Econometrics Lecture 11

Assessing Linear Regression Models

Hang Miao

Rutgers University

April 13, 2021

Overview

1. Introduction
2. Threats to Internal Validity
   - Omitted Variable Bias
   - Wrong Functional Form
   - Errors-in-Variables Bias
   - Sample Selection Bias
   - Simultaneous Causality Bias
3. Threats to External Validity
4. Appendix

Introduction

Introduction

- Lets step back and take a broader look at regression. Is there a systematic way to assess (critique) regression studies? We know the strengths of multiple regression, but what are the pitfalls?
- We will list the most common reasons that multiple regression estimates, based on observational data, can result in biased estimates of the causal effect of interest.
- In the test score application, well try to address these threats as best we can and assess what threats remain. After all this work, what have we learned about the effect on test scores of class size reduction?

A Framework for Assessing Statistical Linear Regression Models: Internal and External Validity

Definition (Internal validity)

the statistical inferences about causal effects are valid for the population being studied.

Definition (External validity)

the statistical inferences can be generalized from the population and setting studied to other populations and settings, where the setting refers to the legal, policy, and physical environment and related salient features.

Threats to Internal Validity

Omitted Variable Bias: Definition

Definition (Omitted Variable Bias)

Omitted variable bias arises if:

1. There exists a latent regressor which is a determinant of y.
2. This latent regressor also correlated with at least one included regressor.

Example

Suppose the model we estimate is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + v$
While the true model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \beta_{p+1} x_{p+1} + u$

- If $\beta_{p+1} \neq 0$, then the latent feature x_{p+1} is a determinant of Y. Condition 1 satisfied.
- If there exists a $j \in \{1, 2, \cdots p\}$ such that $\mathrm{Cov}(x_j, x_{p+1}) \neq 0$, then the latent regressor x_{p+1} is correlated with included regressor x_p. Condition 2 satisfied.

Omitted Variable Bias: Consequence

The consequence is that Assumption 3: Strict Exogeniety dose not hold anymore. i.e $\mathrm{E}[\boldsymbol{u}|\boldsymbol{X}] \neq 0$. This can be showed from the above example.

$$\begin{aligned} \mathrm{E}[v|\boldsymbol{X}] &= \mathrm{E}[\,\beta_{p+1}x_{p+1} + u \,|\, \boldsymbol{X}] \\ &= \beta_{p+1}x_{p+1} + \mathrm{E}[\,u \,|\, \boldsymbol{X}] \\ &= \beta_{p+1}x_{p+1} \\ &\neq 0 \end{aligned}$$

As a result, $\hat{\boldsymbol{\beta}}_{ols}$ is biased and inconsistent. The bias could be calculated by the following formula:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_{x_j}}\right) \rho_{x_j u}$$

where σ_u is the standard deviation of uncertainty u, σ_{x_j} is the standard deviation of the j-th feature x_j , $\rho_{x_j u}$ is the correlation between x_j and u.

Omitted Variable Bias: General Expression for Bias and Inconsistency

$$
\begin{aligned}
\hat{\beta}_{ols} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{v}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{v} \\
&= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{v} \\
\mathrm{E}[\hat{\beta}_{ols}|\boldsymbol{X}] &= \mathrm{E}[\beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{v}|\boldsymbol{X}] = \beta + \mathrm{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{v}|\boldsymbol{X}] \\
&= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ \mathrm{E}[\boldsymbol{v}|\boldsymbol{X}] \\
&\neq \beta \\
\hat{\beta}_{ols} &= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{v} \\
&= \beta + \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}\left(\frac{\boldsymbol{X}'\boldsymbol{v}}{n}\right) \\
&\xrightarrow{p} \beta + \mathrm{E}[\boldsymbol{x}\boldsymbol{x}']^{-1}\mathrm{E}[\boldsymbol{x}v]
\end{aligned}
$$

Solutions to Omitted Variable Bias

- If the omitted causal variable can be measured, include it as an additional regressor in multiple regression;
- If you have data on one or more controls and they are adequate (that is, if conditional mean independence plausibly holds), then include the control variables;
- Possibly, use panel data in which each entity (individual) is observed more than once;
- If the omitted variable(s) cannot be measured or adequately controlled for, use instrumental variables regression;
- Run a randomized controlled experiment.

Solutions to Omitted Variable Bias: Control Variable

Motivation

- It is hard to conduct the randomized controlled experiment for social social science object
- Normally the latent variable is omitted not by accident but it is intrinsically hard to measure.
- Heavily rely on econometric techniques like introducing controlled variable, instrumental variable and panel data to solve OVB.

Definition (Control Variable)

Control Variable w_k is a regressor such that:

1. correlated with the omitted causal factors
2. itself not necessary a causal factor or determinant of y.

Solutions to Omitted Variable Bias: Control Variable

Example: California test score data

$$TestScore = \underset{(5.6)}{700.2} - \underset{(0.27)}{1.00} STR - \underset{(.033)}{0.122} PctEL - 0.547 LchPct + v \qquad \overline{R}^2 = 0.773$$

- STR: student-teacher ratio
- PctEL: percent English Learners in the school district
- LchPct: percent of students receiving a free/subsidized lunch (only students from low-income families are eligible)

Questions

- Which variable is the variable of interest?
- Which variables are control variables? Might they have a causal effect themselves? What do they control for?

Answer

$$\underset{(5.6)}{TestScore = 700.2} - \underset{(0.27)}{1.00STR} - \underset{(.033)}{0.122PctEL} - 0.547LchPct + v \qquad \overline{R}^2 = 0.773$$

- STR is the variable of interest
- PctEL probably has a direct causal effect (school is tougher if you are learning English!). But it is also a control variable: immigrant communities tend to be less affluent and often have fewer outside learning opportunities, and PctEL is correlated with those omitted causal variables. PctEL is both a possible causal variable and a control variable
- LchPct might have a causal effect (eating lunch helps learning); it also is correlated with and controls for income-related outside learning opportunities. LchPct is both a possible causal variable and a control variable.

Omitted Variable Bias: Evaluation for Control Variable

Three interchangeable statements about what makes for an effective control variable:

1. An effective control variable is one which, when included in the regression, makes the error term uncorrelated with the variable of interest.
2. Holding constant the control variable(s), the variable of interest is as if randomly assigned.
3. Among individuals (entities) with the same value of the control variable(s), the variable of interest is uncorrelated with the omitted determinants of y

Omitted Variable Bias: Conditional Mean Independence

If we push the above evaluation criterion to extreme. We got the following mathematical condition under which the variable of interest is uncorrelated with uncertainty u.

Definition (Conditional Mean Independence)

$$\mathrm{E}[\, u \,|\, x_1, x_2, \cdots, x_p, w_1, w_2, \cdots, w_q \,] = \mathrm{E}[\, u \,|\, w_1, w_2, \cdots, w_q \,]$$

Where x_j denote the variable of interest and w_k denote the control variable. Equivalent matrix expression is:

$$\mathrm{E}[\, \boldsymbol{u} \,|\, \boldsymbol{X},\, \boldsymbol{W} \,] = \mathrm{E}[\, \boldsymbol{u} \,|\, \boldsymbol{W} \,]$$

Where $\boldsymbol{X}$, $\boldsymbol{W}$ denote the matrix store the sample data for variable of interest and control variable.

Omitted Variable Bias: Conclusion for Control Variable

Under the situation of Omitted Variable Bias, the original Assumption 3: Strict Exogeniety not valid. However, if we could find control variable w_k such that the Conditional Mean Independence is satisfied, then

- the estimated coefficients $\hat{\beta}_j$ for the variable of interest is still unbiased, consistent and conveys the causal effect $(\{\boldsymbol{x}_i, y_i\} i.i.d)$
- While $\hat{\beta}_k$ for control variable still suffered from OVB: biased, inconsistent and not convey causal relationship.

Wrong Functional Form

Definition (Functional Form Misspecification)

Functional Form Misspecification arises if the functional form is incorrect, for example:

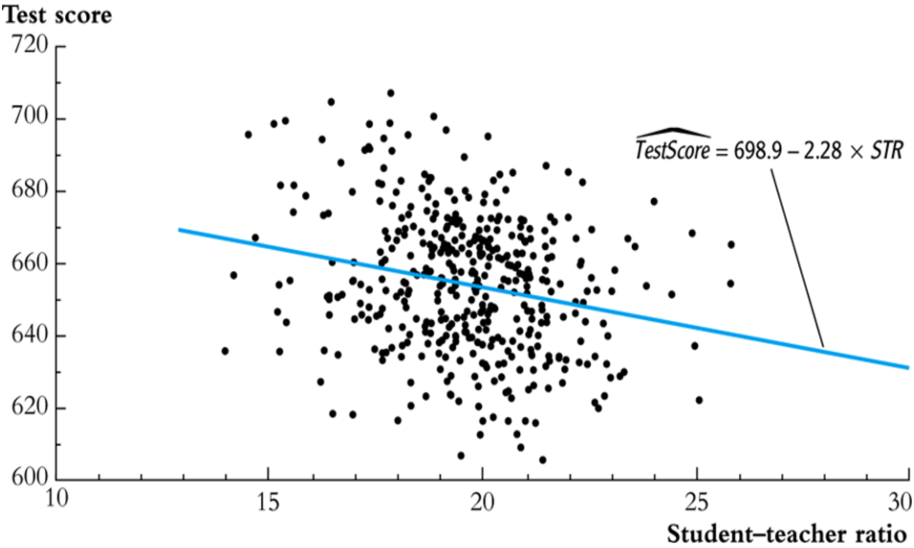
1. misspecify the regression functions form in one or more regressors x_j.
2. an interaction term $x_j x_k$ is incorrectly omitted
3. misspecify the regression functions form in dependent variable y.

Example

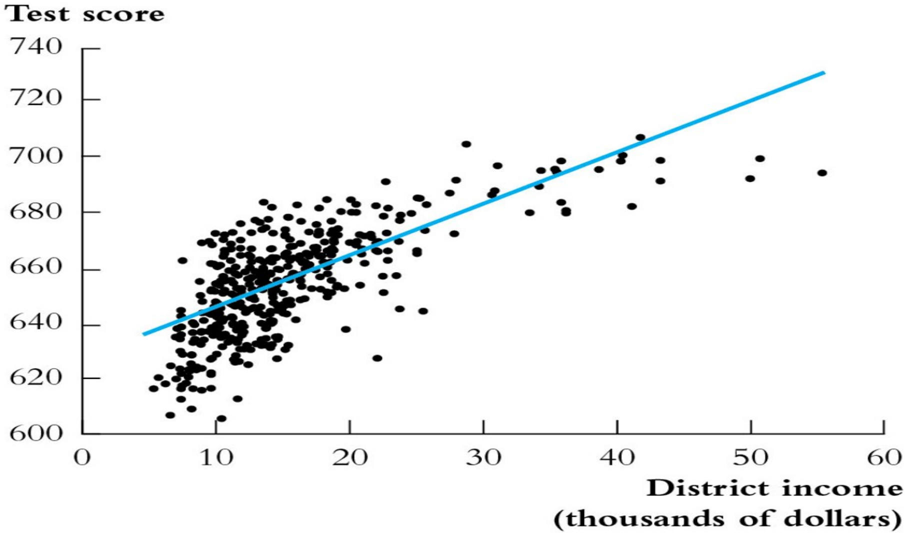
Suppose the model we estimate is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + u$
While the true model is $log(y) = \beta_0 + \beta_1 x_1^2 + \beta_2 log(x_2) + \cdots + \beta_p x_p x_3 + u$
As a result:

- Assumption 1: Linear Model does not hold anymore.
- $\hat{\beta}_{ols}$ is biased and inconsistent

Using Linear Model to Estimate the Linear Relationship



Using Linear Model to Estimate Non Linear Relationship



Solutions to Functional Form Misspecification

1. **Continuous Dependent Variable** y: use the appropriate nonlinear specifications in x_j (logarithms, interactions, etc.)
2. **Discrete (example: binary) Dependent Variable** y: need an extension of multiple regression methods (probit or logit analysis for binary dependent variables).

Errors-in-Variables Bias

So far we have assumed that the sample data $\boldsymbol{X}, \boldsymbol{y}$ is measured without error. In reality, economic data often have measurement error.

Definition (Errors-in-Variables Bias)

Errors-in-Variables Bias arises in the following categories:

1. Data entry errors in administrative data
2. Recollection errors in surveys (when did you start your current job?)
3. Ambiguous questions (what was your income last year?)
4. Intentionally false response problems with surveys (What is the current value of your financial assets? How often do you drink and drive?)

Consequence

- In most cases, $\hat{\boldsymbol{\beta}}_{ols}$ is biased and inconsistent.
- If some extremely rare false measurement pattern is followed, $\hat{\boldsymbol{\beta}}_{ols}$ may be unbiased and consistent. (see HW9 question7(Exercise 9.6))
- If the input-error is involved with dependent variable y only with zero mean and uncorrelated with both independent and dependent variables $(\boldsymbol{x}, y)$, then $\hat{\boldsymbol{\beta}}_{ols}$ is unbiased and consistent
- If the input-error is involved with independent variable $\boldsymbol{x}$, then $\hat{\boldsymbol{\beta}}_{ols}$ is biased and inconsistent.

Solutions to Errors-in-Variables Bias

1. Obtain better data (often easier said than done).
2. Develop a specific model of the measurement error process. This is only possible if a lot is known about the nature of the measurement error
3. Instrumental Variables Regression.

Sample Selection Bias

Type of Missing Data

Data are often missing. Sometimes missing data introduces bias, sometimes it doesnt. It is useful to consider three cases:

1. Data are missing at random.
2. Data are missing based on the value of one or more x_j
3. Data are missing based in part on the value of y or u

Definition (Sample Selection Bias)

- Sample Selection Bias arises if data are missing in the way of Type 3.
- Missing data in the way of Type 1 and Type 2 don't introduce bias. But the standard deviation would be larger.

Example of Type 1: Missing Data at Random

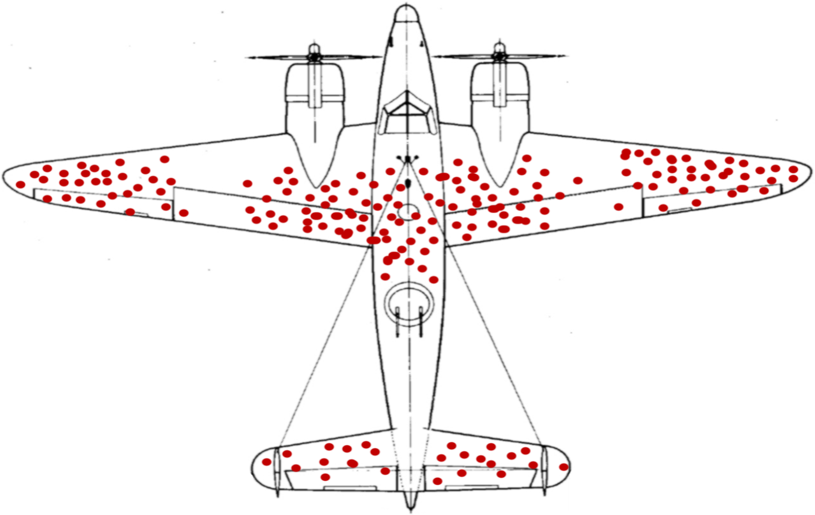
Suppose you took a simple random sample of 100 workers and recorded the answers on paper but your dog ate 20 of the response sheets (selected at random) before you could enter them into the computer. This is equivalent to your having taken a simple random sample of 80 workers (think about it), so your dog didnt introduce any bias.

Example of Type 2: Missing Data Based on x_j

In the test score/class size application, suppose you restrict your analysis to the subset of school districts with STR < 20. By only considering districts with small class sizes you wont be able to say anything about districts with large class sizes, but focusing on just the small-class districts doesnt introduce bias. This is equivalent to having missing data, where the data are missing if STR > 20. More generally, if data are missing based only on values of x_j, the fact that data are missing doesnt bias the OLS estimator.

Aerial Dogfight in World War II

Center for Naval Analyses Count the Damage Part of Returning Planes



Example 1 of Type 3: Height of Undergraduates

Your stats prof asks you to estimate the mean height of undergraduate males. You collect your data (obtain your sample) by standing outside the basketball teams locker room and recording the height of the undergraduates who enter. Is this a good design will it yield an unbiased estimate of undergraduate height? Formally, you have sampled individuals in a way that is related to the outcome y (height), which results in bias.

Example 2 of Type 3: Mutual Funds

Do actively managed mutual funds outperform hold-the-market funds?

- Sampling scheme: simple random sampling of mutual funds available to the public on a given date.
- Data: returns for the preceding 10 years.
- Estimator: average ten-year return of the sample mutual funds, minus ten-year return on SP500
- Is there sample selection bias?
- How is this example like the survival plane example?

Example 3 of Type 3: Returns to Education

What is the return to an additional year of education?

- Sampling scheme: simple random sample of employed college grads (employed, so we have wage data)
- Data: earnings and years of education
- Estimator: regress ln(earnings) on years of education
- Ignore issues of omitted variable bias and measurement error is there sample selection bias?
- How does this relate to the survival plane example?

Solutions to Sample Selection Bias

- Collect the sample in a way that avoids sample selection.
  - Survival plane example: Follow Wald's advice: add armor to the undamaged part
  - Basketball player example: obtain a true random sample of undergraduates, e.g. select students at random from the enrollment administrative list.
  - Mutual funds example: change the sample population from those available at the end of the ten-year period, to those available at the beginning of the period (include failed funds)
  - Returns to education example: sample college graduates, not workers (include the unemployed)
- Randomized controlled experiment.

Simultaneous Causality Bias

Definition (Simultaneous Causality Bias)

- So far we have assumed that x_j causes y.
- Simultaneous Causality Bias arises if y causes x_j, too

Example: Class size effect

- Low STR results in better test scores
- But suppose districts with low test scores are given extra resources: as a result of a political process they also have low STR
- What does this mean for a regression of TestScore on STR?

Simultaneous Causality Bias

1. Causal effect on y of x_j: $y = \beta_0 + \beta_1 x_j + u$
2. Causal effect on x_j of y: $x_j = \gamma_0 + \gamma_1 y + v$

- Large u means large y, which implies large x_j (if $\gamma_1 > 0$) Thus $\mathrm{Cov}(x_j, u) \neq 0$
- Thus $\hat{\beta}_1$ is biased and inconsistent

Example: Class size effect

A district with particularly bad test scores given the STR (negative u) receives extra resources, thereby lowering its STR; so STR and u are correlated

Solutions to Simultaneous Causality Bias

- Run a randomized controlled experiment. Because x_j is chosen at random by the experimenter, there is no feedback from the outcome variable to y (assuming perfect compliance).
- Develop and estimate a complete model of both directions of causality. This is the idea behind many large macro models (e.g. Federal Reserve Bank-US). This is extremely difficult in practice.
- Use instrumental variables regression to estimate the causal effect of interest (effect of x_j on y, ignoring effect of y on x_j).

Threats to External Validity

Threats to External Validity

Assessing threats to external validity requires detailed substantive knowledge and judgment on a case-by-case basis.

Example

How far can we generalize class size results from California?

- Differences in populations
  - California in 2019?
  - Massachusetts in 2019?
  - Mexico in 2019?
- Differences in settings
  - different legal requirements (e.g. special education)
  - different treatment of bilingual education
- Differences in teacher characteristics

Appendix

$\boldsymbol{X}'\boldsymbol{X} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'$

$$
\begin{aligned}
\boldsymbol{X}'\boldsymbol{X} &= \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & x_{3p} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} & \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & \dots & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \boldsymbol{x}_3' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'
\end{aligned}
$$

Matrix Expression for $x_i x_i'$

$$\begin{aligned}
\boldsymbol{x}_i \boldsymbol{x}_i' &= \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \begin{pmatrix} 1 & x_{i1} & x_{i2} & \ldots & x_{ip} \end{pmatrix} \\
&= \begin{bmatrix} 1 & x_{i1} & x_{i2} & \ldots & x_{ip} \\ x_{i1} & x_{i1}^2 & x_{i1}x_{i2} & \ldots & x_{i1}x_{ip} \\ x_{i2} & x_{i2}x_{i1} & x_{i2}^2 & \ldots & x_{i2}x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{ip} & x_{ip}x_{i1} & x_{ip}x_{i2} & \ldots & x_{ip}^2 \end{bmatrix}
\end{aligned}$$

Matrix Expression for $\boldsymbol{X}'\boldsymbol{X}$

$$\begin{aligned}
\boldsymbol{X}'\boldsymbol{X} &= \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \\
&= \sum_{i=1}^{n} \begin{bmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{ip} \\ x_{i1} & x_{i1}^2 & x_{i1}x_{i2} & \cdots & x_{i1}x_{ip} \\ x_{i2} & x_{i2}x_{i1} & x_{i2}^2 & \cdots & x_{i2}x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{ip} & x_{ip}x_{i1} & x_{ip}x_{i2} & \cdots & x_{ip}^2 \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ip} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ip} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ip} & \sum_{i=1}^n x_{ip}x_{i1} & \sum_{i=1}^n x_{ip}x_{i2} & \cdots & \sum_{i=1}^n x_{ip}^2 \end{bmatrix}
\end{aligned}$$

$\frac{\boldsymbol{X}'\boldsymbol{X}}{n} \xrightarrow{p} \mathrm{E}[xx']$

$$
\begin{aligned}
\frac{\boldsymbol{X}'\boldsymbol{X}}{n} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \\
&= \begin{bmatrix}
\frac{\sum_{i=1}^{n} 1}{n} & \frac{\sum_{i=1}^{n} x_{i1}}{n} & \frac{\sum_{i=1}^{n} x_{i2}}{n} & \cdots & \frac{\sum_{i=1}^{n} x_{ip}}{n} \\
\frac{\sum_{i=1}^{n} x_{i1}}{n} & \frac{\sum_{i=1}^{n} x_{i1}^2}{n} & \frac{\sum_{i=1}^{n} x_{i1}x_{i2}}{n} & \cdots & \frac{\sum_{i=1}^{n} x_{i1}x_{ip}}{n} \\
\frac{\sum_{i=1}^{n} x_{i2}}{n} & \frac{\sum_{i=1}^{n} x_{i2}x_{i1}}{n} & \frac{\sum_{i=1}^{n} x_{i2}^2}{n} & \cdots & \frac{\sum_{i=1}^{n} x_{i2}x_{ip}}{n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{\sum_{i=1}^{n} x_{ip}}{n} & \frac{\sum_{i=1}^{n} x_{ip}x_{i1}}{n} & \frac{\sum_{i=1}^{n} x_{ip}x_{i2}}{n} & \cdots & \frac{\sum_{i=1}^{n} x_{ip}^2}{n}
\end{bmatrix} \\
&\xrightarrow{p} \begin{bmatrix}
1 & \mathrm{E}[x_1] & \mathrm{E}[x_2] & \cdots & \mathrm{E}[x_p] \\
\mathrm{E}[x_1] & \mathrm{E}[x_1^2] & \mathrm{E}[x_1 x_2] & \cdots & \mathrm{E}[x_1 x_p] \\
\mathrm{E}[x_2] & \mathrm{E}[x_1^2] & \mathrm{E}[x_2^2] & \cdots & \mathrm{E}[x_2 x_p] \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\mathrm{E}[x_p] & \mathrm{E}[x_1 x_p] & \mathrm{E}[x_2 x_p] & \cdots & \mathrm{E}[x_p^2]
\end{bmatrix} \equiv \mathrm{E}[\boldsymbol{x}\boldsymbol{x}']
\end{aligned}
$$