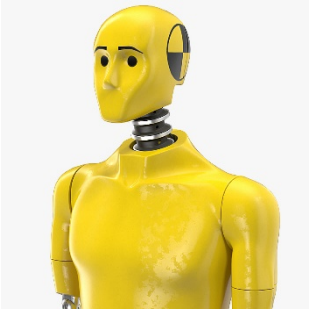
Working with Dummies: Introduction

1) Why are dummy variables called "Dummies"? I have no idea. Maybe it's because only dummies use them? … or maybe it's because you're a dummy not to use them? … or maybe both… or maybe… or maybe. Who knows? But one thing we do know is that they are pervasive in econometrics, and can be extraordinarily useful and powerful analytic tools. So get over the name, and embrace the Dummy![1]



What is a Dummy?

2) A Dummy variable is a binary categorical (or indicator) variable, which takes on one of two values, which are almost always 0 and 1, depending on whether or not an observation falls into a particular category, or not. Typically, a value of 1 indicates the occurrence/presence of a category, event, outcome, characteristic, or thing… or perhaps that a logical statement is TRUE. And the value of 0 indicates the absence of such.

3) In OLS models, dummies capture average differences across categories *controlling for everything else in the model*, and allow you to say things like: controlling for everything else in the model, on average, prices of [insert category 1] are this much higher or lower than prices of [insert category 0].

Dummies in Action, Already!

4) At this point in the semester, you've already seen a bunch of dummies in action:

   a) *AppleMusic* dummy (Spotify v. iTunes… w/ AppleMusic): the estimated AppleMusic dummy coefficient captures the differences in weekly iTunes sales when AppleMusic is on the scene (or not), *controlling for everything else in the model*. It provides an estimate of the impact of AppleMusic on iTunes sales, controlling for… .

   b) *Eurozone* dummy (Sovereign Debt ratings): the estimated Eurozone dummy coefficient captures the average differences in NSRates for Eurozone countries,

---

[1] Merriam-Webster says that the phrase first appeared in 1957. Perhaps we should give credit to Daniel Suits' paper, *Use of Dummy Variables in Regression Equations*, Journal of the American Statistical Association 52 (280), 1957, pp. 548-51. In his paper, Suits discussed what we now call the *Dummy Variable Trap*, even though it isn't really a trap at all. (see below)

relative to all other countries in the dataset, *controlling for everything else in the model*. It provides an estimate of the extent to which S&P may be biased in favor of Eurozone countries, controlling for… .

c) *NewStadium* dummy (NFL Ticket prices): the estimated New Stadium dummy coefficient captures the differences in NFL ticket prices when NFL teams play in new stadiums, *controlling for everything else in the model*. It provides an estimate of the change in ticket prices associated with playing in (or perhaps moving to) a new stadium, controlling for … .

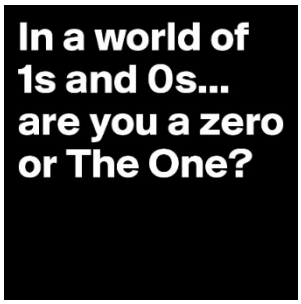
d) Constant/intercept coefficients: In fact, every model you've run has included an estimated ntercept or _cons coefficient… for the "constant dummy". When you estimate a model having a constant term/coefficient, you are basically estimating the coefficient for a dummy variable which always takes on the value 1. As you've seen, the estimated constant coefficient captures the average residual in the model, the part of the dependent variable not explained by the rest of the model. As you'll see, that's what dummies do!

5) *RHS v. LHS*: All of these examples feature dummies on the RHS, as independent explanatory variables in the analysis. We will stay with this case for quite a while, but eventually consider dummies variables on the LHS, as the dependent variables in the model. For reasons that will become clear, OLS models with LHS dummies are called Linear probability models (LPMs)

Dummies are Useful… so call them *Useful Dummies*!

6) It's always risky to attempt to categorize anything, as you always fear the omitted category, so instead of saying that dummy variable uses fall into two categories, let me instead say that here are two important uses of dummies: *Impact/Bias Analysis*, and *Silencing the (Endogeneity) Critics*

7) **Impact/Bias Analysis**: Econometrics methods are perhaps most useful when they provide researchers, policy makers, judicial authorities, etc. estimates of the impacts of certain policies… or capture differential categorical effects, which we might call *biases*, as in, say, the case of discrimination analysis. Models in these cases typically have a single dummy variable of interest on the RHS (the *variable in the spotlight*), with an estimated coefficient (the *favorite coefficient)* that tells you something about estimated impact or bias, controlling for everything else in the model. Those estimated *favorite* coefficients capture average differences (controlling for…). Here are some examples:



a) **Impact**: dummies might capture the presence or absence of gun control laws, legalization of this or that, no texting while driving laws, capital punishment laws, school lunch programs, etc etc … and the estimated coefficients tell us something about the average impact of those programs, controlling for … .

b) **Bias**: dummies might capture binary characteristics, perhaps defined by gender, ethnicity, race, religion, age, etc etc. …and the estimated coefficients might tell us something about whether, say, employment, wages, promotions, termination rates, mortgage rates, etc etc are higher or lower, for specific demographics, controlling for … .

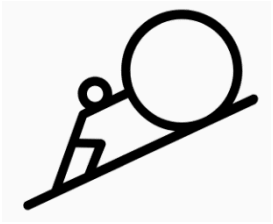
c) So if you are estimating impact or bias, dummy variables will be in your toolbag.

8) **Quieting the Endogeneity Critics**: Every econometrics analysis is subject to the criticism that relevant explanatory factors have been excluded from the analysis, leading to incredible and biased/misleading estimated effects.

a) An example: Suppose you are estimating gender bias in compensation. If the rest of your model is missing a full array of explanatory variables, then you really haven't controlled for much else that might be driving, say, wage differences across gender. Until you do so, endogeneity reigns supreme… and no one should pay any attention to you or your results.

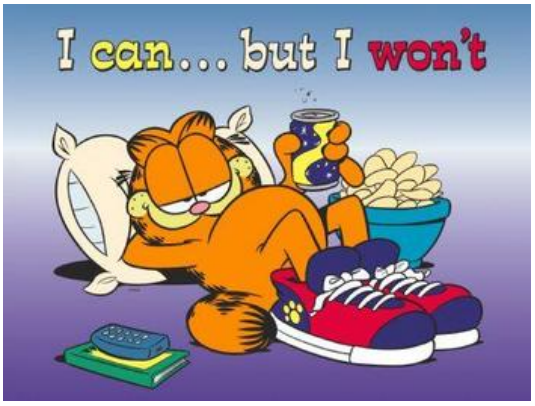
9) And yes, you should lose sleep over the endogeneity issue. With all models, but especially dummy variable models, you are always vulnerable to critics noting that you failed to control for the XYZ factor, which no one would ever ever associate with, say gender bias. And when said factor is added to your model, your gender dummy variable becomes *de minimus* (Latin for *way small*)[2], and loses all statistical significance (at any significance level you want to defend). The fear of being so exposed should keep you up late at night, building the best possible model.

10) **The Hard Working Researcher**: If you are a hard working researcher, you will work hard to grab the relevant heretofore excluded data, bring said data into your analysis, and explore the impact of said data on your estimated coefficients. Some might call this *robustness* analysis. Were your previous coefficient estimates biased by endogeneity? … Have you fixed the issue? … or maybe just partially remedied the situation? If it's the later, get back to work!



[2] Oxford's English Dictionary: *Too trivial or minor to merit consideration*. https://www.lexico.com/en/definition/de_minimis

11) **The Lazy Lazy Researcher**: But if you are the lazy lazy researcher, you just say *Bring on the dummies!* And when the (endogeneity) critics ask if you've controlled for this or that effect, you just say, *Mais oui, but of course!* … quickly followed by *look at all those Fixed Effects in my model*. Fixed Effects are just categorical dummies (a full complement of dummy variables, one for each categorical value)… and yes, there are plenty of examples below.



12) Here's an example, looking at the relationship between New Stadiums and NFL real ticket prices, working with 1996-2018 data… and estimating the new stadium effect:

i) Model (1): regress rprice on newstad: … $9.64*

ii) Model (2): add in yr FEs (fixed effects): … $15.74***

iii) Model (3): add in team FEs: … $8.02*

iv) Model (4): add in yr and team FEs: … $14.23***

	(1)	(2)	(3)	(4)
	rprice	rprice	rprice	rprice
newstad	9.637*	15.74***	8.015*	14.23***
	(2.33)	(4.57)	(2.40)	(5.99)
_cons	80.14***	79.94***	80.20***	55.83***
	(106.63)	(129.87)	(133.18)	(26.77)
Fixed Effects (FEs)				
yr		Yes		Yes
team2			Yes	Yes
N	727	727	727	727
R-sq	0.007	0.355	0.394	0.712
adj. R-sq	0.006	0.334	0.362	0.687

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

a) The results in Model (1) tell you that on average, and controlling for no other explanatory factors, real prices are $9.64 higher for teams playing in new stadiums. But you've controlled for no other explanatory factors, and the critics are quick to point out that you've completely ignored the systematic changes in rprice over time, and no doubt, that omission is biasing your results.

b) So in Model (2) you add yr (Fixed Effects) dummies (one dummy variable for each and every year) to Model (1) and discover that in fact, the new stadium premium is $6 higher, $15.74 (and adjusted R-sq increased from .01 to .33). Clearly, omitting yr effects from the model led to some serious omitted variable bias/impact.

c) But what is driving those yr effects? Model (2) is silent… all it tells you is that yr effects matter. So don't be lazy, try to better understand what yr-related factors might be driving this systematic variation in rprice, and impacting the estimated new stadium premium.

d) You now claim that you have in fact controlled for yr effects, even though you don't really know what they are. The critics, however, are not done with you. Now they point out that you've completely ignored the systematic team-by-team differences in rprice, and no doubt, that omission is seriously biasing your results. No doubt.

e) And so in Model (3) you add team fixed effects to Model (1) (each team gets its own dummy variable) and the estimated premium drops by $1.60 (from Model (1))… no small change to my eyes (and adjusted R-sq increased from .01 to .36). As with yr effects, you know that ticket prices are systematically varying team-by-team… but the team dummies provide no insight into why that's happening, they just tell you that something's going on there.

f) And finally in Model (4), you control for yr and team effects, and have the model with the highest adjusted R-sq and a rprice premium of $14.23. It's probably your best Model in the bunch. You can tell the critics that you controlled for yr and team effects. But when they ask you what drives yr-by-yr and team-by-team rprice effects, you have no reply.

13) Adding Fixed Effect dummies to your model allows you to brag that your model *controlled for this or that effect* (and eliminated these or those sources of omitted variable bias)… but they don't tell you anything about what's in fact driving your results. Maybe you care… and maybe you don't. But be prepared for the critics… and maybe, Don't be so lazy! … and instead try to better understand what actually drives ticket prices and the new stadium price premium.

14) We now turn to a series of examples of dummy variables in action. We'll start with the simplest of models… and slowly work our way back to Fixed Effects. Hold on tight!

Appendix: Creating Dummies in Stata

There are many ways to generate dummy variables in Stata.

Suppose you want to create a {0,1} dummy reflecting whether or not data are from or for the USA (or more specifically, the value of the country variable is "USA"). The following syntax examples will all generate the same usa dummy variable (I assume no missing values for the country variable):

```
gen usa = 0
replace usa = 1 if country == "USA"
```

```
gen usa = 1
replace usa = 0 if country != "USA"
```

```
gen usa = (country=="USA")
```

```
gen usa = (inlist(country,"USA"))
```

And if you want to create a North American dummy, you might try:

```
gen na = 0
replace na = 1 if country == "USA" | country =="Canada" | country ==
"Bermuda"
```

```
gen na = 1
replace na = 0 if country != "USA" & country !="Canada" & country !=
"Bermuda"
```

```
gen na = (country == "USA" | country =="Canada" | country == "Bermuda")
```

```
gen na = (inlist(country,"USA","Canada","Bermuda"))
```