

Simulation of Simpson's Paradox With Palmer Penguin Data

Andy Grogan-Kaylor

4 Jan 2021 14:29:31

Background

Simpson's paradox occurs when a bivariate association is reversed in a multivariate model. This example using the Palmer Penguins Data was inspired by a tweet by Andrew Heiss.

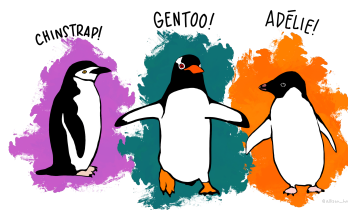


Figure 1: Palmer Penguins Illustration from @allison_horst

To begin, a little definition of penguin terminology is in order. Note the diagram defining culmen depth below.

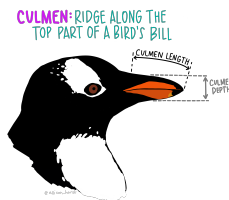


Figure 2: Culmen Depth from @allison_horst

Setup

```
. clear all  
  
. cd "/Users/agrogan/Desktop/newstuff/simpsonsparadox"  
/Users/agrogan/Desktop/newstuff/simpsonsparadox  
  
. use "penguins.dta"
```

Bivariate

```
. twoway (scatter culmen_depth_mm body_mass_g) ///
```

```

> (lfit culmen_depth_mm body_mass_g), ///
> title("Culmen Depth by Body Mass") ///
> caption("Palmer Penguin Data") ///
> scheme(michigan)

. graph export mygraph1.png, width(1000) replace
(file /Users/agrogan/Desktop/newstuff/simpsonsparadox/mygraph1.png written in PNG format)

```

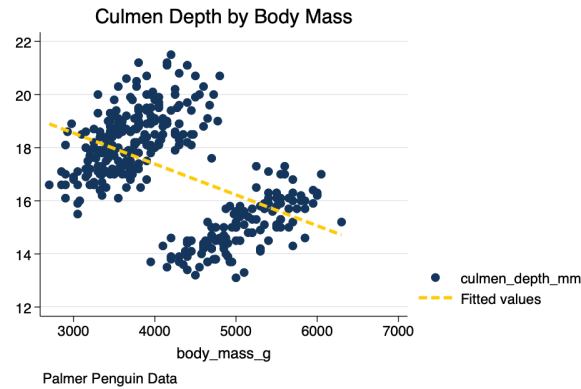


Figure 3: Scatterplot and Linear Fit

```
. regress culmen_depth_mm body_mass_g
```

Source	SS	df	MS	Number of obs	=	342
Model	296.15994	1	296.15994	F(1, 340)	=	97.41
Residual	1033.67459	340	3.04021939	Prob > F	=	0.0000
				R-squared	=	0.2227
				Adj R-squared	=	0.2204
Total	1329.83453	341	3.89980801	Root MSE	=	1.7436

culmen_dep_m	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
body_mass_g	-.0011621	.0001177	-9.87	0.000	-.0013937 -.0009305
_cons	22.03395	.5036206	43.75	0.000	21.04334 23.02455

Multivariate

```

. twoway (scatter culmen_depth_mm body_mass_g) ///
> (lfit culmen_depth_mm body_mass_g), ///
> by(species, title("Culmen Depth by Body Mass") caption("Palmer Penguin Data")) ///
> scheme(michigan)

. graph export mygraph2.png, width(1000) replace
(file /Users/agrogan/Desktop/newstuff/simpsonsparadox/mygraph2.png written in PNG format)

```

The association is reversed when we take into account multiple variables.

```
. regress culmen_depth_mm body_mass_g species
```

Source	SS	df	MS	Number of obs	=	342
Model	759.047284	2	379.523642	F(2, 339)	=	225.41
Residual	570.787248	339	1.6837382	Prob > F	=	0.0000
				R-squared	=	0.5708
				Adj R-squared	=	0.5683
Total	1329.83453	341	3.89980801	Root MSE	=	1.2976

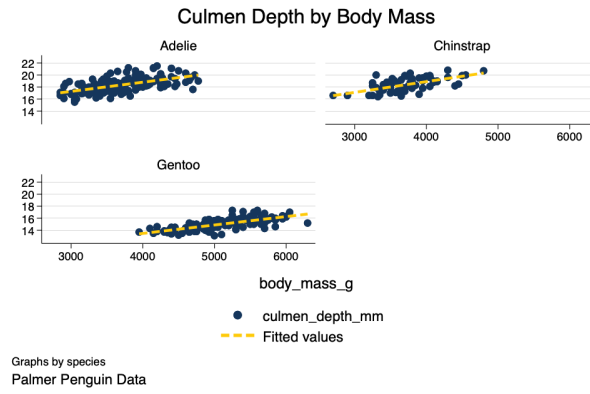


Figure 4: Scatterplot and Linear Fit

culmen_dep_m	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
body_mass_g	.0004877	.0001326	3.68	0.000	.0002269	.0007485
species	-1.974985	.1191142	-16.58	0.000	-2.209281	-1.740689
_cons	18.89014	.4200224	44.97	0.000	18.06396	19.71631