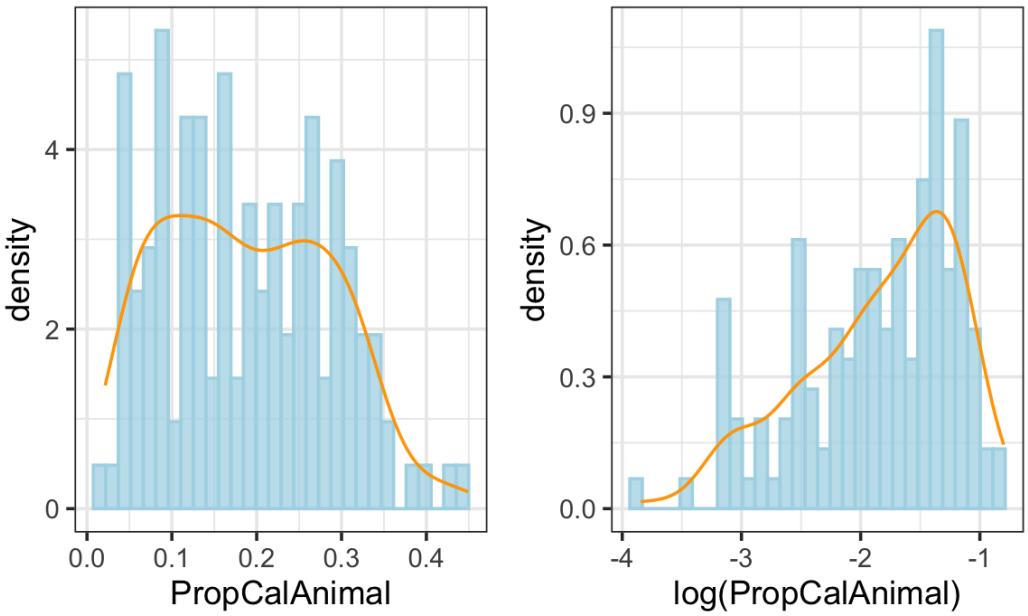

Figure 2: Density Plot of PropCalAnimal and log(PropCalAnimal)

The log(PropCalAnimal) is skewed to the left with one peak, as shown in Figure 2. Meanwhile, the PropCalAnimal distributions appear to have two peaks, which may represent two distributions or indicate that the PropCalAnimal might be affected non-lienarly by other variable(s). Therefore, based only on the distribution, log(PropCalAnimal) is preferable because it more closely resembles a normal distribution.

2. Estimate the following model using OLS

$$\begin{aligned} PropCalAnimal =& \beta_0 + \beta_1 AgLandShare + \beta_2 AgGDPShare \\ &+ \beta_3 ArableLandShare + \beta_4 GdpPerCapita \\ &+ \beta_5 HCI + \beta_6 LifeExp + \beta_7 PopMiddle \\ &+ \beta_8 PopOld + \beta_9 PopFemale + \beta_{10} PopUrban + u \end{aligned} \quad (2)$$

Do you find it likely that the Gauss-Markov assumptions apply to model (2)? Furthermore, in terms of model (2), what would this imply about the estimated parameters?

Answer:

To check whether the model (2) satisfies Gauss-Markov assumptions, we need to breakdown and check each of the all assumptions (MLR.1-MLR.5):

(a) MLR.1: The parameters (β_k) should be linear. It is already satisfied by model (2)

(b) MLR.2: We already discussed that the missing datapoints are missing completely at random (MCAR) on previous question. Therefore, we do not have any sampling problems.

(c) MLR.3: Based on the variables definition, it should be that the explanatory variables on model(2) do not have perfect collinearity. Moreover, subject to sample variance, we can check the correlation matrix of the explanatory variables:

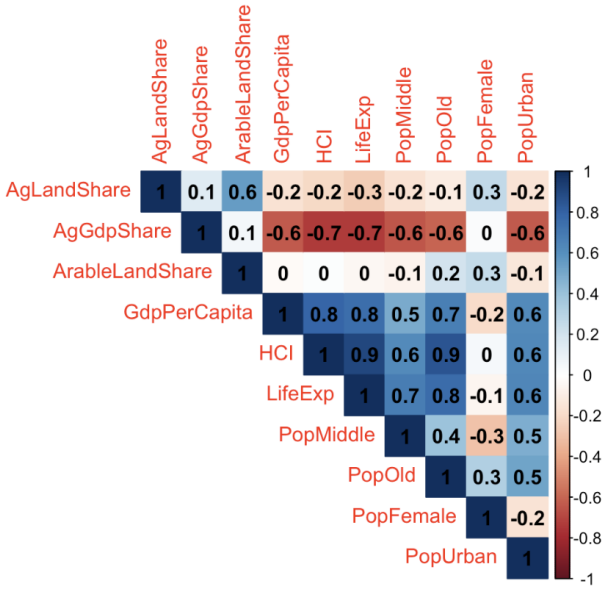

Figure 3: Correlation Matrix of the Explanatory Variables

Figure 3 validates our assupmtion of MLR.3.

(d) MLR.4: The zero conditional mean ($E(u|x_1, ..., x_k) = 0$) seems fulfilled since all of the explanatory variables on model(2) might not have correlation with other unobserved factors.

(e) MLR.5: We assume that the unobserved factors, u has a constant variant $Var(u|x_1, ..., x_k) = \sigma^2$ (homoskedasticity). To check this assumption, we can use Breusch-Pagan (BP) test. This test has null hypothesis:

$$H_0 : E(u^2|x_1, ..., x_k) = \sigma^2$$

by using lmtest package in R, we have $p - value = 0.6571$, which tells us that even at 10% significant level, we fail to reject the null hypothesis of homoskedasticity on model(2).

Thus, according to the checks above, the Gauss-Markov applies to model (2). Therefore, our estimation results of the coefficients of model (2) are unbiased and efficient, which means that on average they are centered around its true value ($E(\hat{\beta}_k) = \beta_k$) and they have the smallest variance. The estimation results by OLS estimator is shown on table 4.

Table 4: Estimation result of model(2) by OLS estimator

	Dependent variable:
	PropCalAnimal
AgLandShare	0.0000 (0.0003)
AgGdpShare	−0.0001 (0.001)
ArableLandShare	−0.001** (0.0004)
GdpPerCapita	0.0000** (0.0000)
HCI	0.39*** (0.10)
LifeExp	−0.002 (0.002)
PopMiddle	0.001 (0.001)
PopOld	0.003* (0.002)
PopFemale	0.005* (0.003)
PopUrban	0.0002 (0.0003)
Constant	−0.25 (0.18)
Observations	140
R^2	0.74
Adjusted R^2	0.72
Residual Std. Error	0.05 (df = 129)
F Statistic	36.66*** (df = 10; 129)
Note:	*p<0.1; **p<0.05; ***p<0.01

3. Re-estimate model (2) with log(PropCalAnimal) as the new dependent variable:

$$\begin{aligned} log(PropCalAnimal) =& \beta_0 + \beta_1 AgLandShare + \beta_2 AgGDPShare \\ &+ \beta_3 ArableLandShare + \beta_4 GdpPerCapita \\ &+ \beta_5 HCI + \beta_6 LifeExp + \beta_7 PopMiddle \\ &+ \beta_8 PopOld + \beta_9 PopFemale + \beta_{10} PopUrban + u \end{aligned} \quad (3)$$

How does the interpretation of the estimated parameters change when going from model (2) to model (3)? For instance, how would you interpret the estimated coefficient on AgLandShare and LifeExp in model (3) as compared to model (2)?

Answer:

Table 5: Estimation result of model (3) by OLS estimator

	Dependent variable:
	log(PropCalAnimal)
AgLandShare	−0.0003 (0.002)
AgGdpShare	−0.01* (0.005)
ArableLandShare	−0.01*** (0.003)
GdpPerCapita	0.0000 (0.0000)
HCI	1.96*** (0.69)
LifeExp	−0.01 (0.01)
PopMiddle	0.02*** (0.01)
PopOld	0.02** (0.01)
PopFemale	0.03* (0.02)
PopUrban	0.001 (0.002)
Constant	−5.01*** (1.20)
Observations	140
R^2	0.73
Adjusted R^2	0.71
Residual Std. Error	0.35 (df = 129)
F Statistic	34.53*** (df = 10; 129)
Note:	*p<0.1; **p<0.05; ***p<0.01

Since the PropCalAnimal is transformed into log(PropCalAnimal) in model (3), the interpretation of each parameters are changed. In model (3), one unit change of a explanatory variable, holding other explanatory variables fixed, is responsible to the percentage change of PropCalAnimal.

For comparison, in model (2) we interpret an unit change of $LifeExp$, holding other factors fixed, corresponds to -0.002 constant change of PropCalAnimal. However, in model (3), while holding fixed all other factors, an unit change of $LifeExp$ corresponds to $(exp(-0.01) - 1) = -0.00995 = -0.995\%$ constant rate change of PropCalAnimal.

4. Compare the distribution of residuals from model (2) and model (3). Which model do you prefer and why? Are there any issues related to taking the logarithm of the dependent variable in model (3)?

Answer:

Figure 4 depicts the distribution of the estimated residuals for models (2) and (3). The distributions of the two appear to resemble a normal distribution. However, the residual distribution of model (2) seems to have less variance than model (3). This indicates that our estimations of (β_k) are more concistent in model (2) (lower in variance). Therefore, model(2) is preferable compared to model(3)

In addition, there might be two issues on using log(PropCalAnimal) as dependent variable. First, it might suffer from numerical stability if the value of PropCalAnimal is very small. Second, since the PropCalAnimal is already in percentage, it is really hard to interpret the "percentage change of percentage".

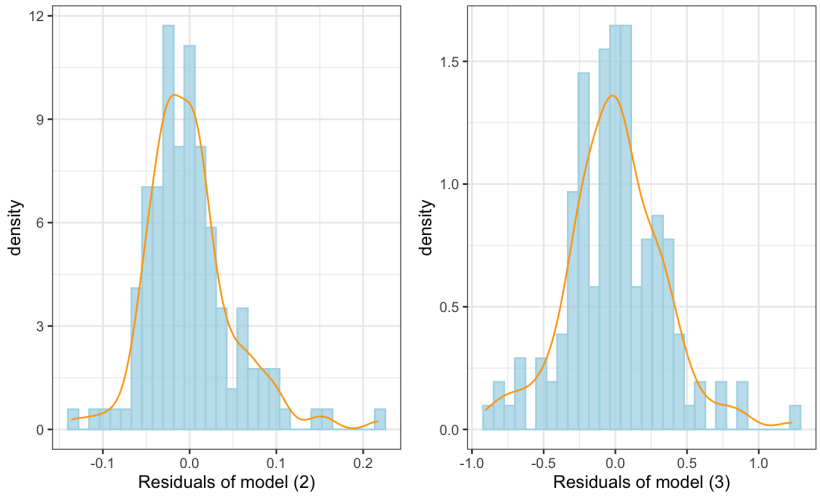

Figure 4: The distribution of $\hat{u}$ of model(2) and model(3)

5. Carry out a test for conditional heteroskedasticity in model (2) and model (3). What do you conclude?

   Answer:

   To examine the heteroskedasticity on the two models, we use the Breusch-Pagan (BP) test. To do the test, we regress u^2 on all of the explanatory variables, β_k:

$$\begin{aligned} u^2 =& \delta_0 + \delta_1 AgLandShare + \delta_2 AgGDPShare \\ &+ \delta_3 ArableLandShare + \delta_4 GdpPerCapita \\ &+ \delta_5 HCI + \delta_6 LifeExp + \delta_7 PopMiddle \\ &+ \delta_8 PopOld + \delta_9 PopFemale + \delta_{10} PopUrban + v \end{aligned}$$

   With the null hypothesis:

$$H_0 : \delta_0 = \delta_1 = ... = \delta_{10} = 0$$

   Hence, if we fail to reject the null, all of the explanatory variables do not correlated with the unobserved factors, u. Thus, we have homoskedasticity on our model.

   To run the test, we run it with lmtest package in R which gives us p-value of 0.6571 and 0.01602 for model(2) and model(3) respectively. Therefore, we can conclude that for model(2) we fail to reject the null even at 10% level, hence we have homoskedasticity on that model. However, in model(3) we can reject the null even at 1% level; thus, model(3) exhibits heteroskedasticity.

6. Using model (2), test for the joint significance of AgLandShare and AgGdpShare. State the relevant null- and alternative hypotheses carefully. Why might carrying out the same test in model (3) be problematic?

   Answer:

To test the joint significance of AgLandShare and AgGdpShare, we can use restricted F-test. Since we can use model(2) as our unrestricted model, we are left with the construction of the restricted model:

$$\begin{aligned} PropCalAnimal_{restricted} =& \beta_0 + \beta_1 ArableLandShare + \beta_2 GdpPerCapita \\ &+ \beta_3 HCI + \beta_4 LifeExp + \beta_5 PopMiddle \\ &+ \beta_6 PopOld + \beta_7 PopFemale + \beta_8 PopUrban + u \end{aligned}$$

Then, we can construct the hypotheses:

$$\begin{aligned} H_0 :\ & \beta_{AgLandShare} = \beta_{AgGdpShare} = 0 \\ H_A :\ & \beta_{AgLandShare} \neq 0 \text{ or } \beta_{AgGdpShare} \neq 0 \end{aligned}$$

Using R for the F-test, it gives p-value of 0.9739, which means that we fail to reject the null even at 10% level. Hence, we can conclude that the two are not jointly significant. The joint significant test result agrees with the individual t-test of AgLandShare and AgGdpShare as shown on table 4. Therefore, we can confidently remove the two variables from model (2).

In addition, carrying out the same test on model (3) would be problematic due to the fact that the AgGdpShare is statistically significant at the 10% level, as shown in table 5. Thus, it makes no sense to examine the joint significance test of AgLandShare and AgGdpShare on model(3).

7. Now, estimate the model:

$$\begin{aligned} PropCalAnimal =& \beta_0 + \beta_1 ArableLandShare + \beta_2 log(GdpPerCapita) \\ &+ \beta_3 HCI + \beta_4 LifeExp + \beta_5 PopMiddle \\ &+ \beta_6 PopOld + \beta_7 PopFemale + \beta_8 PopUrban + u \end{aligned} \tag{4}$$

In model (4), how do you interpret the estimated coefficient on log(GdpP erCapita)? What does this imply about the relationship between the proportion of calories from animal products and the GDP per capita?

Answer:

As a result of transforming GdpPerCapita to log(GdpPerCapita) in model (4), we now have level-log models between PropCalAnimal and GdpPerCapita. Based on the estimation result from table 6, hodling all other factors constant, 1% change in GdpPer-Capita corresponds to 0.0003 change in PropCalAnimal, or doubling GdpPerCapita corresponds to 0.03 change in PropCalAnimal.

8. Formally test the hypothesis that the proportion of calories from animal products is increasing until a certain level of human capital (HCI), at which point it decreases by using and eventually extending model (4). Remember to state the relevant null- and alternative hypotheses carefully. Discuss your findings. What do you conclude about the initial hypothesis from question five in the preliminary analysis?

Answer:

Table 6: Estimation result of model (4) by OLS estimator

	Dependent variable:
	PropCalAnimal
ArableLandShare	-0.001^{**} (0.0004)
log(GdpPerCapita)	0.03^{**} (0.01)
HCI	0.40^{***} (0.10)
LifeExp	-0.002 (0.002)
PopMiddle	-0.0002 (0.001)
PopOld	0.002 (0.002)
PopFemale	0.003 (0.002)
PopUrban	0.0000 (0.0003)
Constant	-0.30^{*} (0.17)
Observations	140
R^2	0.74
Adjusted R^2	0.72
Residual Std. Error	0.05 (df = 131)
F Statistic	46.07^{***} (df = 8; 131)
Note:	*p<0.1; **p<0.05; ***p<0.01

Since the hypothesis stated that the HCI increases PropCalAnimal until a "certain level", it might be that PropCalAnimal and HCI correlates in a quadratic form where we exepect that the coefficient of HCI^2 is negative (concave down) and the coefficent of HCI should be positive to make sure that the inflicted point is positive on the HCI axis. Thus, we can validate it using one-sided t-test by the following hypotheses:

For β_{HCI}:

$$H_0 : \beta_{HCI} = 0$$
$$H_A : \beta_{HCI} > 0$$

For β_{HCI^2}:

$$H_0 : \beta_{HCI^2} = 0$$
$$H_A : \beta_{HCI^2} < 0$$

Table 7 shows the estimated extended version of model (4) with HCI^2. To be statistically significant, $t_{\widehat{\beta}_{HCI}}$ should be greater than a critical value and $t_{\widehat{\beta}_{HCI^2}}$ should be lesser than a critical value. With significant level at 10% and $df = 140 - 9 - 1 = 130$, we have a critical value of $|c_{t_{130}}| = 1.29$.

To get the t-statistic we can use $t_{\widehat{\beta}} = \frac{\widehat{\beta}}{se(\widehat{\beta})}$. Thus, $t_{\widehat{\beta}_{HCI}} = 0.4/0.46 = 0.86$ and $t_{\widehat{\beta}_{HCI^2}} = 0.01/0.38 = 0.026$, subject to rounding error. To be statistically significant, $t_{\widehat{\beta}_{HCI}} > c_{t_{130}}$ and $t_{\widehat{\beta}_{HCI^2}} < -c_{t_{130}}$ which in fact they are not. Thus, we fail to reject the null for β_{HCI}, and we also can not reject the null for β_{HCI^2}.

In addition, if we look at the sign of HCI^2 it is not what we expected since after the inflection point the PropCalAnimal is increasing rather than decreasing due to

environmental awareness. Finally if we look at the marginal effect of HCI, holding other effect fixed, $\frac{\Delta PropCalAnmial}{\Delta HCI} = 0.02HCI + 0.4$ and search for the inflection point $\frac{\Delta PropCalAnmial}{\Delta HCI} = 0$, we have $HCI = -20$ which is impossible since HCI is ranged between 0 and 1.

Table 7: Estimation result of model (4) with quadratic HCI by OLS estimator

	Dependent variable:
	PropCalAnimal
ArableLandShare	-0.001^{**} (0.0004)
log(GdpPerCapita)	0.03^{**} (0.01)
HCI	0.40 (0.46)
I(HCI^2)	0.01 (0.38)
LifeExp	-0.002 (0.002)
PopMiddle	-0.0002 (0.002)
PopOld	0.002 (0.002)
PopFemale	0.003 (0.002)
PopUrban	0.0000 (0.0003)
Constant	-0.30^{*} (0.17)
Observations	140
R^2	0.74
Adjusted R^2	0.72
Residual Std. Error	0.05 (df = 130)
F Statistic	40.64^{***} (df = 9; 130)
Note:	*p<0.1; **p<0.05; ***p<0.01

In conclusion, our previous hypothesis, which stated that environmental awareness in countries with a high HCI could cause people to consume fewer animal products at a certain point, is not supported by the discussion above. It may be because the decision to avoid animal product consumptions due to environmental preferences is too complex to be captured by the model and the data availability.

2.3 Model selection

We begin the model selection with showing the comparison of model(2), model(3), and model(4) on the table 8. According to the table results, all of the three models have overall statistically significant regression even at 1% level, thus one or more independent variables of each model could help to explain their respective dependent variable, which is a good start.

Subsequently, it is observed that based on the BP test, both model (2) and model (4) do not exhibit statistical significance, even when considering a significance level of 10%. Therefore, we do not have sufficient evidence to reject the null hypothesis of the BP test for the two models. Consequently, it can be concluded that both model (2) and model (4) exhibit homoskedasticity. In contrast, the BP test result of model(3) is statistically significant even at 5% level. Therefore, we can reject the null of BP test of it, concluding the model(3) has heteroskedasticity. This implies that for the model(3), we need to be careful for omitting variables due to its statistical significant level from the t-test. Thus, using the robust-heteroskedasticity standrad-error is more preferable. However,

we can see that there is not any significant difference on robust-heterosekdasticity SE and standard SE of model(3) as shown on table 9. In addition, since the model(3) has heteroskedasticity, the the OLS estimator is no longer BLUE on it.

In addition, based on the general functional test with RESET for quadratic form, model(2) and model(4) are statistically significant at 10% level, while model(3) is not. Consequently, we can reject the null for model(2) and model(4), which means that there might be misssepcification of functional form of the two models. Meanwhile for model(3) we fail to reject the null which shows an evidence in favor of correct functional form.

Furthermore, The differences in the Adjusted R^2 among the three models do not differ that much, which means that the three models are roughly equivalent regarding how much their respective dependent variable variation is explained by their independent variables.

Model interpretation is also an important factor for model selection. At first, we need to consider how would we interpret the ceteris paribus effect of $GdpPerCapita$ since this variable might be very important representation of macroeconimic condition. In model(2) and model(3) it does not make sense to represent a change in a dollar since the range of $GdpPerCapita$ is very large ($118,211.2). This argument is also supported by the fact that the predicted coefficient of $GdpPerCapita$ on both two models are very small. Therefore, It is more prefarable to transform it into $log(GdpPerCapita)$.

As a consequence, based on the previously mentioned reasons, it can be argued that model (4) appears to be the most suitable for analyzing the correlation between the percentage of animal products consumptions and diverse macroeconomic and sociodemographic factors. In addition, we might also remove several variables from model(4) which are LifeExp, PopMiddle, PopOld, PopFemal, and PopUrban since they are all statistically insignificant even at 10% level.

Despite our chosen model(4) looks good in terms of explaining the variation of PropCalAnimal as well as easy to interpret, there are several limitations. First, this model heavily relies on log(GdpPerCapita) as macroeconomic indicators. Despite its usefulness to representing the average wealth, it does not show us how the economic within country is distributed. If the wealth is more evenly distributed, more people might be able to consume animal products which in turns increasing the PropCalAnimal. Thus, including the GiniIndex might be a good idea if it is available. Another limitation of the chosen model is it does not capture the correlation of PropCalAnimal with demographic factors. It might be due to the complexity of how those factors affecting consumption behaviours.

Table 8: Comparison of estimation results of model(2), model(3), and model(4) by OLS estimator

	Dependent variable:		
	PropCalAnimal	log(PropCalAnimal)	PropCalAnimal
	(2)	(3)	(4)
AgLandShare	0.00003	−0.0003	
	(0.0003)	(0.002)	
AgGdpShare	−0.0001	−0.008*	
	(0.001)	(0.005)	
ArableLandShare	−0.001**	−0.010***	−0.001**
	(0.0004)	(0.003)	(0.0004)
GdpPerCapita	0.00000**	0.00000	
	(0.00000)	(0.00000)	
log(GdpPerCapita)			0.026**
			(0.011)
HCI	0.388***	1.956***	0.405***
	(0.103)	(0.694)	(0.101)
LifeExp	−0.002	−0.013	−0.002
	(0.002)	(0.010)	(0.002)
PopMiddle	0.001	0.023***	−0.0002
	(0.001)	(0.008)	(0.001)
PopOld	0.003*	0.023**	0.002
	(0.002)	(0.011)	(0.002)
PopFemale	0.005*	0.029*	0.003
	(0.003)	(0.017)	(0.002)
PopUrban	0.0002	0.001	0.00004
	(0.0003)	(0.002)	(0.0003)
Constant	−0.252	−5.006***	−0.298*
	(0.179)	(1.204)	(0.166)
BP Test P-value	0.657	0.016	0.856
RESET (power=2) Test P-value	0.085	0.284	0.040
Observations	140	140	140
R^2	0.740	0.728	0.738
Adjusted R^2	0.720	0.707	0.722
Residual Std. Error	0.052 (df = 129)	0.351 (df = 129)	0.052 (df = 131)
F Statistic	36.655*** (df = 10; 129)	34.526*** (df = 10; 129)	46.072*** (df = 8; 131)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 9: Model (3) with standard and heteroskedasticy-robust standard error

	Dependent variable:	
	log(PropCalAnimal)	
	(standard SE)	(heteroskedasticity-robust SE)
AgLandShare	−0.0003	−0.0003
	(0.002)	(0.002)
AgGdpShare	−0.008*	−0.008*
	(0.005)	(0.005)
ArableLandShare	−0.010***	−0.010***
	(0.003)	(0.004)
GdpPerCapita	0.00000	0.00000
	(0.00000)	(0.00000)
HCI	1.956***	1.956***
	(0.694)	(0.661)
LifeExp	−0.013	−0.013
	(0.010)	(0.011)
PopMiddle	0.023***	0.023***
	(0.008)	(0.009)
PopOld	0.023**	0.023**
	(0.011)	(0.011)
PopFemale	0.029*	0.029**
	(0.017)	(0.012)
PopUrban	0.001	0.001
	(0.002)	(0.002)
Constant	−5.006***	−5.006***
	(1.204)	(1.122)
Observations	140	140
Note:		*p<0.1; **p<0.05; ***p<0.01

References

Nyaradi, Anett et al. (2013). "The role of nutrition in children's neurocognitive development, from pregnancy through childhood". In: *Frontiers in Human Neuroscience* 7. DOI: 10.3389/fnhum.2013.00097.

Wooldridge, Jeffrey M. (2018). *Introductory econometrics: A modern approach (seventh edition)*. Cengage.