The Linear Model and OLS

IRIII.2 – Quantitative Methods in the Study of International Relations

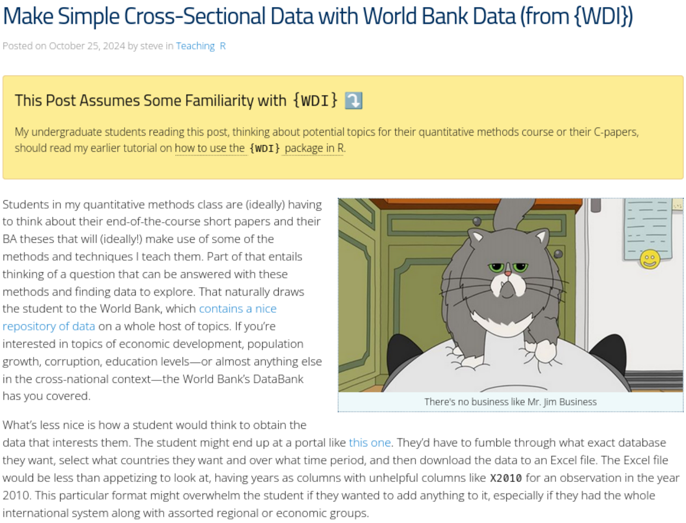
Steven V. Miller

Department of Economic History and International Relations

Goal(s) for Today

1. Introduce the basic intuition behind linear regression.
2. Give an applied example and unpack it.

Elsewhere on My Blog



Make Simple Cross-Sectional Data with World Bank Data (from {WDI})

Posted on October 25, 2024 by steve in Teaching R

This Post Assumes Some Familiarity with `{WDI}` ⤵️

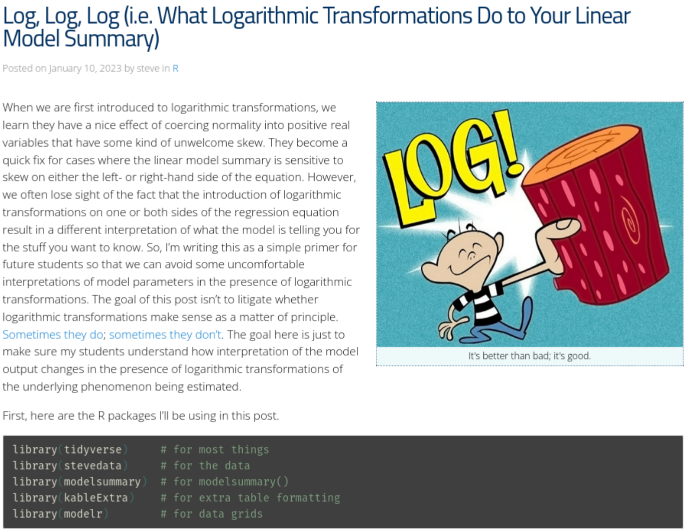
My undergraduate students reading this post, thinking about potential topics for their quantitative methods course or their C-papers, should read my earlier tutorial on how to use the `{WDI}` package in R.

Students in my quantitative methods class are (ideally) having to think about their end-of-the-course short papers and their BA theses that will (ideally!) make use of some of the methods and techniques I teach them. Part of that entails thinking of a question that can be answered with these methods and finding data to explore. That naturally draws the student to the World Bank, which contains a nice repository of data on a whole host of topics. If you're interested in topics of economic development, population growth, corruption, education levels—or almost anything else in the cross-national context—the World Bank's DataBank has you covered.

There's no business like Mr. Jim Business

What's less nice is how a student would think to obtain the data that interests them. The student might end up at a portal like this one. They'd have to fumble through what exact database they want, select what countries they want and over what time period, and then download the data to an Excel file. The Excel file would be less than appetizing to look at, having years as columns with unhelpful columns like `X2010` for an observation in the year 2010. This particular format might overwhelm the student if they wanted to add anything to it, especially if they had the whole international system along with assorted regional or economic groups.

Elsewhere on My Blog



Log, Log, Log (i.e. What Logarithmic Transformations Do to Your Linear Model Summary)

Posted on January 10, 2023 by steve in R

When we are first introduced to logarithmic transformations, we learn they have a nice effect of coercing normality into positive real variables that have some kind of unwelcome skew. They become a quick fix for cases where the linear model summary is sensitive to skew on either the left- or right-hand side of the equation. However, we often lose sight of the fact that the introduction of logarithmic transformations on one or both sides of the regression equation result in a different interpretation of what the model is telling you for the stuff you want to know. So, I'm writing this as a simple primer for future students so that we can avoid some uncomfortable interpretations of model parameters in the presence of logarithmic transformations. The goal of this post isn't to litigate whether logarithmic transformations make sense as a matter of principle. Sometimes they do; sometimes they don't. The goal here is just to make sure my students understand how interpretation of the model output changes in the presence of logarithmic transformations of the underlying phenomenon being estimated.

It's better than bad; it's good.

First, here are the R packages I'll be using in this post.

```
library(tidyverse)     # for most things
library(stevedata)     # for the data
library(modelsummary)  # for modelsummary()
library(kableExtra)    # for extra table formatting
library(modelr)        # for data grids
```

Correlation to Linear Regression

Correlation has a lot of nice properties.

- It's another "first step" analytical tool.
- Useful for detecting **multicollinearity**.
  - This is when two independent variables correlate so highly that no partial effect for either can be summarized.

However, it's neutral on what is *x* and what is *y*.

- It won't communicate cause and effect.

Fortunately, regression can do that for us (under ideal conditions).

Demystifying Regression

Does this look familiar?

$$y = mx + b$$

Demystifying Regression

That was the slope-intercept equation.

- b is the intercept: the observed y when $x = 0$.
- m is the familiar “rise over run”, measuring the amount of change in y for a unit change in x.

Demystifying Regression

The slope-intercept equation is, in essence, the representation of a regression line.

- However, statisticians prefer a different rendering of the same concept measuring linear change.

$$y = a + b(x)$$

The *b* is the **regression coefficient** that communicates the change in *y* for each unit change in *x*.

- However, this is a deterministic function. We live in a stochastic world.

A Full Statement of the Regression Formula

If you've followed that, we're just going to add two more things:

$$\hat{y} = \hat{a} + \hat{b}(x) + e$$

...where:

- $\hat{y}$, $\hat{a}$ and $\hat{b}$ are estimates of *y*, *a*, and *b* over the data.
- *e* is the error term.
    - It contains random sampling error, prediction error, and predictors not included in the model.

We can further extend this out by including more *x* variables into our equation.

- Mechanically: there's a lot to unpack. Conceptually: not really (at this level).

Getting a Regression Coefficient

How do we get a regression coefficient for more complicated data?

- Start with the **prediction error**, formally: $y_i - \hat{y}$.
- Square them. In other words: $(y_i - \hat{y})^2$
    - If you didn't, the sum of prediction errors would equal zero.

The regression coefficient that emerges minimizes the sum of squared differences ($(y_i - \hat{y})^2$).

- Put another way: "ordinary least squares" (OLS) regression.

An Applied Example: The Correlates of Tourism

Figure 1: I'll be honest that a trip to the Canaries sounds mighty nice in February...

An Applied Example: The Correlates of Tourism

Academic interest in the international relations/IPE of tourism may owe to H.P. Gray (1966).

- Tourism as a form of “soft power” projection.
    - (i.e. you’ve seen the YouTube ads for Türkiye or Arsenal’s “Visit Rwanda” patch)
- Tourism-dependent countries have *a lot* of subtle risks.
    - Behave like rentier states; economy is dependent on foreign receipts.
    - Environmental sustainability concerns are paramount.
    - *Very* sensitive to massive shocks like the pandemic or political violence.

Thus, tourism is a kind of barometer or measuring stick for important questions of development and peace.

- *ed. notice how I’m selling you on the value of doing this in the first place...*

A Simple Exercise

I gathered data from the World Bank's repository for a simple cross-sectional exercise.

- *DV*: number of arrivals for international tourism (logged)
- *IVs*: conceptually (operationally) `[expected effect]`
    - Ease of access/infrastructure for visitors (passengers carried by air transport, logged) `[+]`
    - Relative price to USD (price level ratio of PPP to market exchange rate) `[-]`
    - Economic development (GDP per capita, logged) `[+]`
    - Political security (political stability/absence of violence, terrorism) `[+]`
    - Dummy variable for "fragile/conflict affected states (FCAS)" `[-]`

I use the same lag and group-by fill I describe in the blog post I reference earlier.

- Referent year: 2019 (or shortly before it)

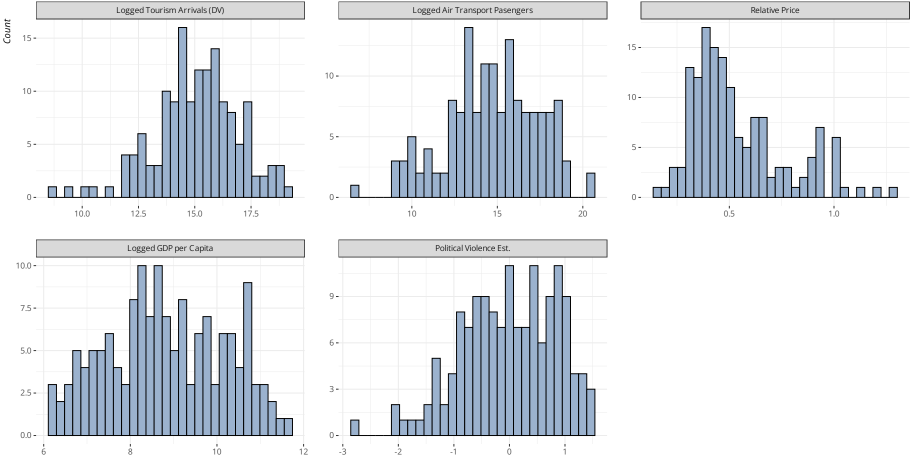
The Model(s)

We'll run two linear models.

1. Bivariate model with just relative price.
2. Full model with all the other stuff.

Faceted Histograms of the Important Variables

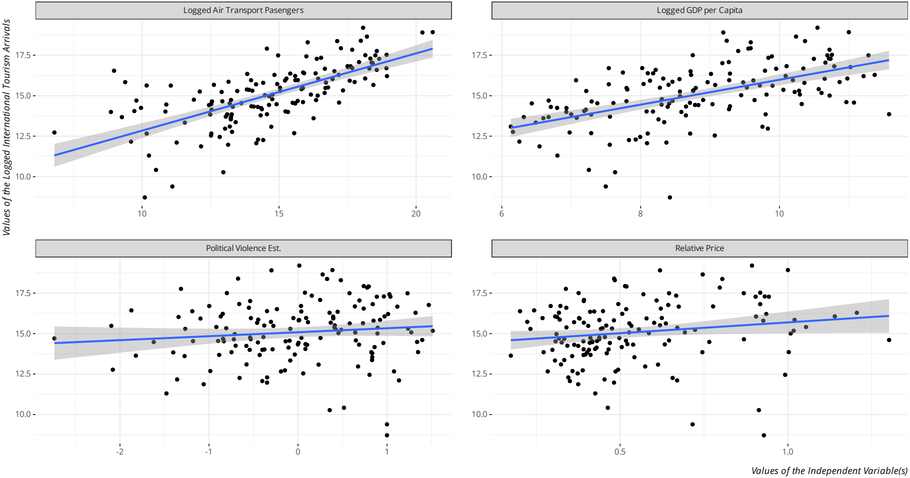
The price variable has relatively few countries that are pricier than the U.S. Some Oceania countries have few visitors and Syria was a conspicuously unsafe country at this time.



Data: World Bank Data repository, through various other places.

Faceted Scatterplots of Bivariate Relationships

All relationships are positive. The, the relative price correlation is a bit surprising (prima facie). In all cases, the line drawn is the OLS line for a simple bivariate model.

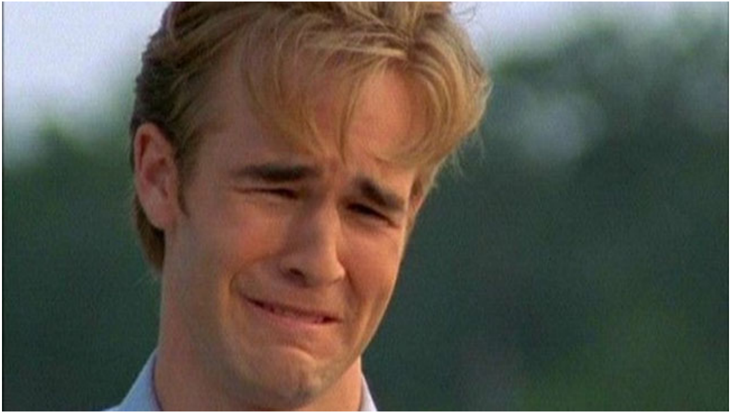


Data: World Bank Data repository, through various other places.

Table 1: Cross-National Correlates of (Logged) International Tourism Arrivals

	Bivariate	**Full Model**
Relative Price	1.333*	-1.565*
	(0.664)	(0.671)
Air Transport Passengers (Logged)		0.283***
		(0.047)
GDP per Capita (Logged)		0.763***
		(0.136)
Political Stability		-0.631***
		(0.182)
Fragile/Conflict-Affected State		-1.360***
		(0.366)
Intercept	14.358***	5.216***
	(0.392)	(0.911)
Num.Obs.	149	149
R2	0.027	0.627
R2 Adj.	0.020	0.614
F	4.026	48.172

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$



How to Interpret a Regression Table Like This

1. Find the variable(s) of interest.
2. Look for direction (positive/negative)
3. Look for “stars” (to determine statistical significance)

Table 2: Cross-National Correlates of (Logged) International Tourism Arrivals

	Bivariate	Full Model
Relative Price	**1.333***	**-1.565***
	(0.664)	**(0.671)**
Air Transport Passengers (Logged)		**0.283*****
		(0.047)
GDP per Capita (Logged)		**0.763*****
		(0.136)
Political Stability		**-0.631*****
		(0.182)
Fragile/Conflict-Affected State		**-1.360*****
		(0.366)
Intercept	14.358***	5.216***
	(0.392)	(0.911)
Num.Obs.	149	149
R2	0.027	0.627
R2 Adj.	0.020	0.614
F	4.026	48.172

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Being More Careful with Our Takeaways

“Number goes (up/down); other number goes (up/down); has (no) stars” is fine when you’re getting started.

- But let’s do more.

We need to be smart with how we communicate this.

- Our DV is log-transformed and so are a few of our IVs.

Our plan of attack:

1. Start with the two variables that are log-transformed.
2. Talk about the variables that aren’t log-transformed.

The DV and IV are Both Log-Transformed

Air transport passengers: a 1% increase in IV coincides with estimated .283% increase in DV.

- Alternatively, less helpfully: a unit increase log(x) increases log(y) by an estimated .283.

GDP per capita: a 1% change in GDP per capita -> .763% change in tourism arrivals.

- Alternatively, with more precision: `(1.01^(.763)-1)*100` = .762% change in tourism arrivals for 1% increase in GDP per capita.

Be mindful of the percentages!

- “one percent change in *x* -> estimated (regression coefficient)% change in *y*”

The DV is Log-Transformed, but the IVs are Not

Relative price (Full): a change from 0 to 1 in relative price -> est. -1.56% decrease in tourism arrivals.

- Be mindful: 0 = conceptual extreme; 1 = same price level as the U.S.

Political stability: a change from 0 to 1 in political stability -> -.631% change in tourism arrivals.

- Alternatively: `exp(-.631)` = .532. A one-unit change in stability multiplies expected tourism arrivals by .532.
- This interpretation works the same way for relative price because it's not log-transformed.

FCAS: Being a FCAS (e.g. Sudan) versus not being one (e.g. Sweden) decreases est. international tourism arrivals by est. 1.36%.

Here: unit changes in x -> (regression coefficient)% changes in y

Dont’ Read Much into the Intercept

Don’t bother interpreting the intercept.

- Nominally, it tells you the estimated value of *y* when *every* covariate is set to 0.
- In our case: a country has no air passengers, is *infinitely* cheaper than the U.S., has no logged GDP per capita, has basically a middle level of political security and is not a FCAS.

There are advanced things you can do here, but don’t bother at this stage.

- Just know what this is ultimately communicating.

The Goodness of Fit Statistics

R-square: proportion of variation in *y* accounted for in the model.

- In the bivariate model, it's quite literally Pearson's *r*, squared.

Adj. R-square: Includes a downweight for more (redundant) parameters.

- Consider this the "default" R-square for your linear model.
- The more junk you have in the model, the greater the separation between it and R-square.

F: "overall model fit" against guessing the mean.

- I'm including this because it's in the output. You're free to ignore it.

Assumptions and Diagnostics

There are *myriad* assumptions of OLS. The ones we impress on you:

- `L`: the outcome *y* model is a “l”inear (and additive) function of the regressors.
- `I`: the error term is “i”ndependent/not correlated across observations.
- `N`: the error term is “n”ormally distributed.
- `E`: the error term has “e”qual/constant variance (i.e. no heteroskedasticity).

Let me bore you with `L` and `N` at this round.

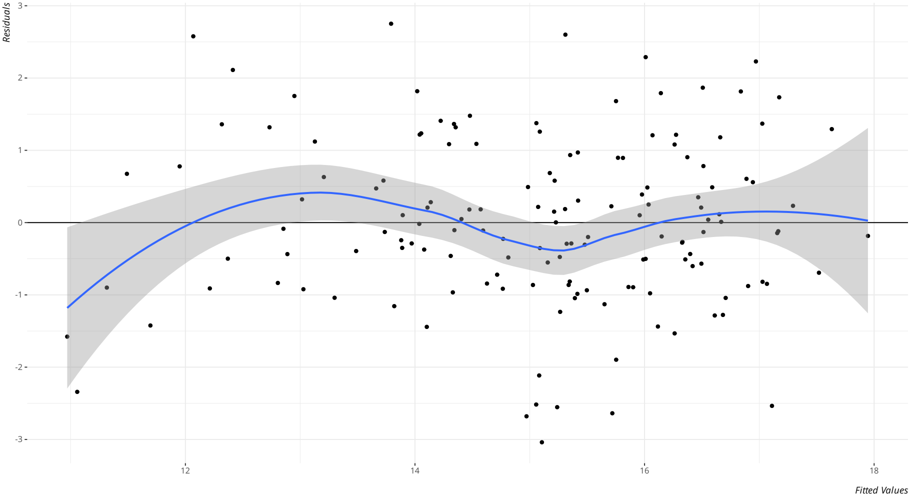
- Save `E` for me tormenting you in the C-paper stage.
- `I` will matter a great deal for more advanced uses (i.e. MA-level).

Diagnostics You Should Run

1. Fitted-residual plot (overall, and/or by regressor)
    - This is the most “bang for your buck” OLS diagnostic. You should always run it.
2. Residual density plot (or QQ plot)
    - Useful for the `N` part, for as unimportant as that assumption mostly is.

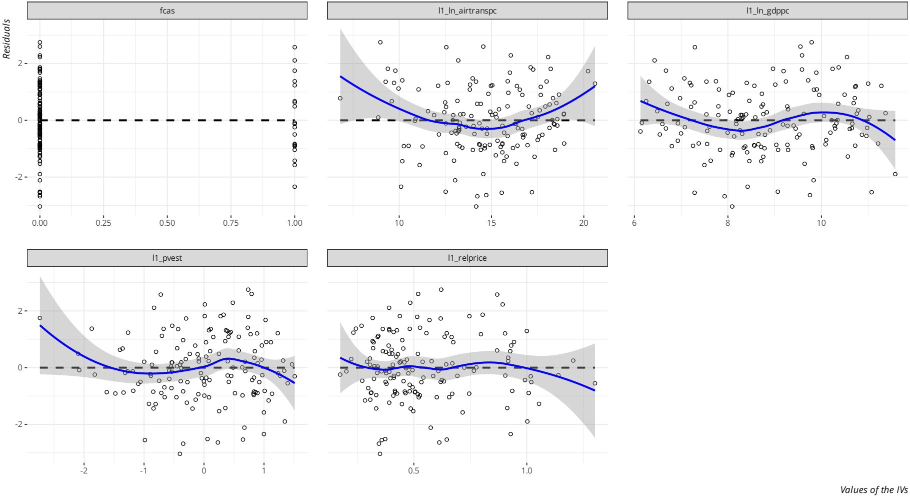
A Fitted-Residual Plot of Our Full Model

The linear line is flat at 0 by definition. Ideally a LOESS smoother agrees with it, but does not here. I don't see a heteroskedasticity concern, fwiw.



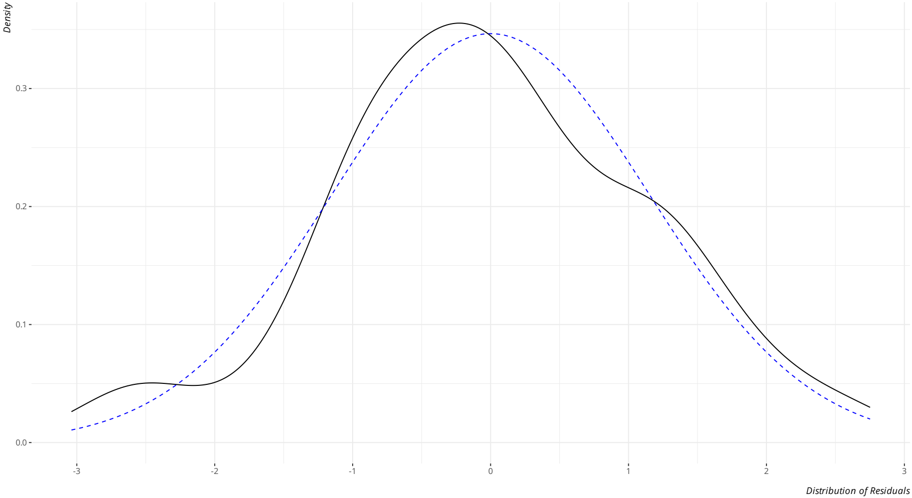
A Fitted-Residual Plot Won't Tell You Where Non-Linearity Is; This Plot Can Help

At most, I see some issues in the air passengers variable. Tail observations will do what they do.



A Residual Density Plot Can Assess Whether Your Residuals Approximate a Normal Distribution

In practice, this is the least important of OLS' major assumptions. It doesn't hurt to look, though.



Conclusion

There's *a lot* I crammed into this lecture, but:

- If you remember the slope-intercept equation, the intuition behind linear regression isn't much.
- OLS gives you the line of best fit that minimized squared prediction errors.
- You gotta get comfortable interpreting regression output.
- Logarithmic transformations proportionalize changes on their raw scale.
  - This might take some getting used-to, but you should know it.
  - I'll extend grace as you get started, but do wrestle with it. Economists definitely do.
- Some parts/assumptions of the linear model are more important than others.

Table of Contents

Introduction

The Linear Model and OLS

- Demystifying, with the Quickness
- An Applied Example: The Correlates of Tourism
- How to Interpret Regression Output
- The Intercept and Goodness of Fit Notes
- Assumptions and Diagnostics

Conclusion