ECON 442 / 642 Term Paper Instructions

Last updated **February 10, 2022**

This document is long but you should carefully read all of it!

Overview and Learning Goals

- Identify a relevant question in development economics that is answerable using standard methods of quantitative economic research.
- Find, understand, and describe an economic dataset.
- Apply econometric methods to an economic dataset in an attempt to answer a research question.
- Interpret the results from applying econometric methods to an economic dataset.
- Understand and discuss the limitations of these results.

General Rules

- You may use any textbooks, materials from this or other classes, etc. while preparing the term paper.
- You may consult with other students and faculty members while preparing the term paper.
- You may use a dataset given to you by another student or faculty member, subject to some additional requirements discussed at the end of this document.
- The written submissions (proposal and final paper), presentation slides, and script file must all be your own unaided work.
- You may not use an assignment completed for another course without both instructors' permission.
- If any part of the term paper assignments is plagiarized, the assignment will receive a zero grade and you will be reported to the Office of Student Conduct. If you have any doubt about what constitutes plagiarism, consult http://library.duke.edu/research/plagiarism/index.html or https://twp.duke.edu/twp-writing-studio/resources-students/sources. Accidental plagiarism due to unfamiliarity with rules is still plagiarism and will be treated accordingly.
- You may use any generally accepted style of academic citation. I encourage you to use an author-date system for in-text citation (Marx, 1867). For reports without a listed author, use the name of the agency publishing the report (African Development Bank, 2015).
- The term paper proposal is due at *16:00* on *Tuesday 15 March*.
- The term paper presentations will take place between *Friday 1 April* and *Wednesday 20 April*.
- The final term paper is due at *16:00* on *Monday 25 April*.
- Assignments that are less than 24 hours late will receive a 25 percentage point penalty. Assignments that are 24 to 48 hours late will receive a 50 percentage point penalty. Assignments that are more than 48 hours late will receive a zero grade.

How To Get Help

- Jennifer Kades, a doctoral student in the economics department, will hold office hours each week to advise on term paper preparation. Jennifer's office hours will be Mondays 11:00 - 12:00 and Fridays 12:00 - 13:00. You can contact Jennifer at mailto:jennifer.kades@duke.edu.

- I will answer questions about term papers in office hours and, for short questions only, by email.
- You can ask Jennifer or me questions about how to identify a research question; find data; understand, analyze, and interpret your data; and construct your script file (more details below).

What is a Research Question?

This may sound obvious but often causes confusion. Here are two simple examples of good research questions:

- "What is the relationship between earnings and the highest level of schooling attained?" This question examines a relationship between two well-defined and relatively easily measured concepts: earnings and grade attainment. The relationship is based in economic theory: education may increase job-relevant skills, raising the workers' marginal revenue product, which raises their earnings under additional assumptions. The answer is not obvious: the skills learned in some grades of some schools may have little use in the labour market, or students may learn little in some schools. The question allows you to do difficult econometric analyses if you choose to, e.g., thinking carefully about omitted variable bias or nonlinearity.
- "How did income inequality in country X change from year A to year B?" This question studies a well-defined and important economic concept. The answer is not obvious: it may be different depending on the measure of inequality you use and may be different depending on which people in the country you study. The question allows you to do difficult econometric analyses if you choose to, e.g., thinking carefully about measurement error and missing data.

Some warning notes about research questions:

- Questions with obvious answers are seldom worth asking.
- Questions may be about describing a single economic concept or about the relationship between two economic concepts (how much more/less do workers with more schooling earn?).[1]
- Questions may concern an unconditional economic relationship or conditional economic relationship (e.g. the earnings-schooling relationship conditional on experience). In practice, you should typically consider both unconditional and conditional forms of your relationship of interest. The choice of conditioning variables or control variables can be very important.
- Causal relationships, such as the "effect" of schooling on earnings, are extremely difficult to study. I recommend that you examine a relationship and discuss the conditions under which it might be interpreted as causal. This theme will recur through the course.
- Choose a question you are interested in answering! Investing large amounts of time in something you don't care about can be frustrating.

Final Term Paper

[1] Most papers in this course focus on linear relationships. But economic relationships may be *nonlinear* or even *nonmonotonic*. A nonlinear relationship between Y and X has a magnitude that varies with X. A nonmonotonic relationship between Y and X has a sign that varies with X. For example, a log function allows a nonlinear but monotonic relationship and a quadratic function allows a nonmonotonic relationship.

These are the minimum requirements for the term paper:

- The paper should be a reasonable attempt to answer a well-posed research question in development economics. The paper will not provide a comprehensive or flawless answer and you should not try and ignore the inevitable flaws of the paper. You may interpret development economics broadly: anything that is related to the topics covered in the course is probably acceptable. Consult the instructor if you are in doubt.
- The answer must be based on analysis of an appropriate quantitative dataset. The data may be generated from a census or survey of firms, households, students, etc., administrative records on clinics, firms, schools, etc., or aggregate data on cities, countries, or regions. The data may be from replication files for a published research paper. The methods will probably include summary statistics and linear regression models. You are welcome to use more advanced regression-based methods (multilevel models, panel data methods, nonlinear regression models, instrumental variables, etc.) or alternative methods such as matching. But you are definitely not required to go beyond the empirical methods discussed in this class and in your previous econometrics class(es).
- The answer and/or empirical analysis may be based in part on a theoretical framework or formal model. But the paper is primarily an empirical exercise and any model should not come at the expense of good empirical work.
- The paper should include a brief discussion of prior research on the topic. This does not need to be a comprehensive review of the existing literature. It should show evidence of some research beyond the assigned readings for the class. It should help to motivate why the research question is important and why you have chosen your dataset and empirical strategy.
- The paper does not need to generate new knowledge and may ask the same question and use similar data and methods to existing research. It should not be a direct replication of an existing research paper.
- The paper should include a section on limitations of the answer, due to omitted variables, missing data, imperfect link between theory and data, etc. The paper should frankly discuss these limitations rather than trying to hide them. *All research has problems and being a good researcher means understanding the problems you face!*
- The paper must include a detailed discussion of the dataset source and structure. Readers of the paper should be able to find the dataset and understand how the dataset was created, what units are observed and analyzed, and how the key variables are defined. If you want to use proprietary or restricted-access data, consult the instructor.
- The paper must be accompanied by a data file and a script file that generates all results, tables, and/or figures used in the term paper.[2] Readers should be able to exactly replicate all results in the paper by running the script file on the data file.
- The paper should not exceed 20 pages using 11 point or larger font, single spaced lines, and 2.5cm / 1 inch margins. This length limit includes tables, figures, and references. There is no penalty for producing short papers! Well-written papers with fewer than 15 pages have scored very high grades in

[2] You may use any statistical analysis package of your choice, provided it supports scripting. Support will be provided for Stata and minimal support might be provided for Python or R. If you are using Stata, you should submit a `.dta` and a `.do` file, not a `.log` file.

past years.

- The paper should be submitted as a pdf file. The script file should be submitted in a readable format (`.do`, `.py`, etc.).

Here is one example of an appropriate term paper structure, though this structure is not required and is *not intended to be prescriptive*:

Introduction: State the research question, motivate why it matters, preview the dataset and empirical strategy, preview the results, contextalize the results relative to the existing literature. Approximately 4 pages.

Data: Explain the dataset, how it was obtained, what the unit of observation is, what population it represents, what questions were used to measure the variables of interest, and how new variables have been constructed from the raw data. Approximately 3 pages.

Empirical strategy: Explain precisely how the relationship of interest will be measured: what variable will be regressed on what other variables, what parameters of the regression will be relevant, etc. Write out any regression models you plan to use. Approximately 3 pages.

Results: Discuss the results from applying the empirical strategy to the data. The results may be presented as tables and/or figures and should be discussed in text. Approximately 7 pages including figures/tables.

Interpretation and limitations: Discuss how your results answer (or fail to answer) your research question and any limitations of your results. The limitations may reflect features of the data (e.g. missing responses) or features of the empirical strategy (omitted variables). Your discussion of the limitations should draw on critiques of papers and methods discussed in class. For example, there will almost certainly be a set of variables omitted from a regression model that might be correlated with both the outcome and the included regressor of interest. Think about which omitted variables might be most important for your research question, data, and empirical strategy. Approximately 3 pages.

It is sometimes effective to swap the order of the data and empirical strategy sections. Combining the results and interpretation/limitations sections can work well.

Term Paper Proposal

The term paper proposal is a short description of your research question, the dataset(s) and methods you propose to use to answer the question. The proposal may not be longer than four pages with 11 point or larger font, single spacing, and 2.5cm / 1 inch margins. You may write the proposal in any format. Here is one possible format that, once again, *is not intended to be prescriptive*:

Section 1: Clearly state the research question and briefly explain why this question is important for, or at least relevant to, economic development.

Section 2: Explain the empirical strategy you will use answer this question. Be specific about what regressions you propose to run, using what variables, and how you will interpret them.

Section 3: Discuss the dataset(s) you may use to answer the question. You may suggest multiple datasets - this is a good opportunity to get feedback on their relative merits. You should discuss the unit of observation in the dataset (individual, household, firm, etc.), what population the datset is meant to represent (e.g. representative survey of households in Mali), and what variables you will use (e.g. female enrollment and number of children in the household as functions of adult education, adult earnings, and household wealth). You should discuss any concerns about measurement error (e.g. some values of earnings implausibly high)

or missing data (e.g. highest grade attained was not asked of respondents older than 35).

If you are deciding between two research questions, you can submit a term paper proposal that discusses both of the two questions.

You may be advised to change your research question, dataset(s), and/or methods if your proposal looks infeasible. This does not mean you have asked a bad question or selected bad methods. This means the combination of question, data and methods are unlikely to yield a term paper by the end of the semester.

The graded proposals will be returned by Friday 26 March. Everyone should meet with the instructor and/or term paper consultant before or after (ideally before and after) submitting the proposal.

Term Paper Presentation

You will be randomly assigned to a presentation date between 1 and 20 April. You may swap your presentation date with another student, subject to mutual consent. You should inform me about any swaps at least 24 hours before the presentation. The presentation is intended to:

- Clearly and efficiently communicate the question of your term paper, the data and methods you use to answer this question, and the (preliminary) answers you reach.
- Apply general presentation skills to a research presentation.
- Interact effectively with an audience on a research topic.

Class participation is strongly encouraged during the presentations. You may ask questions or offer suggestions during other students' presentations, though you may not interrupt the presenter. Particularly insightful or useful suggestions will earn extra credit toward your participation grade.

As the presenter, you should plan the time allocation for your presentation carefully. All presentations will be cut off after 25 minutes. You control the presentation and can choose whether to accept questions/comments from other students. You should accept and answer/address at least three questions/comments. This means that your prepared presentation should not take longer than 20 minutes.

You will not have time to discuss your data and results in detail. Here is one possible *but not required* format for the presentation: Begin by clearly stating your research question, explaining why the audience should care about this question, and previewing the answer. Then explain the data and methods you use to answer the question. Conclude by presenting your results and acknowledging any limitations of the data or methods that might cast doubt on the methods.

You can use any presentation software, within reason. You should email your presentation to the instructor by 12:30 on the day of your presentation and should be saved as a PDF (strongly recommended), HTML, LibreOffice, or OpenOffice file. If you create the presentation in PowerPoint, save it as a PDF file.

Grading criteria

The term paper proposal will be graded on five criteria. Each criterion will be graded on a scale of 0 to 4 points, giving a final score between 0 and 20.

- Economic relevance and motivation for the topic
- Feasibility of the topic
- Choice and understanding of the appropriate data source
- Choice and understanding of the empirical/econometric methods
- Link between the topic, dataset, and empirical/econometric methods

The term paper presentation will be graded on five criteria. Each criterion will be graded on a scale of 0 to 4 points, giving a final score between 0 and 20.

- Presentation style and slide organization
- Economic relevance and motivation for the topic
- Choice and understanding of the appropriate data source
- Choice and understanding of the empirical/econometric methods
- Interpretation of the results

The term paper will be graded on seven criteria. Each criterion will be graded on a scale of 0 to 5 points, giving a final score between 0 and 35.

- Writing style and paper organization
- Economic relevance and motivation for the topic
- Choice and understanding of the appropriate data source
- Choice and understanding of the empirical/econometric methods
- Interpretation of the results
- Incorporation of feedback from the presentation
- Clarity and accuracy of the script file

Possible Data Sources

Here is a list of popular datasets used for development economics research. This is not close to comprehensive and do not be worried or offended if your favorite dataset is not included on the list! The list includes some panel/longitudinal datasets but these typically require a major time investment to set up before you can analyze the panel dimension. Use panel/longitudinal data at your own risk. It is very difficult to write a good term paper in this class using time-series data (e.g. annual GDP for a single country in multiple years). You should be very careful about using a dataset if you cannot see the questionnaire and some description about how the dataset was collected.

- The World Bank's data portal includes both microeconomic survey and census data from a large number of countries (http://microdata.worldbank.org/index.php/home) and aggregate country-level data, including the World Development Indicators (http://data.worldbank.org/).
- The DataFirst Research Unit maintains a similar data portal for surveys conducted in (mainly Southern) Africa (http://datafirst.uct.ac.za/).
- The Penn World Tables provide country-level economic and socioeconomic measures dating as far back as 1950 (https://www.rug.nl/ggdc/productivity/pwt/).

- The Living Standards and Measurement Surveys are household surveys conducted in 37 countries since 1985. Almost all can be accessed at https://microdata.worldbank.org/index.php/catalog/lsms. They include measures of many economic and socioeconomic variables, with a focus on assets, consumption, and income. Some countries use slightly different names for these surveys - see examples at the end of this document.
- The Demographic and Health Surveys are household surveys conducted in nearly 100 countries since 1985. Almost all can be accessed at http://www.dhsprogram.com/. The set of measures is slightly narrower than the LSMS surveys but are still very useful.
- There are (roughly) nationally representative panel datasets available for several middle-income countries:
  - Indonesian Family Life Survey at http://www.rand.org/labor/FLS/IFLS.html.
  - Malaysian Family Life Survey at http://www.rand.org/labor/FLS/MFLS.html.
  - Mexican Family Life Survey at http://www.ennvih-mxfls.org/english/index.html.
  - South African National Income Dynamics Study at www.nids.uct.ac.za.
- There are several international education datasets available online:
  - Programme for International Student Assessment (PISA) at https://www.oecd.org/pisa/data/.
  - Southern and Eastern African Consortium for Monitoring Educational Quality (SACMEQ) at http://www.sacmeq.org/sacmeq-data. (The online registration system has problems and may not be accessible this semester.)
  - Trends in the International Mathematics and Science Study (TIMSS) at http://timssandpirls.bc.edu/timss2011/international-database.html.
- The Learning and Educational Acheivements in Punjab Schools is a very usful economic dataset on education in a developing country context. This is available on the World Bank's data portal.
- Several economists maintain lists of development economics datasets:
  - Sebastian Bauhoff at https://scholar.harvard.edu/bauhoff/misc.html.
  - Masayuki Kudamatsu at http://devecondata.blogspot.com/.

Some academic journals also require authors to share their data online at the time of publication. If a particular paper discussed in class or cited in a reading interests you, look at the journal's and the authors' websites to see if the data are available. These datasets are often released only to allow the results in the paper to be replicated, so they may include only a few variables from the original survey.

Finding and understanding a new dataset is a core part of the term paper assignment. You may use a pre-prepared dataset that you have obtained from another faculty member for a class, research position, etc. However, in this case *you must also identify a new dataset that could be used to answer the same or a closely related question.* You must provide a short (2-4 pages) description of the new dataset, where to find it, how the dataset was created, what units are observed and analyzed, and how the key variables are defined. You can submit this with the term paper proposal or with the final term paper.

Resources for Data Analysis

Here is a list of some online guides to that might be useful as you analyze your data and prepare your script

files.

- Duke Center for Data and Visualization Sciences at https://library.duke.edu/data/consulting
- Stata guides online at https://stats.idre.ucla.edu/stata/, https://www.ssc.wisc.edu/sscc/pubs/sfr-intro.htm, and http://www.princeton.edu/ otorres/Stata/.
- Stata listserv where you can ask questions and, more importantly, see if anyone else has already answered the same question at https://www.statalist.org/.
- R guides line at https://www.r-project.org/mail.html and https://online.duke.edu/coursera-for-duke/ (search for R).
- R listserv where you can ask questions and, more importantly, see if anyone else has already answered the same question at https://www.r-project.org/mail.html.

There is a guide to important Stata commands on the course's Sakai website.

Software for Data Analysis

You can download R and Python for free.

You can buy a copy of Stata for $50 from Duke's software shop. This is a 75% discount on the normal student price of Stata.

You can also access Stata for free through Duke's Center for Data and Visualization Science. To do that: install the Duke VPN on your computer, reserve a Stata image at the VCM site, use a RDC client to connect to your Stata image. There's a helpful set of instructions for the whole process in this video. This process is complicated to set up if you're not used to using remote software. But it's very easy to use after setting it up.

Examples of Research Questions from Previous Classes

- Do participants in conditional cash transfer programmes understand the conditions, and do participants who understand the conditions behave differently to participants who do not? Used household survey data from a journal article about a randomised controlled trial in Morocco.
- Does the relationship between children's test scores and their families' socioeconomic status differ in tracked and untracked education systems? Used data from PIRLS, PISA, and TIMSS (acronyms defined in list of datasets).
- Does the relationship between economic growth, governance, and inequality? Used country-by-year data from the World Development Indicators and Country Policy and Institutional Assessment.
- How does perceived well-being differ by race? Used household survey data from South Africa's National Income Dynamics Study.
- How did healthcare spending change after the 2008 financial crisis? Used household survey data from the Tajikistan Living Standards Survey.
- What is the relationship between aid receipt and economic growth? Used country-by-year data from the World Development Indicators, World Governance Indicators, and the AidData research centre.
- What is the relationship between contraceptive use and women's employment? Used household survey

data from the Moldovan Reproductive Health Survey.

- What is the relationship between education and homicide rates? Used country-by-year data from the World Development Indicators.
- What is the relationship between elite running results and GDP per capita? Used country-by-year data from the World Development Indicators and Track and Field News.
- What is the relationship between female autonomy and female self-employment? Used household survey data from the Indian Human Development Survey.
- What is the relationship between handwashing and children's health? Used household survey data from a journal article about a randomised controlled trial in Kenya.
- What is the relationship between job churn and economic growth? Used country-by-year data from the World Development Indicators and OECD.
- What is the relationship between maternal education and infant mortality? Used household survey data from the Guatemalan Survey of Family Health.
- What is the relationship between remittances and household spending? Used household survey data from the Migration and Remittances Household Survey in Burkina Faso, Kenya, Nigeria, Uganda and Senegal
- What is the relationship between renewable energy generation and GDP per capita? Used county-by-year data from the World Development Indicators, International Energy Agency, and Enerdata.
- What is the relationship between women's inheritance rights and household spending patterns? Used household survey data from the Indian Human Development Survey.

Tips for Writing and Presenting Well in This Class

(a) Our goals for writing and presenting

  (a) Clear, precise explanation of complex ideas

  (b) Honest discussion of strengths of weaknesses of the work

(b) Not our goals

  (a) Beautiful formatting

  (b) Complex language

  (c) Over-selling

(c) Implications of these goals

  (a) Think carefully about what you want to communicate and what your reader knows.

  (b) Spend time on writing and rewriting.

  (c) Think carefully about how to display graphs and tables.

  (d) Keep sentences short and simple.

  (e) Dont spend time making beautiful slides or colourful headers.

  (f) Dont oversell your topic or results: You just need to show that the topic matters. You don't need to show that this is the most important topic, or your results are a massive new contribution.

(d) Examples

   (a) Paper introduction

      (a) Assume the reader wont read anything else not true in this class but a good life rule.

      (b) First few paragraphs should make the reader care, but not oversell.

      (c) Then explain what you do data sources, regressions, etc.

      (d) Then explain what you find.

   (b) Data section

      (a) Reader needs to understand exactly what your measures are and where you got them.

      (b) Think carefully about how they interrelate can the reader understand the first measure before you explain the second measure?

(e) How I write

   (a) Keep very short, rough, ugly notes on literature, data, and data analysis while I do them.

   (b) Decide on what the main tables and figures of the paper/presentation.

   (c) Write a paper skeleton with one sentence per planned paragraph. This often ends up as the first sentence of the paragraph.

   (d) Revise steps 2 and 2 repeatedly.

   (e) Write a full draft, quickly and at low quality.

   (f) Edit the full draft.

   (g) Ask a friend or colleague to read it, and tell me what they dont understand

   (h) Repeat step 6 often, sometimes repeat step 7.

Examples of bad, adequate, and good tables, showing the same information. To generate tables like the good example, learn to use commands like **esttab** or **estout** in Stata.

```
      Source |       SS           df       MS      Number of obs   =       699
-------------+----------------------------------   F(4, 694)       =    591.76
       Model |  1053.49623         4  263.374056   Prob > F        =    0.0000
    Residual |  308.876699       694   .44506729   R-squared       =    0.7733
-------------+----------------------------------   Adj R-squared   =    0.7720
       Total |  1362.37292       698  1.95182367   Root MSE        =    .66713

------------------------------------------------------------------------------
         lny | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         lnk |   .6007211   .0123676    48.57   0.000     .5764387    .6250035
          x1 |   .0047493   .0254689     0.19   0.852     -.045256    .0547545
          x2 |   .0199695   .0263162     0.76   0.448    -.0316994    .0716384
          x3 |  -.0261793   .0248847    -1.05   0.293    -.0750377    .0226792
       _cons |  -.3456336   .0915659    -3.77   0.000    -.5254131   -.1658542
------------------------------------------------------------------------------
```

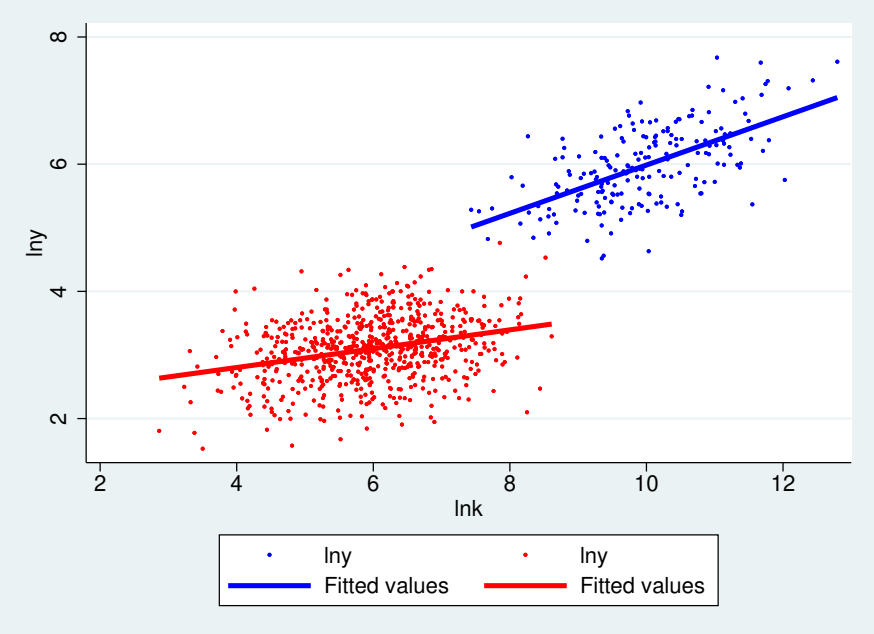
```
      Source |       SS           df       MS      Number of obs   =       699
-------------+----------------------------------   F(4, 694)       =    591.76
       Model |  1053.49623         4  263.374056   Prob > F        =    0.0000
    Residual |  308.876699       694   .44506729   R-squared       =    0.7733
-------------+----------------------------------   Adj R-squared   =    0.7720
       Total |  1362.37292       698  1.95182367   Root MSE        =    .66713

------------------------------------------------------------------------------
         lny | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         lnk |   .6007211   .0123676    48.57   0.000     .5764387    .6250035
          x1 |   .0047493   .0254689     0.19   0.852     -.045256    .0547545
          x2 |   .0199695   .0263162     0.76   0.448    -.0316994    .0716384
          x3 |  -.0261793   .0248847    -1.05   0.293    -.0750377    .0226792
       _cons |  -.3456336   .0915659    -3.77   0.000    -.5254131   -.1658542
------------------------------------------------------------------------------
```

	(1)	(2)	(3)	(4)	(5)
	Log output pc	Log output pc	Log output pc	Log output pc	Log output pc
Log capital pc	0.582***	0.591***	0.594***	0.601***	0.601***
	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)
Primary school		-0.008	-0.004	0.005	0.005
completion rate		(0.023)	(0.024)	(0.025)	(0.025)
Life expectancy			0.019	0.020	0.020
			(0.026)	(0.026)	(0.026)
Unemployment rate				-0.026	-0.026
				(0.025)	(0.025)
Constant	-0.255**	-0.292***	-0.308***	-0.346***	-0.346***
	(0.082)	(0.086)	(0.089)	(0.092)	(0.092)
Observations	900	799	749	699	699

Table shows results from regressing log output per capita on log capital per capita and selected control variables, using a sample of 180 countries from 2016 to 2020. The sample size changes between columns due to missing values for some covariates. Data are sourced from the World Development Indicators. Heteroskedasticity-robust standard errors are shown in parentheses. */**/*** denote statistical significance at the 10/5/1% level.

Examples of bad and good figures, showing the same information. To generate figures like the good example, learn to use commands like **graph twoway** in Stata and **ggplot** in R.



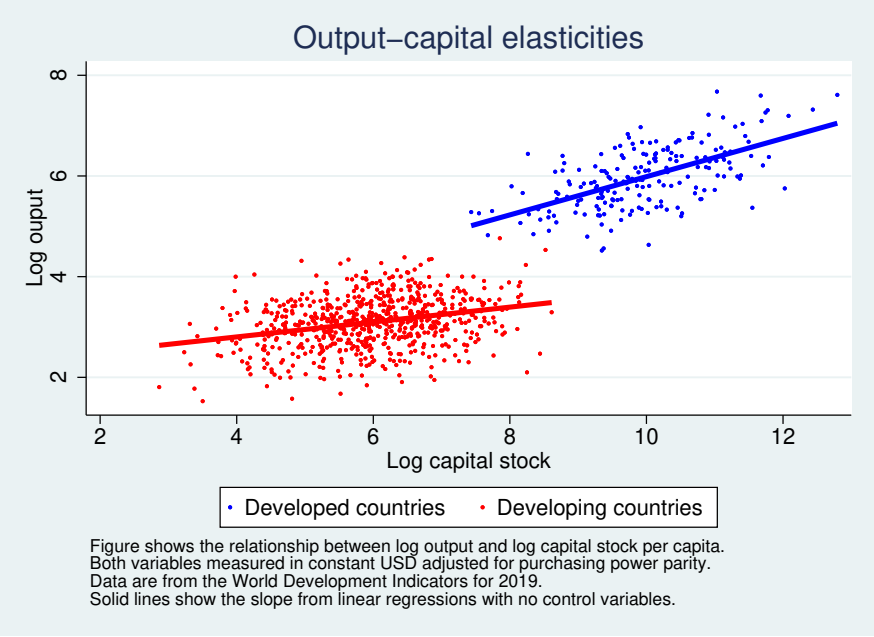


Figure shows the relationship between log output and log capital stock per capita.
Both variables measured in constant USD adjusted for purchasing power parity.
Data are from the World Development Indicators for 2019.
Solid lines show the slope from linear regressions with no control variables.