DLSU Business & Economics Review 33(1) 2023 p. 26–33

RESEARCH ARTICLE

A Pedagogical Note on Linear Regressions

Andrew Adrian Yu Pua
Xiamen University, Xiamen, Fujian, China
andrewypua@outlook.com

JEL Classification: A2, C01

I write this note to pay tribute to Dr. Tereso Tullao's efforts to democratize economic education in the Philippines. In the main part of the article, I emphasize the fact that linear regressions are data summaries of a particular form, without resorting to too much mathematical burden on the part of the students. An advantage of the presented approach is that it becomes easier to get students "onboard" and get them to run their own data analysis as fast as possible, while still being able to peek into the black box of linear regression and to gain confidence in their own understanding of the material. I encourage other instructors to incorporate the presented materials into their own teaching.

Looking back and taking stock

Dr. Tereso Tullao, Jr. was the instructor for the Development Economics class (ECONDEV) for economics majors. I still remember our small class in the seminar room on the second floor of the LS Building. I recall an article on absorptive capacity, and the reading was from an issue of the Journal of Development Economics (JDE). At that time, I was already savvy enough to use electronic databases like EBSCOHost. But the article was not available electronically, so it was the first time I had ever accessed the physical copies of issues from JDE.

Most people fondly remember Dr. Tullao as an animated and engaging instructor. I do not remember a lot of back-and-forth discussions during the ECONDEV classes, but it did push me into reading more scholarly articles and reduce my reliance on textbooks. I have also observed and tried to engage with his philosophy of democratizing the teaching of economics (like his Filipino language principles textbook, *Mga Prinsipyo sa Ekonomiks*) to almost everyone. It is unfortunate that I am not the best Filipino native speaker--there are times I think that *tuhod* is the hind leg! If I were any better in the native language, I would have drafted this article in Filipino.

In deciding what to do for his festschrift, I initially wanted to write a retrospective on his oeuvre in international economics and development. But my expertise is lacking in these areas. I also wanted to engage more on the topic of the intellectualization of the Filipino language, especially given my teaching experience in different countries and exposure to a wide variety of international students, including Filipinos. My earliest exposure to intellectualization stems from Dr. Teresita Fortunato's class on *Filipino Para sa Iba't Ibang Disiplina* (FILIPI2 at that time). Looking back, I was intrigued by this area, but I must still resolve in my mind the issue of as to whether the Filipino language is already intellectualized or whether a more inclusive pidgin language would be the more appropriate choice of language for educational purposes. But again, I must devote years of study to meaningfully engage with this topic further.

When I was a faculty member at what used to be the Economics Department at De La Salle University, I was part of the fixed thesis committee for a sub-cohort of undergraduate students writing their theses. The members included Dr. Tullao, Dr. Mitzie Irene

P. Conchada, and myself. About fifteen years ago, Dr. Tullao and I had more of an exciting back-and-forth discussion on statistical significance. Things have not changed much over the past 15 years, and statistical significance and reproducibility of findings are re-emerging themes in our current research environment (see Volume 73, Issue sup1 of *The American Statistician* in 2019, an entire issue devoted to the subject, along with the statement by the task force appointed by the president of the American Statistical Association (Benjamini et al. 2021)). But there is already a lot of material available on this topic, despite the abundance of confusion.

In the end, I decided to write about something more compatible with Dr. Tullao's efforts to democratize economics education. Since my expertise is in econometrics, I am going to focus on linear regressions which, for better or worse, are the bread and butter of both theoretical and applied work in economics. I want to demystify and deepen the understanding of linear regression by laying bare its core to high school students, undergraduates, graduates, and even research professionals, without relying on calculus and matrix algebra. I expand on the material I have developed in teaching students with extremely heterogeneous backgrounds (Pua, 2022).

In this article, I will illustrate a major, but underappreciated, point: different linear regressions you calculate on the same dataset are mutually compatible. Without additional information (such as assumptions, biases, prejudices, thoughts, and details about the context), we cannot rule out any of those linear regressions a priori. In fact, highlighting one (or even a subset) of these linear regressions implicitly suggests a model in the mind of the person communicating the results.

What I am writing about here is not entirely new; Malinvaud (1970) has a first chapter dedicated to "*Econometrics without stochastic models*". The treatment of this topic requires a good intuitive grasp of linear algebra. But I avoid linear algebra and use the computer for illustration instead. I also limited the mathematics to only the summation notation.

All analyses were performed using R Statistical Software v4.2.2 (R Core Team 2022), and tables are displayed using `stargazer` v5.2.3 (Hlavac 2022). To reproduce the computer output, you need to load the `TeachingRatings` dataset from Stock and Watson (2012) and install R along with the packages `foreign` and `stargazer`.

```
ratings <- foreign::read.dta("TeachingRatings.dta")

with(ratings, c(mean(course_eval), tapply(course_eval,

female, mean)))

                0        1

3.998272 4.069030 3.901026
```

Linear regression is a data summary

At their core, linear regressions are not fundamentally different from usual summaries of the data. The issue is to demystify how these summaries are made. Let me illustrate using an extract by Stock and Watson (2012) from a course evaluation dataset collected by Hamermesh and Parker (2005). The unit of analysis is the course given by an instructor. Course evaluations (`course_eval`) are on a scale of 1 to 5, and the sex of the instructor (`female`) is coded as 1 if the instructor of course is female and 0, if male.

With this kind of data, a natural starting point is to look at the overall mean and the group means.

Refer to the computer output in the previous page. The first number is the overall mean, the second number is the mean of course evaluations when female is equal to 0 (meaning the instructor is male), and the third number is the mean of course evaluations for females. How does linear regression connect with these summaries?

I will now compute two sets of linear regressions using the built-in R function lm(). Below you may find the code and the table presenting the results.

```
reg1 <- lm(course_eval ~ 1, data = ratings)

reg2 <- lm(course_eval ~ female, data = ratings)

stargazer::stargazer(reg1, reg2, type = "text", style=

"aer",

                     report=  "vc", omit.stat=  "all",

no.space = FALSE,

                     column.labels = c("reg1", "reg2"),

                     omit.table.layout = "n", model.numbers

= FALSE)

=====================

         course_eval

         reg1   reg2

---------------------

female         -0.168


Constant 3.998 4.069

---------------------
```

Notice how the numbers reported in the table and the preceding result are tied together. `reg1` is the linear regression of course evaluations with an intercept (coded `1`, which is a column of ones if you imagine a spreadsheet) and the results reproduce the overall mean of course evaluation. `reg2` is the linear regression of course evaluations on the sex of the instructor, including an intercept. The reported constant reproduces the mean of course evaluations for male instructors. The reported coefficient for female is a difference of means. It compares the average for females (3.901) to the average of males (4.069) in a particular way: (3.901 - 4.069 = -0.168). In other words, female instructors score about 0.17 points lower than male instructors on average.

Both sets of linear regressions `reg1` and `reg2` are compatible with each other. There is no reason to prefer one linear regression over the other, unless we bring in either our own views of the world, some assumptions, our biases, and so on. Because linear regression in this context produces only a comparison of averages, it should not be surprising that some female instructors can score higher than male instructors.

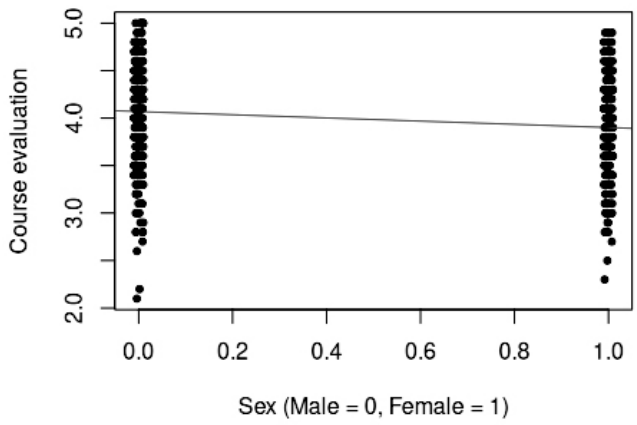
The scatterplot presents a more complete picture of what is happening. The red line is a visualization of the linear regression. For convenience, let Y_t represent the tth course evaluation and X_{1t} the sex of the instructor teaching the tth course. The equation of the red line is given by:

$$\hat{Y}_t = 4.069 - 0.168 * X_{1t}$$

The notation $\hat{Y}_t$ means that the reported linear regression is not about the actual value of course evaluations; rather it is about the fitted value of course evaluations. Of course, the actual values Y_t are used to obtain the reported results. We could calculate a fitted value for each instructor. The difference between the actual value and the fitted value is called the residual. These objects have the following algebraic properties by design:[1]

1. The average of the fitted values is equal to the average of the actual values of *Y*.
2. The average of the residuals is equal to zero.

```
with(ratings, plot(jitter(female, factor = 0.05),
course_eval,
                   xlab = "Sex (Male=  0, Female = 1)",
                   ylab = "Course evaluation", pch=  20))
abline(reg2, col=  "red")
```



Short and long regressions

There is another common theme shared by these results. Refer to `reg1` and `reg2` once more. In both results, you have two different intercepts with different interpretations but which are mutually compatible descriptions of the data.

To dig deeper into how they are algebraically related, we look at a simpler version of the setup of short and long regressions. `reg1` is the short regression, while `reg2` is the long regression. The terminology arises from the longer list of regressors in `reg2`, namely 1 and `female`, compared to `reg1` (which is just `1`). Let $\tilde{Y}_t = \tilde{\beta}_0$ be the short regression and let $\widehat{Y_t} = \widehat{\beta_0} + \widehat{\beta_1} X_{1t}$ be the long regression. We obtain $\widehat{\beta_0}$, $\widehat{\beta_0}$, and $\widehat{\beta_1}$ from an appropriate use of `lm()`. Note further that $\tilde{Y}_t$ and $\widehat{Y_t}$ are different fitted values. How is $\widehat{\beta}_0$ related to $\tilde{\beta}_0$ algebraically?

We can obtain a linear regression of X_1 on a column of ones, i.e., $\hat{X}_{1t} = \hat{\delta}_0$. Here we obtain $\hat{\delta}_0$ from an appropriate use of `lm()`. Next, X_{1t} differs from the fitted value $\hat{X}_{1t}$ by a residual. Therefore,

$$\widehat{Y_t} = \widehat{\beta_0} + \widehat{\beta_1} X_{1t} = \widehat{\beta_0} + \widehat{\beta_1}\left(X_{1t} - \widehat{X_{1t}}\right) + \widehat{\beta_1}\widehat{X_{1t}} = \widehat{\beta_0} + \widehat{\beta_1}\left(X_{1t} - \widehat{X_{1t}}\right) + \widehat{\beta_1}\widehat{\delta_0}$$

Taking averages of both sides of the previous equation, we have

$$\frac{1}{n}\sum_{t=1}^{n} \widehat{Y_t} = \widehat{\beta_0} + \widehat{\beta_1}\frac{1}{n}\sum_{t=1}^{n}\left(X_{1t} - \widehat{X_{1t}}\right) + \widehat{\beta_1}\widehat{\delta_0}$$

$$\Rightarrow \frac{1}{n}\sum_{t=1}^{n} \widehat{Y_t} = \widehat{\beta_0} + \widehat{\beta_1}\widehat{\delta_0} \Rightarrow \overline{Y} = \widehat{\beta_0} + \widehat{\beta_1}\widehat{\delta_0}$$

In the second equality, we used the property that residuals should sum up to zero. In the third equality, we used the property that the average of the fitted values is equal to the average of the actual values of . Similarly, we also have

$$\frac{1}{n}\sum_{t=1}^{n} \tilde{Y}_t = \tilde{\beta}_0 \Rightarrow \overline{Y} = \tilde{\beta}_0$$

Since the average of the actual values of *Y* is a unique value, then we must have

$$\tilde{\beta}_0 = \hat{\beta}_0 + \hat{\beta}_1\hat{\delta}_0$$

This algebraic relationship has a very neat (which may be surprising to some) interpretation in our course evaluation example. The overall mean $\overline{Y}$ or $\widetilde{\beta_0}$ is equal to the average for males $\widehat{\beta_0}$ plus the proportion of females $\widehat{\delta_0}$ multiplied by the difference in the average for females and the average for males $\widehat{\beta_1}$. Not surprisingly, this is also the same as a weighted average: the proportion of females multiplied by the average for females and the proportion of males multiplied by the average for males.

But, more importantly, this algebraic relationship tells us that different sets of linear regressions are connected to each other in a very particular way. Furthermore, the connections depend on what you put into your list of regressors. I do not illustrate any further; but I mention that the algebraic relationships which connect the short

regression with the long regression holds in the general case. In fact, these algebraic relationships form the basis for discussing omitted variable biases and sensitivity of regression results (for example, see Cinelli and Hazlett (2020).

Sure, but the regressor is discrete.

In our context, the regressor `female` takes on only two values and would be classified as a discrete-valued regressor. A fair question to ask is whether what I discussed in this article holds in the case of continuous-valued regressors. The answer is yes but with some conceptual differences:

1. When `female` equal to zero, it points to a specific observable subgroup of the data. In the continuous case, finding a regressor taking on a value exactly equal to zero might be difficult. In addition, that zero value might not even make sense! Therefore, the interpretation of the constant in the regression tables and results might not even make sense.
2. The difference of means analogously extends to the continuous case. But now you are comparing means of subgroups of the data which might not necessarily have an observable counterpart. Regression lines use their shape to "fill up" or extrapolate to those subgroups.

I will elaborate on the second point. For example, the `TeachingRatings` dataset includes a measure of the "beauty rating" of the instructor. We could compute the linear regression of course evaluations on beauty rating. Finding two subgroups whose beauty ratings differ by exactly 1 point may or may not be possible. As you noticed, the comparison will be based on a vanishingly small subset of the data. In some sense, some of these subgroups may be considered hypothetical.

Wrapping it all together using some algebra

Recall that our linear regression was $\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_{1t}$. In the context of this article, it is possible to derive the regression slope $\widehat{\beta}_1$, along with its alternative forms:

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n}\left(X_{1t} - \overline{X}_1\right)\left(Y_t - \overline{Y}\right)}{\sum_{t=1}^{n}\left(X_{1t} - \overline{X}_1\right)^2} = \frac{\sum_{t=1}^{n}\left(X_{1t} - \overline{X}_1\right)Y_t}{\sum_{t=1}^{n}\left(X_{1t} - \overline{X}_1\right)^2} = \sum_{t=1}^{n} w_t\, Y_t$$

where

$$w_t = \frac{\left(X_{1t} - \overline{X}_1\right)}{\sum_{t=1}^{n}\left(X_{1t} - \overline{X}_1\right)^2}$$

The w_t's are available for every observation and serve as "weights" which combine with the actual values Y_t. More importantly, the w_t's depend only on the regressors. Furthermore, it depends on how far the tth observation of X_{1t} is from its mean, how spread out *all* the observations of X_{1t} from its mean, and the total number of observations available for X_{1t}.

In our course evaluation example, things are simpler to present because we only have two values for the : one for male instructors (268 of them) and another one for female instructors (195 of them). Observe that multiplying each of the weights with the corresponding course evaluation and then adding them up together exactly matches the difference of means we have obtained earlier.

```
weights <- with(ratings, (female -

mean(female))/sum((female-mean(female))^2))

table(weights)

weights

-0.00373134328358209  0.00512820512820513

                2681                   95

with(ratings, sum(weights*course_eval))

[1] -0.1680042
```

These weights are also related to leverage points in regression analysis. Traditionally, these leverage points help detect influential observations in a regression analysis. Recent research on leverage points is connected to which standard errors to use when reporting results of a regression analysis (Hansen, 2022).

Concluding remarks

In this note, I present some materials about linear regression that other instructors could use in their classes. I think that the note could be used for self-study as well. I hope that the use of open source statistical software along with minimal mathematical knowledge is more than enough to democratize the teaching of linear regressions. I avoid the baggage of introducing econometric models, as models can distract (if not well-understood) and often adds unnecessary mystery to a commonly used tool. Should the student desire to pursue econometric models, seeing how linear regressions work can help shed light on what linear regressions help us learn about those models.

Notes

[1] There is an exception when it comes the case of regression without an intercept, but this is very rare in practice.

References

Benjamini, Yoav, De Veaux, Richard D., Efron, Bradley, Evans, Scott, Glickman, Mark, Graubard, Barry I., He, Xuming et al. (2021). The ASA president's task force statement on statistical significance and replicability. *The Annals of Applied Statistics, 15* (3), 1084–85. https://doi.org/10.1214/21-AOAS1501.

Cinelli, Carlos, and Hazlett Chad. (2020). Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82* (1),39–67.

Hamermesh, Daniel S., and Parker, Amy. (2005). Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity. *Economics of Education Review, 24* (4), 369–76.

Hansen, Bruce E. (2022). Jackknife Standard Errors for Clustered Regression. Working paper. University of Wisconsin.

Hlavac, Marek. (2022). *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Social Policy Institute. https://CRAN.R-project.org/package=stargazer.

Malinvaud, E. (1970). *Statistical Methods of Econometrics*. Second revised. North-Holland Publishing Company.

Pua, Andrew Adrian. (2022). Lecture Materials for Applied Econometrics 1 (2022 Version). https://applied-metrics.neocities.org/.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stock, James H., and. Watson, Mark M. (2012). *Introduction to Econometrics*. Pearson.