Analysis of Economics Data
Chapter 13: Multiple Regression Case Studies

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 13: Multiple Regression Case Studies

1. School Academic Performance Index
2. Cobb-Douglas Production Function (natural logarithms)
3. Phillips Curve (omitted variables bias)
4. Automobile Fuel Efficiency (natural logarithms; clustered errors)
5. Rand Health Insurance Experiment (randomized control trial)
6. Health Care Access and Outcomes (difference in differences)
7. Gains from Political Incumbency (regression discontinuity design)
8. Institutions and Country GDP (instrumental variables)
9. From Raw Data to Final Data

Examples 5-8 provide causal estimates using methods summarized in chap. 17.5.
Datasets: API, COBBDOUGLAS, PHILLIPS, AUTOSMPG, HEALTHINSEXP, HEALTHACCESS, INCUMBENCY, INSTITUTIONS

13.1 Case Study 1: School Performance Index

- How do we encourage schools to become better?
- Many U.S. states score schools based on student performance on standardized tests
  - in key subjects such as math and English conducted each year.
- Schools are expected to improve their scores over time.
  - failure to do so can lead to intervention by state authorities.

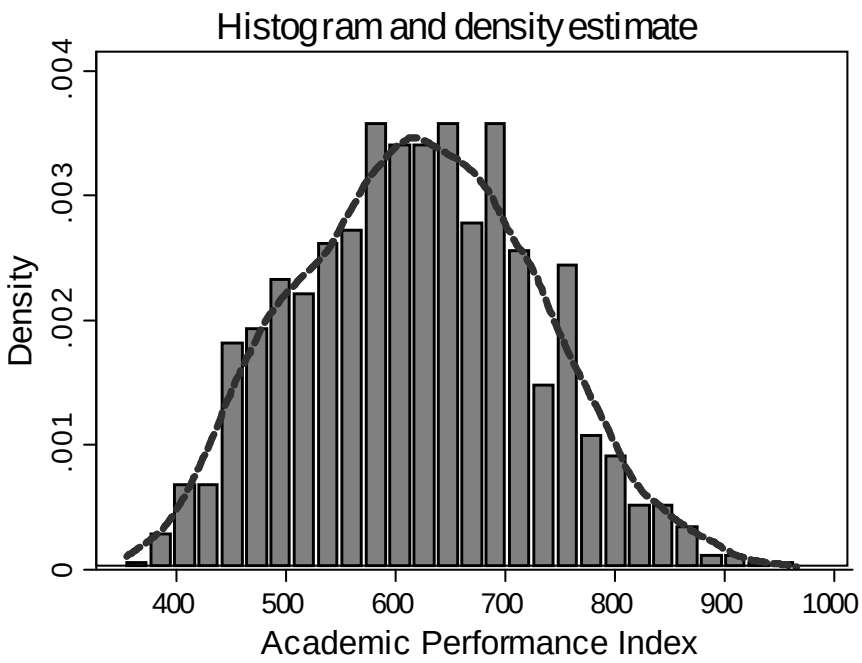
California Academic Performance Index

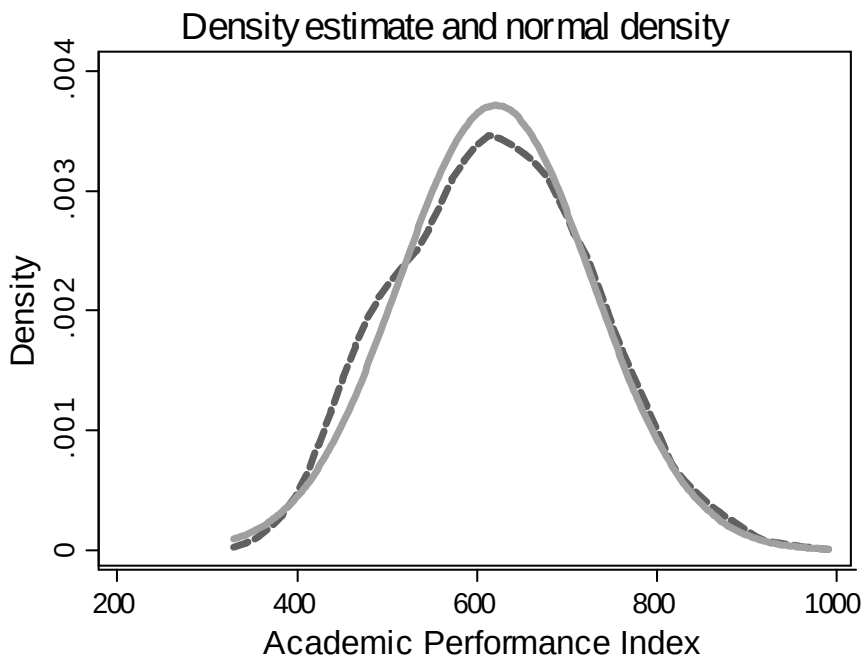
- Dataset API99 has data for 807 high schools in California in 1999 on
  - API (Academic Performance Index) in range 200 to 1000
    - ★ goal is for API > 800
  - socioeconomics variables *Edparent*, *Meals* and *Englearn*
  - school variable *Yearround*
  - teacher variables *Credteach* and *Emerteach*.

Variable	Definition	Mean	Standard deviation	Min	Max
Api	Academic Performance Index	620.94	107.44	355	966
Edparent	Average years schooling of parents	12.84	1.23	9.62	16
Meals	% of students in lunch program	21.92	23.67	0	98
Englearn	% of students English learners	14.00	12.79	0	66
Yearround	= 1 if multi-track year-round school	0.02	0.15	0	1
Credteach	% of teachers with full credentials	89.84	8.44	33	100
Emerteach	% of teachers with emergency creds	10.47	8.21	0	56

Univariate Analysis

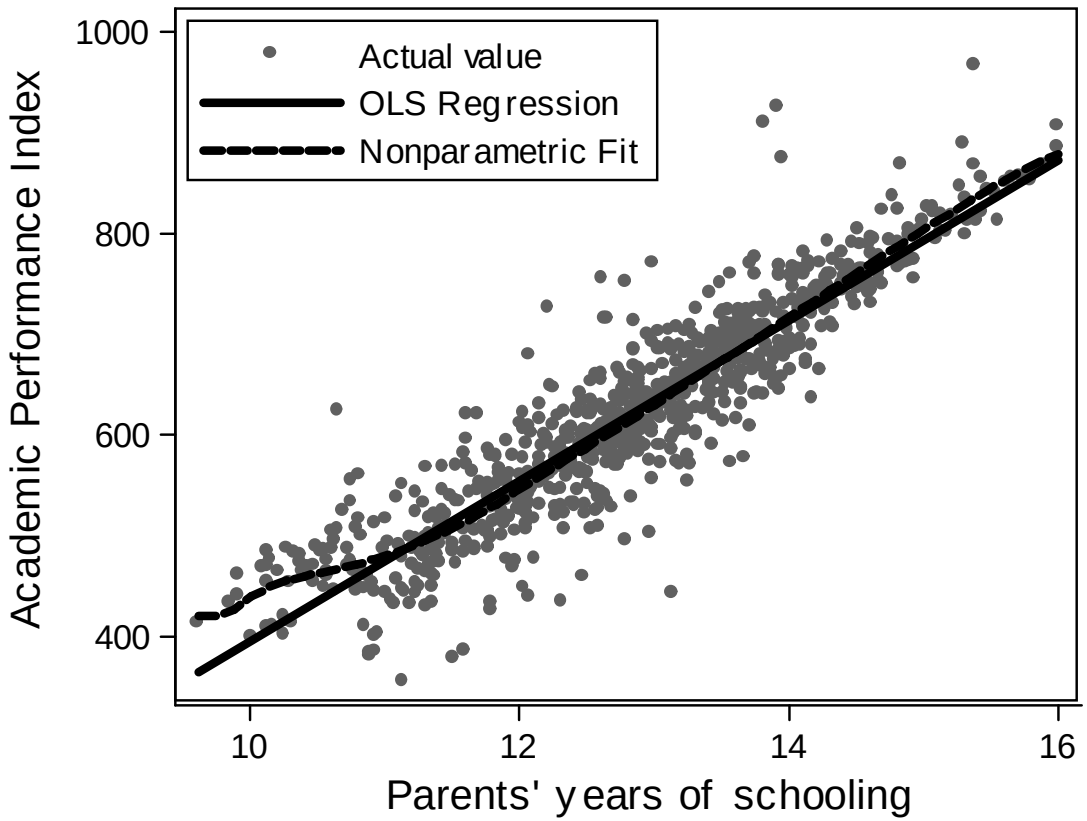
- Data are approximately normally distributed (by design)





Bivariate Analysis

- $\widehat{Api} = \underset{(15.99)}{-400.31} + \underset{(1.22)}{79.53} \times Edparent$, $s_e = 43.674$, $R^2 = 0.835$, $\bar{R}^2 = 0.834$ (heteroskedastic-robust se's in parentheses)
  - One more year of parent education associated with 80 more points!



Correlations

- Pairwise correlations also moderate to high for several other variables

	Api	*Edparent*	*Meals*	*Englearn*	*Yrrd*	*Cred*	*Emer*
Api	1						
Edparent	.91*	1					
Meals	-.54*	-.60*	1				
Englearn	-.66*	-.71*	.56*	1			
Yearround	-.19*	-.25*	.29*	.22*	1		
Credteach	.46*	.40*	-.27*	-.26*	-.18*	1	
Emerteach	-.45*	-.37*	.22*	.20*	.09*	-.82*	1

Multiple Regression

- Regress API on other regressors with default se's
  - *Edparent* coefficient little change from 79.53 to 73.94
  - all six regressors jointly statistically significant $F = 771.4$
  - subset of five regressors other than *Edparent* statistically significant $F = 14.80$ has $p = 0.000$
  - but R^2 only increases to 0.853 from 0.835 with just *Edparent*.

Variable	Coefficient	St. Error	t-stat	p-value	95% conf. int.	
Edparent	73.942	1.835	40.29	0.000	70.339	77.545
Meals	0.079	0.092	0.86	0.390	-0.102	0.260
Englearn	-0.358	0.177	-2.02	0.044	-0.706	-0.010
Yearround	25.956	10.752	2.41	0.016	4.850	47.062
Credteach	0.287	0.349	1.11	0.268	-0.298	1.073
Emerteach	-1.470	0.358	-4.11	0.000	-2.174	-0.767
Intercept	-345.328	44.027	-7.84	0.000	-431.750	-268.905
$n =$	$F(6, 22) =$	$R^2 =$	$\bar{R}^2 =$	$s_e =$		
807	771.4	.853	.852	41.4		

Conclusion

- Very strong association of API with socioeconomic characteristics
  - here parental education.
- Makes it difficult to calculate the separate role of other educational inputs
  - such as teacher quality.
- California also produced a "similar schools" index
  - this controls for socioeconomic characteristics.

13.2 Cobb-Douglas Production Function

- Important issue for determining market structure is whether or not returns to scale are constant, increasing or decreasing.
  - e.g. with increasing returns to scale a natural monopoly may arise.
- A production function models output (Q) as a function of capital (K) and labor (L)
  - plus possibly extra inputs such as land.
- The **Cobb-Douglas production function** specifies

$$Q = \alpha K^{\beta_2} L^{\beta_3}.$$

- With constant returns to scale doubling both inputs leads to exactly doubling output
  - for Cobb-Douglas this is the case if $\beta_2 + \beta_3 = 1$
  - versus increasing if $\beta_2 + \beta_3 > 1$ and decreasing if $\beta_2 + \beta_3 < 1$.

Natural Logarithm Transformation

- The model for Q is nonlinear in K and L
  - so OLS multiple regression seems impossible.
- But OLS is possible once take logs

$$\begin{aligned} \ln Q &= \ln(AK^{\beta_2}L^{\beta_3}) \\ &= \ln A + \ln(K^{\beta_2}) + \ln(L^{\beta_3}) \\ &= \ln A + \beta_2 \ln K + \beta_3 \ln L \\ &= \beta_1 + \beta_2 \ln K + \beta_3 \ln L, \end{aligned}$$

  where $\beta_1 = \ln \alpha$.
- This result uses the properties of natural logarithm that $\ln(a \times b) = \ln a + \ln b$ and $\ln a^b = b \ln a$.
- So do OLS regression of $\ln Q$ on an intercept, $\ln K$ and $\ln L$.

Example: Original Cobb-Douglas Study

- Dataset COBBDOUGLAS has U.S. aggregate data on manufacturing for the 24 years from 1899 to 1922.
  - From C.W. Cobb and P.H. Douglas (1928), "A Theory of Production," American Economic Review," pages 139-165.

Year	Q	K	L	Year	Q	K	L
1899	100	100	100	1911	153	216	145
1900	101	107	105	1912	177	226	152
1901	112	114	110	1913	184	236	154
1902	122	122	118	1914	169	244	149
1903	124	131	123	1915	189	266	154
1904	122	138	116	1916	225	298	182
1905	143	149	125	1917	227	335	196
1906	152	163	133	1918	223	366	200
1907	151	176	138	1919	218	387	193
1908	126	185	121	1920	231	407	193
1909	155	198	140	1921	179	417	147
1910	159	208	144	1922	240	431	161

Regression Results

- Regression results
  - HAC-robust standard errors (lag length 3) in parentheses

$$\widehat{\ln Q} = \underset{(.398)}{-.177} + \underset{(.062)}{.233} \times \ln K + \underset{(.134)}{.807} \times \ln K, \; s_e = 0.0581, \; R^2 = 0.957.$$

- The model fits the data very well
  - high R^2
  - coefficients of $\ln K$ and $\ln L$ are reasonably precisely estimated and highly statistically significant at level 0.05.
- The residuals are only slightly correlated with first three autocorrelations 0.11, -0.16 and -0.16
  - use lag length $m = 3$ as $0.75 \times 24^{1/3} = 2.16$.

Test of Specified Parameter Values

- Cobb and Douglas did not estimate this model by linear regression
  - instead set $\beta_2 = .25$ and $\beta_3 = .75$.
- Estimated coefficients are $b_2 = 0.233$ and $b_3 = 0.807$
- Test whether individually different from these values at 5%
  - e.g. test $H_0 : \beta_3 = .75$ against $H_a : \beta_3 \neq 0$
    - ★ $t = (.807 - 0.75)/.134 = 0.425$ with $p = 0.675$
    - ★ not statistically different from 0.75 at level 5%.
- Joint test of $H_0 : \beta_2 = .25, \beta_3 = .75$ against H_a : at least one of $\beta_2, \beta_3 \neq 0$
  - $F = 0.12$ with $p = \Pr[F_{2,21} > 0.12] = 0.889$.
  - the restrictions are not rejected at significance level 0.05.

Test Constant Returns to Scale

- Constant returns to scale if $\beta_2 + \beta_3 = 1$.
- $b_2 + b_3 = .233 + .807 = 1.040$ is close to 1.
- Formal test of $H_0 : \beta_2 + \beta_3 = 1$ against $H_0 : \beta_2 + \beta_3 \neq 1$
  - $F = 0.23$ and $p = \Pr[F_{1,21} > 0.23] = 0.636$.
  - restrictions are not rejected at significance level 0.05.
- The data are consistent with constant returns to scale.

Predicted Output

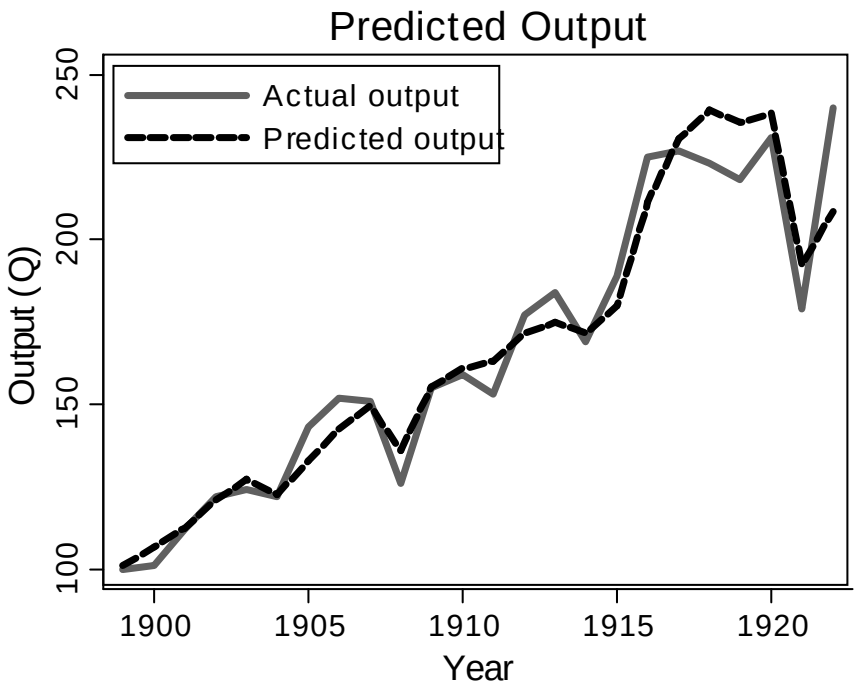
- $\widehat{\ln Q} = -.177 + .233 \ln K + .807 \times \ln K$
- In prediction allow for retransformation bias (chapter 15.5)

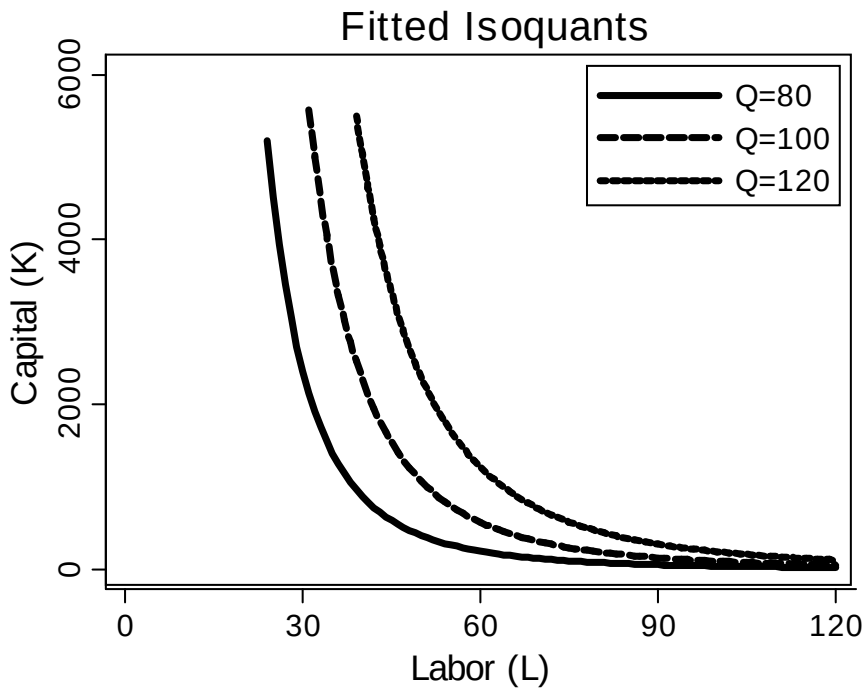
$$\begin{aligned} \widehat{Q} &= \exp(s_e^2/2) \times \exp(-.177 + .233 \ln K + .807 \times \ln K) \\ &= \exp(.0581^2/2) \times \exp(-.177) \times K^{.233} \times L^{.807} \\ &= .839 \times K^{.233} L^{.807}. \end{aligned}$$

- Gives sample mean of $\widehat{Q}$ equal to 166.0, quite close to mean of Q of 165.9.
- First panel of next figure plots actual Q and predicted Q against time
  - fit is quite good aside from final year.

Figures

- First panel plots actual Q and predicted Q against time.
- Second panel gives isoquants obtained next.





Fitted Isoquants

- Isoquants gives K as function of L for different values of Q

$$\begin{aligned} & Q & = \alpha K^{\beta_2} L^{\beta_3} \\ \Rightarrow \quad & K^{\beta_2} & = Q/(\alpha L^{\beta_3}) \\ & & = \alpha^{-1} Q L^{-\beta_3} \\ \Rightarrow \quad & K & = \alpha^{-1/\beta_2} Q^{1/\beta_2} L^{-\beta_3/\beta_2}. \end{aligned}$$

- Fitted values gives $K = 2.140 \times Q^{4.29} \times L^{-3.46}$.
  - ignores log transformation bias for simplicity
    - ★ small as $\exp(.0581^2/2) = 1.0017$ is close to 1.
- As expected isoquants do not cross.

HAC-robust Standard Errors

- For time series data concern about serially correlated errors.
- Less of a problem here as residual autocorrelations $\widehat{\rho}_1 = 0.11$, $\widehat{\rho}_2 = -0.16$, $\widehat{\rho}_3 = -0.16$
  - we nonetheless used them with $m = 3$.
- Robust standard errors of b_1 and b_2 are
  - default: 0.064 and 0.145.
  - heteroskedastic-robust: 0.105 and 0.216
  - HAC $(m = 3)$: 0.062 and 0.134.

13.3 Case Study 3: Phillips Curve

- **Phillips curve** plots price inflation against unemployment.
- A. W. Phillips (1958) found a negative relationship
  - an increase in money supply may stimulate the economy in the short-run
    - ★ leading to lower unemployment
    - ★ accompanied by some increase in prices
- Importance
  - can lower unemployment at the mild expense of somewhat higher price inflation.
  - but fierce debate as to whether this relationship holds in the long-run.

Example: U.S. Price Inflation

- Dataset PHILLIPS has annual U.S. data from 1949 to 2014
  - inflation based on GDP implicit price deflator.
- Later analysis uses expectations of future price inflation
  - 1. Survey of Professional Forecasters from Federal Reserve Bank of Philadelphia
  - 2. ad hoc measure weighted average of inflation over past 4 years
    - ★ $\dot{p}_t^e = 0.4\dot{p}_{t-1} + 0.3\dot{p}_{t-2} + 0.2\dot{p}_{t-3} + 0.1\dot{p}_{t-1}$, where $\dot{p}_t$ is inflation rate in year t.

Variable	Definition	Obs	Mean	St.Dev.	Min
Urate	Civilian unemployment rate (%)	66	5.87	1.63	2.70
Inflation	Annual inflation rate	66	3.20	2.32	-1.97
Expinflation	Forecast of one-year ahead *Inflation*	45	3.31	2.05	1.14
Pastinflation	Average of *Inflation* over past 4 years	63	3.65	2.04	1.48

Phillips Curve pre-1970

- OLS regression 1949 to 1969 looks good
  - the predicted negative relationship between inflation and unemployment
  - t-statistics in paranetheses based on HAC standard errors with $m = 3$.

$$\widehat{Inflation} = \underset{(4.49)}{7.111} - \underset{(-3.17)}{1.030} \times Urate, \; s_e = 1.32, \; R^2 = 0.454, \; n = 21,$$

Phillips Curve post-1970

- OLS regression 1970 to 2014 (HAC t-statistics with $m = 5$ in parentheses) looks bad

$$\widehat{Inflation} = \underset{(1.87)}{1.923} + \underset{(1.03)}{0.266} \times Urate, \ s_e = 2.44, \ R^2 = 0.258, \ n = 45.$$

- Positive though statistically insignificant relationship
  - "breakdown" of the Phillips curve.

Augmented Phillips Curve

- The problem is that people's inflation expectations also matter
  - add this as a regressor
- OLS regression 1970 to 2014 (HAC t-statistics with $m = 5$ in parentheses) looks good

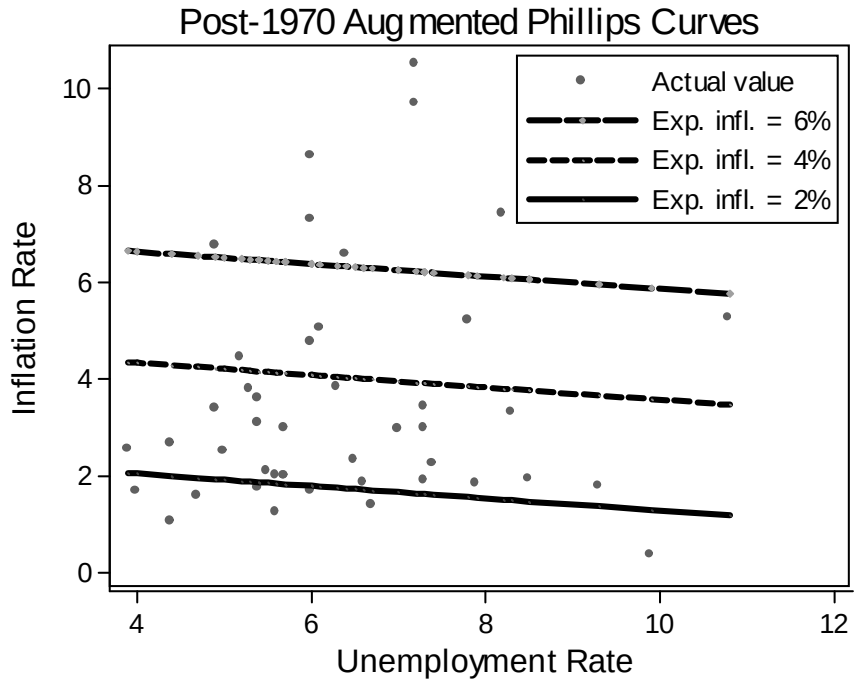
$$\widehat{Inflation} = \underset{(0.43)}{0.270} - \underset{(1.54)}{0.128} \times Urate + \underset{(13.58)}{1.147} \times Expinflation,\ s_e = 0.86,\ R^2$$

- *Urate now* negative, though statistically insignificant at 5%
  - and *Expinflation coefficient* is close to 1.
- Augmented Phillips curve relationship can be represented by a series of regular Phillips curves
  - each curve is given for a different expected inflation rate
  - e.g. for expected inflation rate of 2.0% we have

$$\begin{aligned} \widehat{Inflation} &= 0.270 - 0.128 \times Urate + 1.147 \times 2 \\ &= 2.559 - 0.128 \times Urate. \end{aligned}$$

Figures

- First panel shows time series plot
- Second panel shows augmented curve for 3 expected inflation rates.



Omitted Variables Bias

- Observed sign reversal for the coefficient of *Urate* is a classic example of **omitted variables bias**.
- True model: $Inflation = \beta_1 + \beta_2 \times Urate + \beta_3 \times Expinflation + u_t$.
- Incorrect bivariate model: $Inflation = b_1 + b_2 \times Urate$.
- Omitted variables bias from chapter 16.3: $\mathrm{E}[b_2] = \beta_2 + \beta_3 \gamma$
  - γ is the coefficient of *Urate* in a regression of *Expinflation* on *Urate*.
  - here bivariate regression of *Expinflation* on *Urate* has slope of .343.
- Then $\widehat{\mathrm{E}[b_2]} = -.128 + 1.147 \times .343 = 0.266$
  - equals estimated coefficient of *Urate* from bivariate regression of *Inflation* on *Urate*.

13.4 Automobile Efficiency

- Was better fuel efficiency of cars negated by switch to bigger more powerful cars?
- Dataset AUTOSMPG has annual data on most models of cars and light trucks on sale in the U.S. from 1980 to 2006 ($n = 27{,}871$).
- Model fuel efficiency (m.p.g.) which decreases with increased horsepower, car weight and torque.
- Estimate **log-log model**.
- Find that greatly increased fuel efficiency from 1980 to 1960 has been completely negated by heavier more powerful vehicles.
- Use **cluster-robust standard errors** with clustering on car manufacturer
  - because errors are correlated within manufacturer.

13.5 Rand Health Insurance Experiment

- Does better health insurance increase consumption of health care?
- 1970's **randomized control trial experiment** (to give a causal estimate)
  - randomly assign different levels of health insurance to different families.
- Dataset HEALTHINSEXP has 20,203 individual-year observations on 5,915 individuals in 2,205 families in experiment for 3 years or 5 years.
- Use data for the first year of experiment and only selected variables.
  - y = total annual spending on health
  - x includes six different insurance plans ranging from 0% coinsurance (free care) to 95% coinsurance.
- Find that spending increases with better health insurance
  - joint F test finds statistically significant at 5%
  - use cluster-robust standard errors with clustering on family.

13.6 Access to Health Care and Health Status

- Does greater access to health care improve health status?
- 1994 South Africa policy change
  - increase access to health care for children in communities with clinics.
- Use **difference-in-differences method** (to give a causal estimate)
  - change over time for treated (children in communities with clinics) minus change over time for untreated (children in communities without clinics).
- Dataset HEALTHACCESS has data on children ages 0 to 4.
- Outcome is a weight-for-age z-score
  - so normed to have mean 0 and standard deviation 1 for a representative world-wide population.
- Estimate is a 0.522 increase in weight-for-age z-score
  - and increase of 0.516 when controls variables are added.
- Use cluster-robust standard errors with clustering on community.

13.7 Gains to Political Incumbency

- Does being an incumbent increase the probability of winning the next election?
- Use **regression discontinuity method** (to give a causal estimate)
  - compare party vote in subsequent election if party just won the senate seat to that if party just lost the senate seat.
- Dataset INCUMBENCY has data on 1,390 Senate seat elections from 1914 to 2010.
- Estimated effect is a 5% to 7% increase in the vote if win previous election.
- Use heteroskedastic-robust standard errors
  - cluster-robust standard errors with clustering on state are similar.

13.8 Institutions and Country GDP

- Do better institutions lead to higher GDP?
- Use **instrumental variables estimator** (chapter 17.4) rather than OLS (to get a causal estimate).
- Dataset INSTITUTIONS has data on 64 countries settled by Europeans.
  - outcome is log GDP per capita in 1995
  - regressor is average protection against appropriation risk
  - instrument is log settler mortality (many years in the past)
- Find that better institutions lead to higher GDP.

13.9 From Raw Data to Final Data

- Going from raw data to a final dataset for analysis can be difficult
  - recently labelled as **data carpentry** or **data wrangling.**
- First task: **read** any sort of data into a statistical package
  - Excel spreadsheets (with extension .xls or .xlsx)
  - plain text file with character-separated values (with extension .csv)
  - a data file formatted for a commonly-used statistical package
  - a table in a PDF document (with extension .pdf)
  - hardcopy data may be scanned and digitized using e.g. Adobe Acrobat
  - web data obtained using a web scraping program.\
- Second task: **combine data** from multiple sources
  - merging data requires care.
- Third task: **cleaning** the data
  - entails recoding data and detecting data that are in error.
- And in many places: **check the data**.