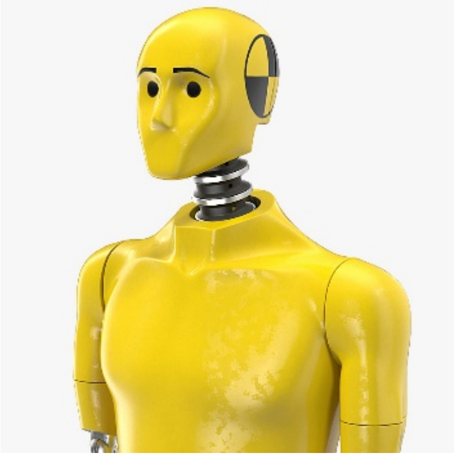
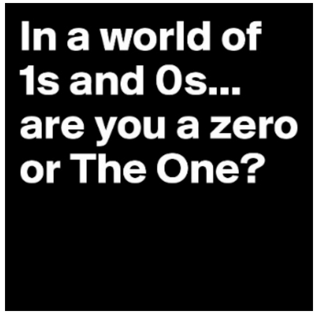
Dummy Variables: *Introduction*

- What's a dummy? … *(binary) indicator variables*
- Dummies revisited… *we've been here before!*
  - *AppleMusic, eurozone, newStadium, tempStadium, reloc, exp, twoteam, cap95, etc etc etc*
- On the RHS… and on the LHS
  - RHS: Capture what was not explained by the rest of the model
  - LHS: Linear probability models (LPM's), and …
- Useful dummies
  - Impact/bias analysis
  - Quieting the endogeneity critics (*Fixed Effects*)
- *You want examples?*
  - *Sovereign debt ratings; Gender bias in wages; Death penalty econometrics...*



What's a dummy? … (binary) indicator variables



- ***Binary categorical* (or *indicator*) variable**
  - 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.
    - *Value = 1*: Typically indicates the occurrence/presence of a category, event, outcome, characteristic, or thing… or perhaps that a logical statement is TRUE.
    - *Value = 0*: … indicates the absence of such.
- ***Average differences, ceteris paribus***: 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 revisited… we've been here before!

- ***AppleMusic*** dummy (*Spotify v. iTunes*): The estimated coefficient…
  - captures the average difference in weekly iTunes sales when AppleMusic is on the scene (or not), *controlling for everything else in the model*…
  - provides an estimate of the impact of AppleMusic on iTunes sales, controlling for…

- ***Eurozone*** dummy (*Sovereign debt ratings*): The estimated coefficient…
  - captures the average difference in NSRates for Eurozone countries, relative to all other countries in the dataset, *controlling for everything else in the model*…
  - provides an estimate of the extent to which S&P sovereign debt ratings may be biased in favor of Eurozone countries, controlling for… .

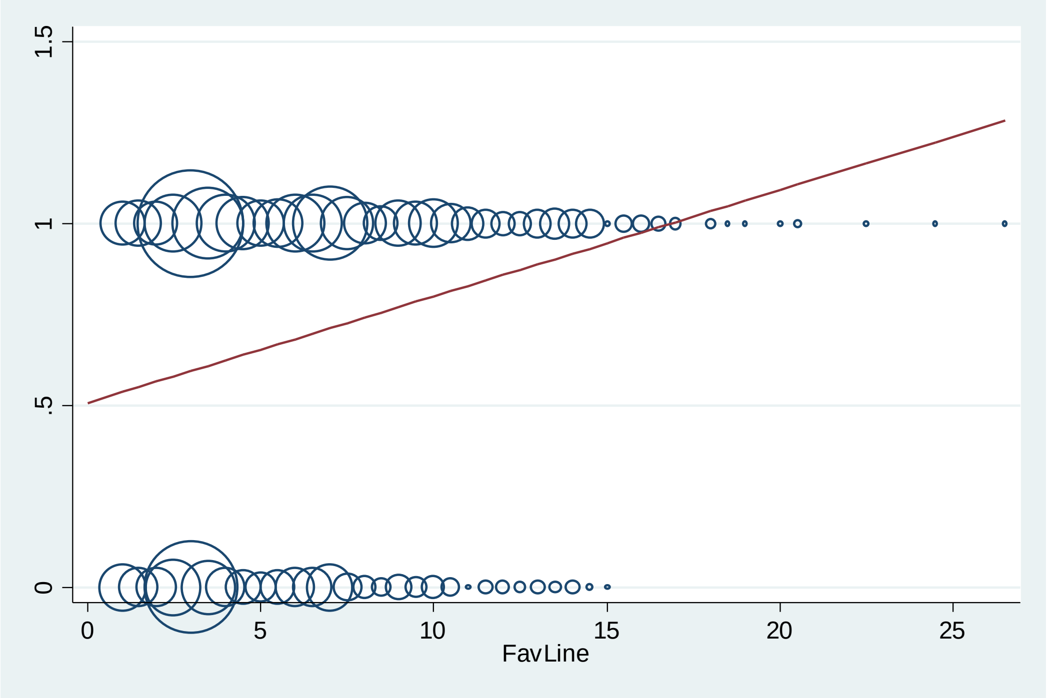
- ***NewStadium*** dummy (*NFL ticket prices*): The estimated coefficient…
  - captures the difference in NFL ticket prices when NFL teams play in new stadiums, *controlling for everything else in the model*…
  - provides an estimate of the change in ticket prices associated with playing in (or perhaps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

… and on the LHS:
Linear Probability Models (LPMs)

	favwins				
favline	=0		=1		nobs
1	69	*53%*	62	*47%*	131
1.5	51	*44%*	65	*56%*	116
2	48	*43%*	63	*57%*	111
2.5	95	*48%*	104	*52%*	199
3	269	*43%*	358	*57%*	627
3.5	92	*37%*	159	*63%*	251
4	49	*32%*	105	*68%*	154
4.5	37	*30%*	87	*70%*	124
5	28	*29%*	70	*71%*	98
5.5	36	*32%*	77	*68%*	113
6	49	*32%*	105	*68%*	154
6.5	50	*32%*	106	*68%*	156
7	67	*27%*	179	*73%*	246
7.5	24	*21%*	93	*79%*	117
8	17	*24%*	53	*76%*	70
8.5	12	*25%*	36	*75%*	48
9	21	*24%*	65	*76%*	86
9.5	13	*17%*	63	*83%*	76
10	17	*19%*	72	*81%*	89
10.5	10	*18%*	47	*82%*	57
11	1	*3%*	33	*97%*	34
11.5	7	*23%*	24	*77%*	31
12	6	*27%*	16	*73%*	22
12.5	4	*18%*	18	*82%*	22
13	6	*20%*	24	*80%*	30
13.5	4	*12%*	29	*88%*	33
14	7	*23%*	23	*77%*	30
14.5	2	*8%*	23	*92%*	25
15	1	*50%*	1	*50%*	2
15+	0	*0%*	38	*100%*	38

```
      Source |       SS           df       MS      Number of obs   =     3,290
-------------+----------------------------------   F(1, 3288)      =    156.39
       Model |  33.1244761         1  33.1244761   Prob > F        =    0.0000
    Residual |   696.42446     3,288  .211807926   R-squared       =    0.0454
-------------+----------------------------------   Adj R-squared   =    0.0451
       Total |  729.548936     3,289  .221814818   Root MSE        =    .46023

------------------------------------------------------------------------------
     favwins |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     favline |   .0292851   .0023418    12.51   0.000     .0246936    .0338765
       _cons |   .5066255   .0152011    33.33   0.000     .4768209    .5364301
------------------------------------------------------------------------------
```



Useful Dummies I: *Estimating Impact/Bias*

- Econometrics can provide researchers, policy makers, judicial authorities, etc. 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.
- ***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 …
  - The estimated coefficients tell us something about the average impact of those programs on whatever, controlling for … .
- ***Bias***: dummies might capture binary characteristics, perhaps defined by gender, ethnicity, race, religion, age, etc etc …
  - 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 … .
- So if you are estimating impact or bias, dummy variables will be in your **toolbar**.

Useful Dummies II: *Quieting the endogeneity critics*

- ***Omitted Variable Bias***: 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 and incredible inferential analysis..
- ***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.

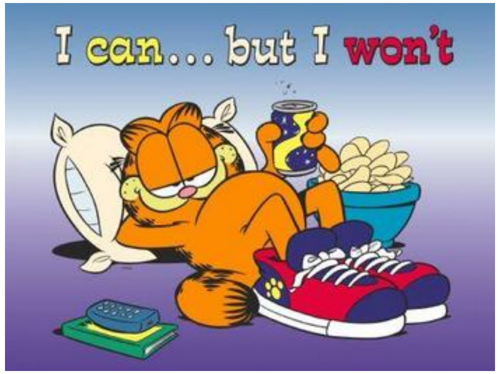
- **The Hard Working Researcher**: If you are a hard working researcher, you will work hard, you hard worker you, to grab the relevant heretofore excluded data, bring said data into your analysis, and explore the impact of said data on your estimated coefficients. Some might call this *robustness* analysis.

Dummies in Action: … *Bring on the Fixed Effects!*

- ***Fixed Effects***: 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*.

- Recall the ***Uber Gender Bias*** analysis:

			Dependent variable			
			Dependent variable: Tip Amount			
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Female Driver	**0.057 *****	**0.057 *****	**0.056 *****	**0.047 *****	**0.046 *****	**0.048 *****
	(0.001)	**(0.001)**	**(0.001)**	**(0.001)**	**(0.001)**	**(0.001)**
	0.47 ***					
	(0.0005)					
Controls/Fixed Effects						
Date		**X**	**X**	**X**	**X**	**X**
Hour of Week			**X**		**X**	**X**
Pick-up Geo				**X**	**X**	**X**
Drop-off Geo						**X**
Observations	**23,146,167**	**23,146,167**	**23,146,167**	**23,146,167**	**23,146,167**	**23,146,167**
R^2	0.0002	0.0003	0.001	0.020	0.020	0.028



- **Fixed Effects (FE's)**: Fixed Effects are just categorical dummies (a full complement of dummy variables, one for each categorical value)… and yes, there will be plenty of examples.
    - 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.
    - You perhaps observe systematic relationships… but ***Where's the Why?***

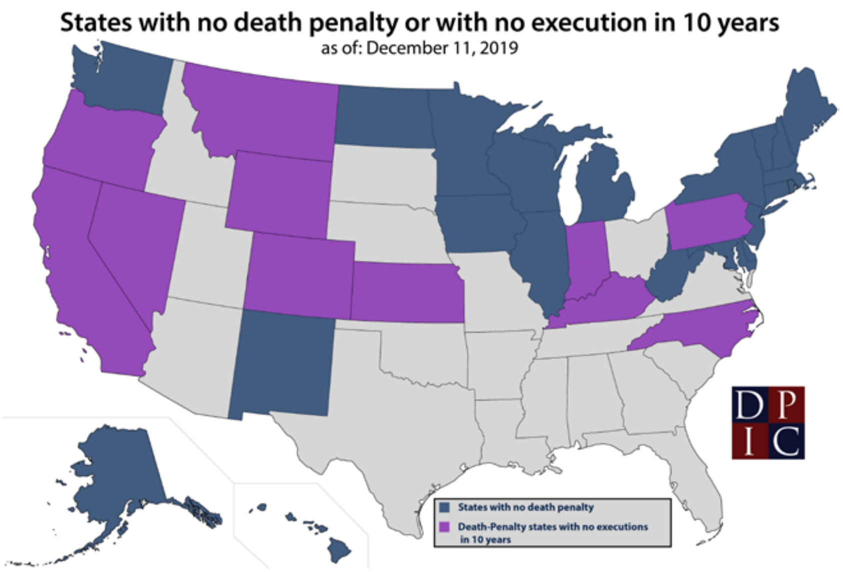
Dummies in Action:
NFL ticket prices and New Stadiums

	(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
Team			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

Dummy Variables: *You want more examples?*



S&P Global Ratings