A. Colin Cameron
Analysis of Economics Data: An Introduction to Econometrics
http://cameron.econ.ucdavis.edu/aed/
Version 1.1 February 2022

Table 13.6: Automobiles fuel efficiency: Summary Statistics (n=14,423)

Variable	**Definition**	**Mean**	**Standard deviation**	**Min**	**Max**
year	Year	-	-	1980	2006
idmfr	Unique numeric identifier for manufacturer	-	-	1	45
mpg	Miles per gallon	27.90	6.43	8.7	76.4
curbwt	Unloaded wight of vehicle in pounds	3019.53	593.64	1450	6200
hp	Horsepower	157.10	76.96	48	660
torque	Torque	238.68	105.13	69.4	1001
d_truck	= 1 if light (pickup) truck, = 0 if car	0.00	0.00	0	0
d_manual	= 1 if manual transmission, = 0 otherwise	0.38	0.49	0	1
time_d_manual	d_manual times (year - 1979)	5.10	8.04	0	27
d_diesel	= 1 if uses diesel fuel, = 0 otherwise	0.03	0.18	0	1
d_turbo	= 1 if turbocharged, = 0 otherwise	0.11	0.31	0	1
d_super	= 1 if supercharged, = 0 otherwise	0.01	0.11	0	1

From the multiple regression we have estimates of β_2 and β_3 of -0.128 and 1.146. Separate bivariate regression of *Inflgdp1yr* on *Urate*, not detailed here, yields a slope estimate of γ of 0.343. Then $\widehat{\mathrm{E}[b_2]} = -0.128 + 1.146 \times 0.343 = 0.266$, which is the previously obtained estimated coefficient of *Urate* from regression of *Inflgdp* on just *Urate*.

13.4 Automobile Fuel Efficiency

A major goal of public policy has been to improve the fuel efficiency of automobiles. The study, reproduced in part below, shows that there has been considerable improvement in fuel efficiency but that much of this has been undone by a consumer switch to bigger and more powerful vehicles.

13.4.1 Data Description

Dataset AUTOSMPG has annual data on most models of cars and light trucks on sale in the U.S. from 1980 to 2006. The original source is the article by Knittel (2012). The original dataset had 27,871 observations. Dataset AUTOSMPG follows the original study and does not include 876 observations that are dropped due to missing data or obvious coding errors. The remaining 26,995 observations are on 14,423 cars (*d_truck*= 0) and 12,572 trucks (*d_truck*= 1) that weigh less than 8,500 pounds. In this section only cars are analyzed.

Table 13.6 presents variable definitions and summary statistics for cars.

13.4.2 Graphical Analysis

Key determinants of fuel efficiency are engine power and vehicle weight.

The first panel of Figure 13.7 provides a plot of *mpg* against *hp* for cars in the years 1980 and 2006 along with Lowess nonparametric regression curves, Note that this figure restricts attention

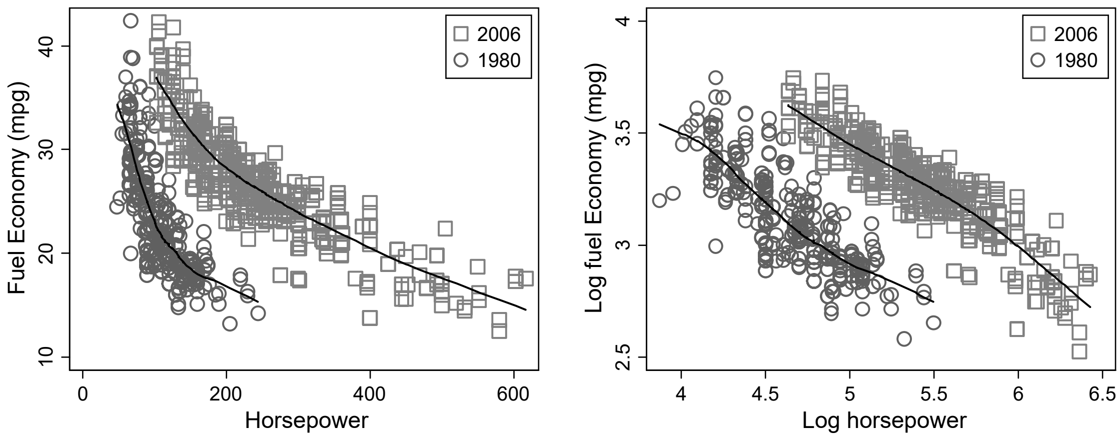

Figure 13.7: Automobiles fuel efficiency: Plot of fule conomy against horsepower

to cars that are fueled by gasoline (about 96% of the cars). The curves show a strong negative relationship. There is marked improvement in fuel efficiency over the 26 years, with an increase of approximately ten miles per gallon for a given level of horsepower. Finally note that horsepower has increased markedly over this period, and the purpose of the article by Knittel was to establish that virtually all the gains in efficiency were negated by a movement to more powerful and larger vehicles.

The second panel of Figure 13.7 is a variation that plots the natural logarithms of *mpg* and *hp*. Comparing the two curves, the fuel efficiency over the 26 years has increased by about 0.5 on a log scale, or by around 65% ($= 100 \times (\exp(0.5) - 1)$.

The second panel of Figure 13.7 indicates a relationship that is close to linear in logs. Similar graphs for *mpg* against *curbwt* and *lmpg* against *lcurbwt* again show a negative relationship that is also close to linear in logs. Thus a log-log regression model is estimated below.

13.4.3 Regression Analysis

The basic model is a log-log model with dependent variable *lmpg* and regressors *lcurbwt*, *lhp* and *ltorque* that control for car weight and engine size. Additional variables are whether the car has a manual transmission (*d_ manual)*, a time trend in this variable (*d_ manual_ time*), and various engine features (*d_ diesel*, *d_ turbo* and *d_ super*).

For all regressions standard errors are clustered on manufacturer. This is essential as a given car model often changes little from year to year, so an additional year's data is not adding much new independent information. And even within a given year some car models are very minor variations on a closely-related model. The correct cluster-robust standard errors are approximately five times incorrect heteroskedastic-robust standard errors; see Chapter 13.4.4.

The first model presented in Table 13.7 adds a linear time trend. All regressors except *lcurbwt*

Table 13.7: Automobile fuel efficiency: Multiple regression estimates with cluster-robust standard errors.

Dependent Variable *lmpg* (ln(*mpg*))						
	Time trend		Time dummies		Mfr fixed effects	
Variable	Coefficient	St. Error	Coefficient	St. Error	Coefficient	St. Error
lcurbwt	-0.415	0.052	-0.398	0.046	-0.383	0.032
lhp	-0.294	0.046	-0.324	0.047	-0.268	0.039
ltorque	-0.047	0.039	-0.019	0.038	-0.064	0.030
d_manual	0.084	0.013	0.087	0.013	0.101	0.013
time_*d_manual*	-0.004	0.001	-0.004	0.001	-0.004	0.001
d_diesel	0.191	0.019	0.196	0.018	0.212	0.017
d_turbo	0.028	0.011	0.025	0.010	0.051	0.010
d_super	0.049	0.019	0.055	0.017	0.034	0.011
year	0.017	0.001	-	-	-	-
intercept	-25.914	1.843	7.857	0.254	7.686	0.200
time dummies	No		Yes		Yes	
manufacturer FEs?	No		No		Yes	
R-squared	0.824		0.838		0.883	
Observations	14423		14423		14423	

are statistically significant at 5%. The coefficients are meaningfully large. Controlling for other regressors, a 10% increase in vehicle weight is associated with a 4.1% decrease in fuel efficiency, and a 10% increase in engine horsepower is associated with a 2.9% decrease in fuel efficiency.

Controlling for the various vehicle attributes there has been a very substantial improvement in fuel efficiency. The coefficient of the time trend is 0.0172, so over the 26 years from 1980 to 2006 car miles per gallon has increased by $26 \times 0.0172 = 0.447$ on a log scale, or by 56 percent ($= \exp(0.447) - 1$)!

The second fitted model in Table 13.7 estimated replaces the linear time trend (variable *year*) with a more flexible set of indicator variables for each year, with a dummy for 1980 the omitted dummy. There is little change in coefficient estimates, which are exactly the same as those for Model 1 in table 2 of the Knittel paper. The dummies increase in an approximately linear fashion. The omitted category is 1980 and the dummy for 2006 has coefficient with 0.522, a difference of 0.522 that is consistent with the linear time trend estimate of a 0.447 increase from 1980 to 2006.

The final model in Table 13.7 adds a set of indicator variables for each manufacturer. Again there is little change in coefficient estimates. The manufacturer dummy coefficients range from -0.401 for Lamborghini to 0.068 for Toyota and can be compared to the omitted manufacturer which was AMC.

13.4.4 Cluster-robust Standard Errors

A given car model often changes little from year to year, so an additional year's data is not adding much new independent information. And even within a given year some car models are very minor variations on a closely-related model. We therefore obtained cluster-robust standard errors with clustering on car manufacturer. This assumes that errors for observations on different manufacturers are uncorrelated, but for different observations (car model and/or year) from the same manufacturer there can be arbitrary correlation and heteroskedasticity.

In this example there are 14,423 observations and 38 manufacturers, so the average cluster size is 14423/38 = 379.6 observations. Consider, for example, the estimated OLS coefficient of *lhp* in the first fitted model in Table 13.7. Specialized commands provide the within-cluster correlation of the OLS residual, $\widehat{\rho}_e = 0.267$, and the within-cluster correlation of *lhp*, $\widehat{\rho}_{lhp} = 0.301$. The approximate standard error inflation formula given in Chapter 12.1.4 yields $\tau_{lhp} \simeq \sqrt{1 + 0.301 \times 0.267 \times (379.6 - 1)} = 5.61$. In fact variable *lhp* had cluster-robust standard error of 0.04552 that was 5.75 times the heteroskedastic-robust standard error of 0.00792 and 7.75 times the default standard error of 0.00587.

13.5 Rand Health Insurance Experiment

Health insurance, like any other form of insurance reduces an individual's exposure to risk. This leads to welfare gain as consumers are usually risk-averse. At the same time health insurance reduces the effective price or net price paid by the consumer. This may lead to a welfare loss due to overconsumption of health services, just as a subsidy on a good can lead to welfare loss.

A big issue in health care policy is therefore measuring how price responsive is the use of health care services. A simple method regresses health care use on the level of health insurance. In countries with universal health insurance all consumers have the same basic insurance, so such analysis is restricted to measuring the effect of supplemental insurance that provides increased cover beyond the basic cover. In the U.S. this approach is less restrictive as there is a very wide range of health insurance policies across individuals (including no insurance).

13.5.1 Experiment and Data Description

In either case, however, regression of health care use on the level of health insurance that is chosen by the individual measures only association, and not causation. Individuals with greater demand for health care can be expected to self select into choosing a higher level of health insurance. A finding that a higher level of health insurance is associated with higher use of health care services combines increase due to self selection and increase due to lower prices. The self selection can be partly controlled for by including as regressors various health status measures, such as the number of chronic conditions, and demographic variables, such as age and education. But there remain unobservable factors that are correlated with health insurance level, so OLS is biased and inconsistent as assumption 2 is violated.

As a result, the U.S. government funded the Rand Health Insurance Experiment conducted in the 1970s. This was a large social experiment where different families were randomly assigned health insurance policies with different levels of insurance cover. Specifically, different policies had different coinsurance rates where the coinsurance rate is the percentage of health costs paid by the individual. For example, with 25% coinsurance the insured pays 25% and the insurance company pays 75%. Experiment details and results are given in Manning, Newhouse, Duan, Keeler,