Health Economics

Dr Pratap C Mohanty

Department of Humanities and Social Sciences,

Indian Institute of Technology Roorkee

Week – 11

Lecture 55- Econometric Modelling in Health

Welcome, friends, to our 11th-week lectures. As promised, we will have a dedicated lecture on Econometric Modelling in Health. So, before going into depth about this, I think a brief recap is required. We especially discussed issues of healthcare through data. We emphasised their sources, importance, micro and macro dimensions and indicators, and also we discussed about their challenges and limitations.

Here, understanding the data (as we mentioned) is the first step in conducting any econometric analysis. The next step is to specify the econometric/ regression model and appropriate estimation technique. Hence, in this lecture, we will introduce econometric modelling used by applied health economists. In the econometric modelling, we will discuss about linear regression techniques and non-linear regression techniques. We also briefly discuss research designs in causal inferences or causal analysis, where we discuss the RCT, known as Randomised Control Trial, and the quasi-natural experiments.

So, the pure form of experimental design is called RCT in social sciences. It is largely used by health professionals, who apply it in clinical trials. However, since our domain is a social science, we are focusing on social issues of healthcare, so we are focusing on randomised control trials for the explanation. Within randomised control trials, we will emphasise how difficult it is, how expensive it is, etc., and quasi- or natural experiments are also suggested. The last one to emphasise is estimation methods.

In the introduction, I wish to explain the econometric model that describes the key relationship under investigation. For example, suppose we have micro-data on a sample population, and we are interested in understanding the role of education in determining the differences in health across populations. In order to do so, we can think of a model summarising the relationship between education and the health of the i^{th} individual that will be presented as:

$$y_i = \alpha + \beta_1 x_{i1} + X_i'\gamma + \varepsilon_i$$

where,

y= dependent variable, the measure of health (e.g.- blood pressure)

x= primary independent variable, which can be proxied by either→ ***year of education or highest level of education or individual attended university*** *etc.*

X = vector of **other explanatory** variables associated with blood pressure such as ***age, sex, dietary quality, etc.***

ε= error term, captures all other determinants of blood pressure that are not included as explanatory variable

β_1= impact of education on blood pressure, controlling all others

So, this is presented in matrix form. You can refer to this form from the book of 'William Green' for econometrics. So, Y is our dependent variable here. When we say its relationship with education and health, we are actually referring to the measure of health, maybe blood pressure, maybe any other indicator, like any morbidity, etc. Whereas, the first small x_i refers to the primary factor, which we want to emphasise, i.e., the primary independent variable, which can be proxied by either 'year of education', or "highest level of education', or 'individual attended university', etc. Since our focus is on education and health, we must first emphasise these variables.

Other control variables are also required, which simultaneously impact healthcare and education. There are multiple relationships. Yes, we just say these (education on health), but that does not mean others are not impacted. These are also impacting. Hence, we need to cross-check how these also have some relationships, and these have some relationship. But 'over the time period' (that refers to the aspect of time-series), we are not discussing at this moment. The time-series component or the serial correlation, etc., we are not discussing at this moment. I am just giving a signal to those who have already read econometrics to some extent. Hence, all those relationships are going to be emphasised.

We also need to cross-check how these two are also related. In all forms, let us emphasise at this moment about X_i' (the capital X), that is all about the vector of other explanatory variables associated with blood pressure, such as age, sex, dietary quality etc. Since our purpose is to discuss $\hat{Y}$. So, the estimated value of Y is also dependent on the dietary quality, sex, age of the person, and other demographic features. Similarly, the model usually possesses error terms. The error term captures all other determinants of blood pressure that are not included in the explanatory variables. Here, β_1 is the coefficient, measuring the impact of education on blood pressure and controlling all other factors (keeping other factors constant).

We will discuss these relationships. Let me give you just a hint. When we are actually trying to check this, we need to check whether the independent factors are actually related or not. So, somewhere, you might have studied about the issue of multicollinearity, which has to be also cross-checked. So, we are saying this is the one (indicated in the lecture) where multicollinearity is shown. Then, whether the explanatory variables are actually linked to the error term or not? In that case, we are actually referring to the context of some issues of endogeneity, which means when this is linked (X and Y are considered to have some reverse causality), so we will just check how far X is dependent on Y, and Y is dependent on X. So, that aspect is going to be emphasised as well.

Then, we must also cross-check how far this follows a distribution. If it follows a normal distribution as per the Gauss-Markov theorem, which is written Gauss-Markov theorem (as per the standard assumptions of the basic or the linear regression). We will discuss everything systematically. At this moment, I have just tried my best to give you some directions. Let us unfold this discussion one by one. Starting with the linear regression. So, linearity, as you might have read, means it is linear with the parameter. So, the parameter should be linear.

Variables might be non-linear, but the parameter we are estimating must be linear. So, it should not be β_1^2 or β_1^3 etc., as that refers to non-linearity or non-linear models. The simplest model specification assumes a linear relationship between dependent and independent variables, but the relationship should be explained with their coefficients. The coefficient should be linear. The linear model relies on several assumptions. You need to refer to the Gujarati. Even the latest books can also be referred, like one by Green (as I already mentioned), for the equation to be written in matrix form.

The ordinary least square method is considered to be the most common regression estimator, where the least square (the square of the error term) is to be estimated at the very lowest level, so that whatever the projection we do, that means the error term is actually at a very low level. So, our projection is considered to be better. Hence, the least is emphasised. So, we aim to take the error distribution's first and second-order derivatives. You can just cross-check with Gujarati; you will get all those details.

I have already mentioned that with OLS, the relationship is considered ordinary in nature, meaning a simple relationship is expected. OLS interpretations are usually easy. For instance, in the earlier case that we have just cited (education and health example), if education is mentioned with a variable called 'years of education', and blood pressure, that is the proxy of the health indicator, is millimetres of the mercury (i.e., mmHg), i.e., measured in a standard unit of blood pressure. The coefficient β_1 represents the impact of unit change or unit increase in education, i.e., one extra year of schooling on blood pressure, keeping every other explanatory variable under control. That is how it is interpreted.

Although linear regressions are a useful measure of econometric analysis, in applied health economics research, many assumptions of linear regression models are violated in practice, particularly the assumption of linearity and exogeneity. As I just said, our projection here is X on Y. So, X should be sufficiently exogenous. But how do we guarantee that variables are exogenous in nature? So, some tests are there. At this moment, we are not emphasising everything but ensuring exogeneity. Then only we can project correctly.

Since there are some biases, there are some challenges with the assumptions. We should have an alternative solution that provides greater flexibility in modelling. Hence, we have to follow non-linear regression. It is sometimes referred to as a non-linear least square equation.

Non-linearity is a common feature of many applications in health economics. Generally, non-linear models are used to specify limited dependent variables. So, the first aspect that you just take note of, we have already highlighted in bold letters- 'limited dependent variable', that is, the dependent variable whose range is restricted in some way. We also referred to the author, who mentioned it clearly.

Most commonly found variables in health economics are usually binary in nature (basically having a choice between a pair of options). For example, an interviewer may ask a

participant that- has a doctor ever told you that you have arthritis? With the response coming as' yes or no'. So its answer is binary.

Similarly, there are multiple responses, maybe in ordinal or nominal forms. Multinomial variables, hence, multinomial regression models are also suggested. In the case of a multinomial variable, the choice is between several options that do not have natural ordering are there. For example, which hospital did you attend for your treatment? So, no order is followed out of the choices (if they are given). Hence, that is purely called the multinomial variable.

Another possible multiple option is there through the ordinal variable. Ordinal, as the word indicates order. It follows an order, which means choosing between several options with a natural ordering. For example, how would you rate your current health with options given as the percentage, or if the responses are as- excellent, very good, good, fair, or poor? This has to be followed accordingly in the non-linear regression model.

Another one is called the count variable. It is not purely continuous, as we used to see in the linear regression model. Count variables are considered to have some sporadic responses that might have some lexicographic choices. Count means it has to be at particular points. Count data typically contains a large proportion of zero observations and the long right tail of observation. We are giving examples as well. Many empirical analyses on the 'use of healthcare services' use dependent variable representing the event's count. For example, how many times you have visited your doctor in last 4 weeks? So, it has to be in count. Then, the ordinary regression model is not used.

Another aspect is 'duration variables', which indeed measure the duration. In applied health economics, many outcomes of interest reflect the time elapsed before an event occurs. For example, time to death. Here, the data are usually censored or restricted. Either left-censored (when we do not know the date of entry, for example- the date of the starting period of smoking, which might have caused the time to death). Or right censored (when we do not know the date of exit, such as the date of quitting smoking, etc.). So, that is also not the typical feature of ordinary regression.

We can present Whatever we discussed here in a systematic structured format in the table below, with regression type, related variables, estimation techniques, etc. A linear regression model is suggested when the dependent variables are continuous or quantitative. Independent variables may be quantitative or may be qualitative. So, in that case, OLS is applied.

Dependent variable	Independent variable	Estimation Method
Linear Regression		
Continuous (Quantitative)	Quantitative/Qualitative	Ordinary Least Squares (OLS)
Non-linear Regression		
Binary (Qualitative)	Quantitative/Qualitative	LPM/ Logit/ Probit Model
Categorical (Qualitative)	Quantitative/Qualitative	Multinomial Logit/ Probit
Ordered Categorical (Qualitative)	Quantitative/Qualitative	Cumulative Logit/ Probit
Repeated Binary (Qualitative)	Quantitative/Qualitative	Panel Logit/Probit
Count (Quantitative)	Quantitative/Qualitative	Hurdle or Two-step Model
Duration (Quantitative)	Quantitative/Qualitative	Survival/ Duration Analysis (E.g.- Cox proportional model)

In other cases, there are several possibilities for non-linear regression. The dependent variable might be binary, categorical, ordered, repeated binary, count, and duration (as we already discussed); however, the independent variable could be either qualitative or quantitative. Estimation techniques are suggested differently. When binary, linear probability model (LPM), logit, and probit, are suggested. In some cases, truncated model is also suggested depending on how far it is censored.

In the case of categorical (qualitative variable), we usually suggest applying multinomial logit or probit and some other interpretation of logit and probit through a logistic regression model, where some aspects of marginal effect are also important; you may follow from our paper. We have several papers in this direction.

In the ordered case, cumulative logit and probit are used. When repeated binary is there, we suggest following panel logit and probit. Poisson regression is suggested in the case of count data, and sometimes it is called a hurdle or two-step model. Even in Poisson, some people will find a negative Poisson regression model; you just go through those papers. For the duration, some of the models are suggested, like survival or duration analysis, and sometimes Cox proportional hazard model is also used.

Now, we are discussing emerging research designs. In recent times, applied health economics has seen increasing research focus towards the identification of causal impact on the outcome of interventions. So, those are considered to be policy variables or based on certain treatments. Identifying causal treatment effects is often illustrated using the potential outcomes framework, as mentioned by Robin (1974).

However, many times, in observational data, the observed relationship between outcome and treatment can be misleading due to unobserved factors and reverse causality, rendering the relationship of endogenous. For example, in assessing the impact of health insurance on healthcare utilisation, it is likely that the omitted variables reflecting the demand for health insurance may also partly explain the observed positive relationship between health insurance and healthcare utilisation. It is also possible that higher healthcare utilisation may be associated with having health insurance as well. In this case, the analyst must find a source of variation in the treatment that is independent of other factors that influence outcomes.

Such sources of variation are mainly derived from the experimental research design, famously known as a 'randomised control trial', and with certain freedom 'quasi-and natural experiments' are also applied. So far as RCT is concerned, these are considered as the gold standard for evaluating causal impact of an intervention. In an RCT, participants are randomly assigned to either a treatment or control group. So, treatment group means receiving the intervention, whereas the control group, where no interventions are applied or given. For example, clinical drug trials of the COVID-19 vaccine (which also we discussed in our earlier chapter).

In the case of random assignment, researchers control the assignment of treatment or exposure, but it simultaneously requires some aspects, such as balancing on measured confounders. They properly conducted random assignment balances unmeasured confounders between groups that minimise the bias. Another aspect of this (RCT) is called 'control group'. RCTs always include a group that is called the control group, which may receive a placebo or the standard treatment. Placebo means those who have actually received no treatment. Hence, we can actually have a clear comparison. You can refer to the work of J-PAL and recent works of Banerjee and Duflo (Esther Doppler), who have received the Nobel Prize using this.

For example, I am just mentioning that a group of researchers conducted an RCT to evaluate the heterogeneity in blood pressure response to four anti-hypertensive drugs. The participants are randomly assigned to different treatment periods, including a placebo washout period, and their blood pressure levels are measured during these times. Findings from that work is that the trial revealed significant variations in blood pressure response among participants, emphasising the importance of personalised treatment approaches. Refer to the study of Sundström et al. (2023), the latest work, and Mann et al. (2023) cited here. I think that will also clarify this further.

Another important aspect is called quasi-or natural experiments, where the situation mentioning quasi-or natural experiments is where persons are assigned to a treatment (or multiple treatments) and a control group, but the assignment is not fully randomised. In other words, quasi-experiments have a source of randomisation, that is 'as if' randomly assigned. While not explicitly controlled by researchers, they approximate randomised designs.

It has some features, like 'natural variation' and 'less susceptible to bias'. Quasi-experiments leverage naturally occurring variations (such as policy change events) to create treatment and control groups. In theory, as far as the less susceptible to bias is concerned, quasi-experiments are less susceptible to bias than traditional observational designs.

An example we are just putting here on this ground is- that a government implemented a new tax policy affecting the price of sugary beverages, as mentioned by Backholer et al. (2018), want to analyse health outcomes, such as obesity rates before and after the policy change to assess its impact. The method taken was like researcher conducted a quasi-natural experiment by comparing health outcomes before and after the policy change. Here, they analysed regions with stricter taxes and regions with milder restrictions. So, by the stricter taxes, they are referred to as the treatment group. In comparison, the mild restriction is the control group. So, they found that the tax policy had small effects on health outcomes, particularly on obesity rates. Although noticeable, it is not that substantial.

We need to note here for further clarification that, while quasi-experiments and natural experiments aim to assess causal effects, the key difference lies in the assignment process only. Natural experiments rely on naturally occurring variations, whereas quasi-experiments involve intentional but imperfect assignment methods, as mentioned in DiNardo (2008), you can refer for further details.

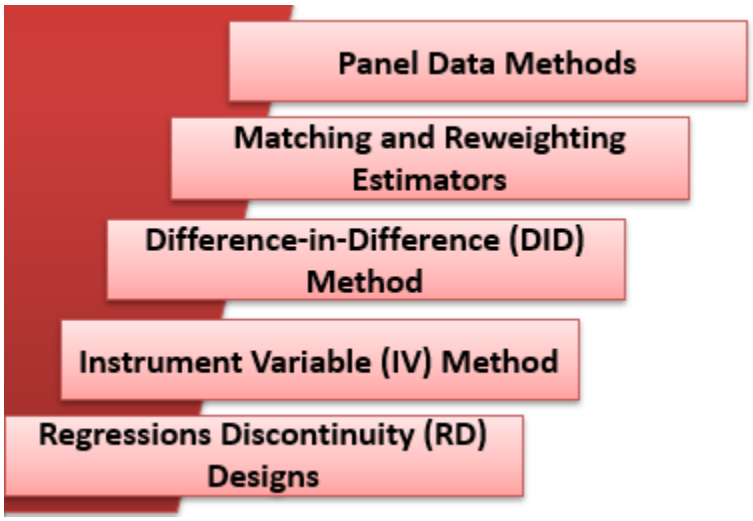
Here, summing up different research experiment designs. These have their objectives, participant assignment, control group, confounding considerations, level of evidence, advantages, and limitations:

Aspect	Experimental Study (a.k.a. Randomized Controlled Trial)	Quasi-Experimental Study
Objective	Evaluate intervention/treatment effect	Evaluate intervention/treatment effect
Participant Assignment	Random assignment	Non-random assignment (participant or researcher choice)
Control Group	Yes	Not always (if present, it strengthens evidence)
Confounding Consideration	No (but few claim, see- Manson et al.)	Yes (statistical techniques used)
Level of Evidence	Highest in hierarchy	One level below experimental studies
Advantages	Minimizes bias and confounding	Feasible when ethics/practicality limit experiments; works with smaller samples
Limitations	High cost; ethical constraints; generalizability issues	Lower ranking due to lack of randomization; susceptible to bias and confounding

Source Website: Quantifying Health by Choueiry, G. (2020)

An experimental study using randomised control (first column) and the second column is on the quasi-experimental study. So, you can see that the control group is strongly defined in the experimental design, whereas quasi is not always required. Confounding is important in quasi-experiments, whereas, in the case of experimental design, it is not considered, but few claims (as mentioned by Manson et al., you can refer to that work). Level of evidence, etc., is also different. You can just see other things, I think. The limitations, especially the RCT, are that it is highly expensive, it has ethical constraints, and generalizability is a major issue since it is based on a particular setting. In contrast, the quasi-experiments have a lower ranking due to the lack of randomisation and are susceptible to bias and confounding. So, you can refer to the work of Choueiry, G. (2020) from this lecture references for further clarification.

Let us discuss post-experiment estimation techniques. These are also important.



Post-experiment estimation techniques, starting with panel data methods, then matching and reweighting estimators, DID (difference in difference) method, instrument variable method, then regression discontinuity designs. So, these are considered to be post-experiment analysis.

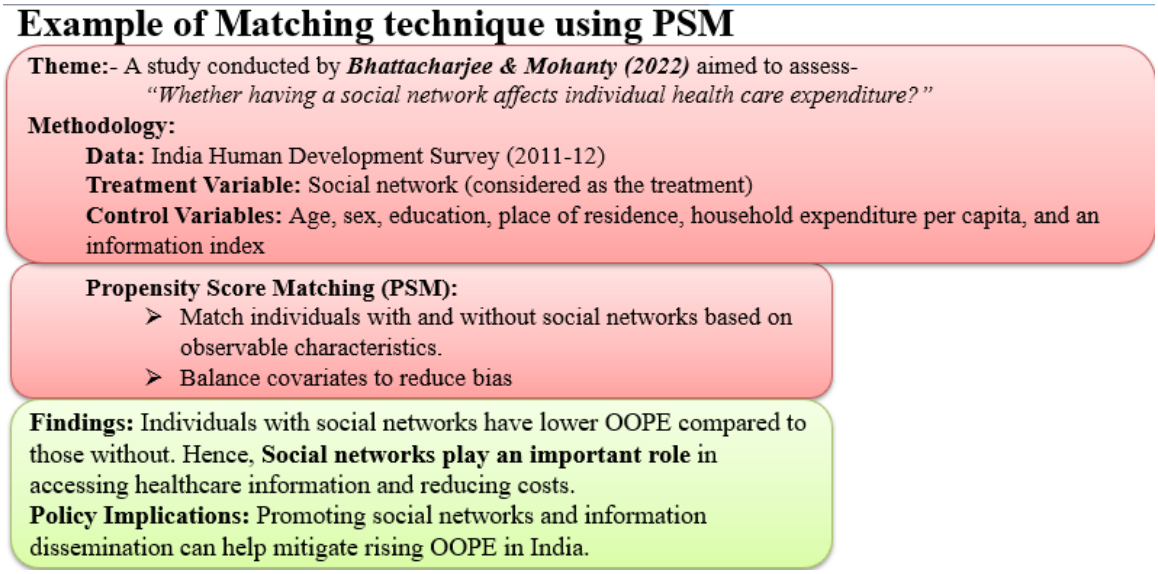
We are starting with panel data method, which we also discussed (to some extent) in earlier modules. However, a post-experimental panel data method involves analysing data with repeated observations on the same entities over time. The entities may be individuals, firms, or countries etc. The methodology usually considered is fixed effect or random effect. Fixed effect, where we control for individual-specific effects by including the fixed effect (i.e., dummy variables) for each entity (i.e., individual or firm). This captures time-invariant characteristics unique to each entity. Since we are not taking the random one, hence it is actually time-invariant.

Whereas in the random effects model, this assumes that individual-specific effects are not fixed. They are actually random and uncorrelated with the regressors. And this also accounts for both time-invariant and time-varying characteristics. That is why this is also

important in the context. The study by F.C. Guanais (2015), quantifies the combined effects of the 'expansion of primary healthcare' and 'conditional cash transfers' on infant mortality in Brazil, 1998-2010 using fixed effects estimation.

Another post-experimental technique is 'matching and reweighting estimators'. These estimators are used to address selection bias in observational studies. It is a quasi-experimental method in which the researcher uses statistical techniques to construct an artificial control group by matching each treatment unit with a non-treated unit of similar characteristics.

One of the most famous matching techniques is the propensity score matching technique, where the matching method is quite important. In propensity score matching (PSM), matches of treated and control units are based on observable characteristics. Reweights are assigned to the observations to create a pseudo-randomised sample. PSMs are used to evaluate program effectiveness when random assignment is not feasible, improving causal inference in non-experimental settings. For some of the examples of the PSM, you can follow our own paper, which we published.



This is our hypothesis, where we discussed whether having a social network affects individual healthcare expenditure. It was published in 2022. We use the data from India's Human Development Survey (IHDS) of 2011-12. The treatment variable was social network. And the control variables we considered were age, sex, education, place of residence, household expenditure per capita, and an information index. So, we applied PSM in this paper, where we matched the individuals with and without social networks based on observable characteristics. Balanced covariates were also used to reduce the bias. Our findings show that individuals with social networks have lower out-of-pocket expenditures compared to those without. Hence, social network plays an important role in assessing the healthcare information and reducing cost. It has strong policy implications because improving social networks and information dissemination would greatly mitigate the out-of-pocket expenditure of individuals and households in India.

Another post-experimental technique is called DID (Difference in Difference) method. DID is a quasi-experimental design that uses longitudinal data. That is important to note. Earlier in the previous method, we did not say that. Longitudinal data is not required. It was applied in one period of time. Longitudinal data from treatment and control groups to estimate causal effects. It is commonly used to assess the impact of specific interventions or treatments. The methodology of this DID is like comparing outcomes before and after treatment for both treated and control groups. It estimates treatment effect as a difference in different approaches. The uses of these are as follows: they will be used in evaluating policy interventions (especially, for example, the health impact of changes in tobacco prices). Analysing natural experiments in these areas is important, as we already mentioned. The specific examples and references you can follow for further clarification of DID from the paper by Tirgil and others, published recently in 2023, that demonstrate the causal impact of family medicine reform on patient satisfaction in Turkey using DID.

Another method is called the Instrumental Variable Method. Usually, this is quite important for controlling endogeneity and omitted variable bias (OVB). In short, it is called IV methods. Usually describe endogeneity and omitted variable bias. An instrument (that is exogenous) variable is used to predict treatment assignment. Estimates causal effects by exploiting variation in the instrument. You can refer to our paper also. I am just citing here for further clarification- Haile and Mohanty published in a top journal on migration and remittances. This is based on work in Ethiopia, published in 2022. We have discussed what the right instrument should be, how far the instrument techniques are suggested, and how to check whether it is actually exogenous enough. Then, some use of this is estimating causal effects when direct manipulation is not possible. Correcting for endogeneity in an econometric model is a must and usually suggested. Especially in the panel model, the endogeneity is expected to be very huge, whereas in cross-sections (since it is for a particular time period), responses are taken, and most of the cross-section variables are considered to have fixity in response, used to be non-dynamic. Hence, endogeneity is expected to be lesser. In the case of the panel model, it is usually suggested that the IV method be applied, with the possibility of some endogeneity within it.

So, we discuss tests and how to check whether there is endogeneity or not. You can refer to our paper as well. Even the study by De Neve and others (2015) established the impact of HIV status knowledge on linkage to HIV treatment in South Africa using the IV approach.

So, the last post-experimental technique we are discussing at this moment is regression discontinuity designs (RD in short). RD design usually evaluates treatment effects near a threshold. Here, participants are assigned to the intervention and the control conditions based on a cut-off score on a pre-intervention measure that typically assesses need or merit.

So, the methodology used explores outcomes around a cut-off point (e.g., the eligibility criteria) and compares units just above and below the threshold. Assessing program impacts near eligibility thresholds is important for policy or program evaluation. Identifying causal effects in situations with clear rules is also important. The study by J. Bor and others

in 2017 showed the causal impact of HIV treatment on mortality, employment and education in South Africa, using the regression discontinuity design approach.

So, that is all I think in this module. The following two lectures will be on data handling packages in healthcare or in health. We will discuss open and licensed data handling packages in the next two lectures, which will also be useful for a researcher to know and apply. I think you can refer to the specific papers in the references. We have highlighted our own paper here as well, as well as its title. And for the instrumental variable, you can also refer (as I told you) to another paper by Haile and Mohanty on remittance flows. So, these are all for today, and I think these references are extremely important for the researchers. Whoever already has some experience, they will also benefit out of it. So, I think that is all for today.

I expect you to listen to the other two lectures. It will be equally important. You can also suggest others to follow. That is all. Thank you.