BLS WORKING PAPERS

U.S. Department of Labor
U.S. Bureau of Labor Statistics
Office of Prices and Living Conditions

On “Incorrect" Signs in Hedonic Regression

Timothy Erickson, U.S. Bureau of Labor Statistics

Working Paper 490
July 2016

On "Incorrect" Signs in Hedonic Regression

Timothy Erickson*
Bureau of Labor Statistics

February 2004
Revised September 2015

Abstract

I compute the equilibrium in a model of a differentiated-product market to show that, even when all consumers in the market agree that more of each product characteristic is preferred to less, an OLS regression of market-clearing prices on product characteristics can yield negative coefficients on the characteristics.

J.E.L. classification: C43.

Keywords: hedonic regression, quality change.

*Please address correspondence to Timothy Erickson, Bureau of Labor Statistics, Room 3105, Postal Square Building, 2 Massachusetts Avenue, NE, Washington, DC 20212-0001. Phone: (202) 691-5145. E-mail: Erickson.Timothy@bls.gov.

I. Introduction

I compute the equilibrium in a model market to show how a hedonic regression of market-clearing prices on product characteristics can yield negative coefficients on "vertical" characteristics, defined to be those having positive marginal utility for all consumers. I show how this outcome can arise when production technology permits increasing the amount of one characteristic only by reducing the amount of another, as in the relationship between vehicle horsepower and miles-per-gallon, or the tradeoff between CPU and GPU within an Intel microprocessor.[1] A first example illustrates how marginal costs influence such an outcome; a second example highlights the influence of markups. Surprisingly, it is possible for *all* the coefficients of a hedonic regression to be negative, even though all consumers prefer more of each characteristic to less.

The market model consists of heterogeneous utility-maximizing consumers who choose between discrete product varieties according to differing amounts of embodied characteristics. Each variety is produced by a different profit-maximizing firm, and a Nash-Bertrand equilibrium determines market prices.

II. The Model

1. The individual's problem and decision rule

My demand model is drawn from Berry, Levinsohn, and Pakes (1995). Each of M individuals faces the choice of buying exactly one unit of a particular type of commodity, or of not buying the commodity at all. The commodity has J varieties. I assume that the indirect utility derived by individual i from buying variety j is

$$u_{ij}(p_j, x_j, y_i) = \begin{cases} 100\log(y_i - p_j) + \beta_{i1}x_{j1} + \beta_{i2}x_{j2} & \text{if } p_j < y_i \\ -\infty & \text{otherwise,} \end{cases} \tag{1}$$

where p_j is the price of variety j, $x_j \equiv (x_{j1}, x_{j2})$ are the quantities of its quality-determining characteristics, (β_{i1}, β_{i2}) are individual i's marginal utilities with respect to these characteristics, and y_i is individual i's income. The individual's indirect utility from *not* buying *any*

[1] My exercise is prompted by the remarks on m.p.g. in Pakes (2001), p. 12.

variety is

$$u_{i0}(y_i) = 100\log(y_i).$$

Indirect utility is therefore

$$u_i(p, x, y_i) = \max\left\{u_{i0}(y_i), u_{ij}(p_j, x_j, y_i), j = 1, \ldots, J\right\},$$

where $p = (p_1, \ldots, p_J)$ and $x = (x_1, ..., x_J)$.

2. Heterogeneity and aggregation of consumers

Let F be the population distribution function of the vectors $(y_i, \beta_{i1}, \beta_{i2})$, each of which completely characterizes an individual consumer. Define

$$A_j(p, x) \equiv \left\{y_i, \beta_{i1}, \beta_{i2} | u_{ij}(p_j, x_j, y_i, \beta_{i1}, \beta_{i2}) > u_{ik}(p_k, x_k, y_i, \beta_{i1}, \beta_{i2}), k \neq j\right\}.$$

Then the demand share for variety j is

$$s_j(p, x) = \int_{A_{ij}(p,x)} dF(y_i, \beta_{i1}, \beta_{i2})$$

3. Production and Market Equilibrium

I assume there are J firms, each the unique producer of one of the product varieties. Each firm has a constant marginal cost c_j of producing its variety. Firms are assumed to set a price and then meet the subsequent demand. I suppose the number of consumers M is sufficiently large that the share equations $s_j(p, x)$ can effectively be treated as continuous in prices. Then, conditional on other firms' prices, the profit-maximizing price for firm j equals that which solves the first-order condition

$$(p_j - c_j)\frac{\partial s_j(p, x)}{\partial p_j} + s_j(p, x) = 0,$$

where

$$\frac{\partial s_j(p, x)}{\partial p_j} \equiv \lim_{\Delta p_j \to 0} \frac{s_j(p_j + \triangle p_j, p_{-j}, x) - s_j(p_j - \triangle p_j, p_{-j}, x)}{\triangle p_j},$$

for p_{-j} denoting the prices of all firms other than j. The Nash-Bertrand market clearing prices $p_i, ..., p_J$ are those that simultaneously satisfy all J firms' first-order conditions.

III. Computing Some Examples

Assume y_i, β_{i1}, and β_{i2} are independent of each other and have lognormal distributions. Note that the lognormal assumption ensures the marginal utilities for each characteristic are positive for every consumer. Make $M = 10,000$ draws from this joint distribution to get the population of consumers. Suppose there are $J = 10$ product varieties.

1. Example 1: Marginal cost depends on *one* product characteristic

Let the parameters of the lognormal distributions be $E(\beta_1) = 4$, $E(\beta_2) = 1$, $\text{var}(\beta_1) = \text{var}(\beta_2) = 2$, $E(y_i) = 30$, and $\text{var}(y_i) = 64$, and the characteristics and marginal costs be

x_{j1}	x_{j2}	c_j
1	9.3883	1.55
2	7.5334	2.20
3	9.7328	2.95
4	9.4068	3.80
5	6.6009	4.75
6	8.9064	5.80
7	7.6817	6.95
8	4.5023	8.20
9	3.3518	9.55
10	1.2470	11.0

The second column is derived from the first according to

$$x_{j2} = \frac{10 - (x_{j1} - 1)^2}{10 - 1} + z_j\sqrt{2}, \tag{2}$$

where the z_j are independent standard normal variables representing differences in firm efficiencies. The first term in (2) can be thought of as the average technological tradeoff between the two characteristics, but the main reason I have included z_j is to reduce the correlation between x_{j1} and x_{j2} to show that the results below do not stem from "excessive" collinearity.

The marginal costs satisfy

$$c_j = 1 + \frac{x_{j1}}{2} + \frac{x_{j1}^2}{20}. \tag{3}$$

The distinguishing characteristic of this example is that marginal cost increases in x_1 only. This tends to produce a positive coefficient on that characteristic.

Given these assumptions, the correlation matrix for (x_{j1}, x_{j2}, c_j) is

$$\begin{pmatrix} 1 & -0.83141 & .99282 \\ -0.83141 & 1 & -0.87557 \\ .99282 & -0.87557 & 1 \end{pmatrix}.$$

The correlation between the regressors in logarithmic form is $\text{corr}\left(\ln(x_1), \ln(x_2)\right) = -0.62883$.

Computing the market equilibrium yields

p_j	s_j	$markup_j$	$profits_j$
1.82	.0609	.26901	163.83
2.22	.0060	.02314	1.39
3.15	.2335	.20169	470.95
3.98	.1884	.17986	338.87
4.85	.0295	.09993	29.48
6.04	.1600	.24403	390.44
7.13	.1004	.18276	183.49
8.33	.0170	.12938	21.99
9.75	.0299	.20090	60.07
11.35	.0209	.34711	72.55

and $s_0 = .1535$.

An OLS regression of p_j on (x_{j1}, x_{2j}) yields a negative coefficient on x_2:

var	$coef$	t
1	1.9725	2.6692
x_1	.92248	14.783
x_2	−.17314	−3.3912

An OLS regression of $\ln(p_j)$ on $(\ln(x_{j1}), \ln(x_{2j}))$ does also:

var	$coef$	t
1	.80076	3.6867
$\log x_1$	.73513	8.7562
$\log x_2$	−.16922	−3.7861

I have not defined a "true" hedonic regression for this market, so I refrain from describing the coefficients produced by OLS as "estimates." Both of these regressions fit the data extremely well. Defining

$$R^2 = 1 - \frac{\text{var}\left(p_j - \hat{p}_j\right)}{\text{var}\left(p_j\right)},$$

the first regression has $R^2 = .98703$ for

$$\hat{p}_j = 1.9725 + .92248x_{j1} - .17314x_{j2}.$$

The second has $R^2 = .98395$, for

$$\begin{aligned} \hat{p}_j &= \exp(.5\operatorname{var}(e_j))[.80076 + .73513\ln(x_{j1}) - .16922(\ln x_{j2})] \\ e_j &= \ln(p_j) - .80076 - .73513\ln(x_{j1}) + .16922\ln(x_{j2}), \end{aligned}$$

where $\exp(.5\operatorname{var}(e_j))$ is a recommended prediction-bias adjustment. The more frequently reported "R-square" for loglinear regressions is $1 - \operatorname{var}(e_j)/\operatorname{var}(p_j) = .96593$. The fitted residuals from the two regressions are

$\underline{linear\ p_j - \hat{p}_j}$	$\underline{loglinear\ p_j - \hat{p}_j}$	$\underline{e_j}$
.54944	.28647	.1765
−.29005	−.42466	−.16968
.096809	−.26424	−.075381
−.053937	−.26496	−.059321
−.59217	−.46057	−.085591
.07862	.27201	.05118
.032827	.50427	.078453
−.2435	.32533	.044973
.05633	.57599	.066018
.36564	−.37232	−.027154

Note that the residuals relevant for judging the correctness of applying OLS to the loglinear model are the e_j.

2. Example 2: Marginal cost is the same for all varieties

Reset $E(\beta_1) = 1$ and

$\underline{x_{j2}}$	$\underline{c_j}$
10	1
9.8889	1
9.5556	1
9	1
8.2222	1
7.2222	1
6	1
4.5556	1
2.8889	1
1	1

The common marginal cost distinguishes this example; price variation now entirely reflects different markups. The new values for x_{j2} are derived from (2) by dropping the random term z_j. This was done to maintain 10 varieties in the market, because when z_j was included some varieties vanished from the market.[2] The correlations between the regressors become $\text{corr}(x_1, x_2) = -.96269$ and $\text{corr}(\ln(x_1), \ln(x_2)) = -.70705$.

The resulting market equilibrium is

p_j	s_j	$markup_j$	$profits_j$
1.0631	.0800	.0631	50.49
1.0684	.1159	.0684	79.24
1.0579	.1078	.0579	62.36
1.0473	.0814	.0473	38.53
1.0421	.0632	.0421	26.59
1.0421	.0441	.0421	18.55
1.0421	.0475	.0421	19.99
1.0526	.0590	.0526	31.03
1.0947	.1286	.0947	121.74
1.2735	.1638	.2735	447.96

(4)

and $s_0 = .1087$.

An OLS regression of p_j on (x_{j1}, x_{2j}) now gives *two* negative slopes

var	$coef$	t
1	1.696	8.8125
x_1	−.043245	−3.0553
x_2	−.05558	−3.4442

and $R^2 = .65524$.

The OLS regression of $\ln(p_j)$ on $(\ln(x_{j1}), \ln(x_{2j}))$ also gives two negative slopes

var	$coef$	t
1	.32511	8.5907
$\log x_1$	−.044936	−3.6705
$\log x_2$	−.10488	−8.8912

This regression fits much better, with $R^2 = .92411$ and $1 - \text{var}(e_j)/\text{var}(p_j) = .89264$. The fitted residuals from the two regressions are

[2]The random term causes some varieties to dominate others in their characteristic content, so no rational consumer would buy them without a sufficiently low price. The firms producing these varieties cannot offer such prices when their marginal costs are no lower than their competitors.

$\underline{linear\ p_j - \hat{p}_j}$	$\underline{loglinear\ p_j - \hat{p}_j}$	$\underline{e_j}$
−.03385	.083624	−.022425
.0084785	.053649	.012485
.022678	.039026	.017214
.024527	.03238	.013866
.019284	.031832	.0093801
.0069485	.03724	.0039726
−.017738	.049758	−.0085451
−.044257	.072642	−.021386
−.051572	.11512	−.024669
.065501	.22164	.020108

The surprising result of negative regression slopes on both vertical characteristics is partly due to the price function having pronounced local modes in characteristics space, a property that neither regression functional form will reveal. These local modes are associated with "extreme" varieties, which tend to be more isolated in characteristic space, i.e., they are farther from and/or have fewer competitors, and thus enjoy bigger markups. In (4) you can see that the highest price is associated with $(x_{j1}, x_{j2}) = (10, 1)$, which has only one immediate neighbor in characteristics space. The smaller price mode is associated with $(x_{j1}, x_{j2}) = (2, 9.8889)$, and although this has neighbors on both sides, the neighbor $(1, 10)$ provides very weak competition, since its x_{j1} content is halved, while its x_{j2} content increases only slightly.

The impact of these local markup modes is more readily seen if the tradeoff between the two characteristics is increased for low values of x_{j1}. Suppose now that $x_2 = (10, 9, 8, 7, 6, 5, 4, 3, 2, 1)'$, the "mirror image" of x_1. Keep all other parameters unchanged. We now have $\text{corr}\,(x_1, x_2) = -1$, $\text{corr}\,(\ln(x_1), \ln(x_2)) = -.83473$, and the market equilibrium

$\underline{p_j}$	$\underline{s_j}$	$\underline{markup_j}$	$\underline{profits_j}$
1.2577	.2385	.2577	614.52
1.0894	.1426	.0894	127.49
1.0421	.0354	.0421	14.89
1.0263	.0055	.0263	1.45
1.0158	.0112	.0158	1.77 .
1.0158	.0066	.0158	1.04
1.0210	.0140	.0210	2.95
1.0368	.0336	.0368	12.37
1.0841	.1425	.0841	119.91
1.2577	.2378	.2378	612.81

Note how the price is highest at the two endpoints of the tradeoff curve in characteristics space, smoothly declining from either endpoint to the smallest prices in the center.

The loglinear regression on this market is

var	*coef*	*t*
1	.66858	24.416
$\log x_1$	−.19634	−22.699
$\log x_2$	−.19463	−19.75

with $R^2 = .97618$ and $1 - \text{var}(e_j)/\text{var}(p_j) = .96721$. The residuals are

loglinear $p_j - \hat{p}_j$	e_j
.22042	.0088629
.10484	−0.019206
.048156	−0.006943
.017663	0.0082936
.0038542	0.0118 .
.0035435	.012111
.016709	.0041094
.046485	−0.010332
.10228	−0.021483
.2165	0.012786

How does this result arise? The one-dimensional tradeoff curve in the $(\log x_{j1}, \log x_{j2})$ plane is concave to the origin. Placing the $\log p_j$ axis perpendicular to this plane, the log-prices all lie above the tradeoff curve. The maximum log-prices lie above the intersections of the curve with the log-characteristics axes, and from either maximum the log-prices fall to their minimum above the intersection of the tradeoff curve and the 45-degree line in the first quadrant. See Figure 1. The least squares fit of a two-dimensional plane through this configuration of $(\log p_j, \log x_{j1}, \log x_{j2})$points will clearly have negative slopes and a positive intercept.

IV. Conclusion

The computed examples show how regression coefficients on vertical characteristics can be negative when production technology permits increasing the amount of one characteristic only by reducing the amount of another. By itself, this tradeoff is not sufficient to produce negative coefficients. The distribution of a finite number of products on the curve is important, too, because the distance between points in the characteristics plane determines

the level of competition. Products distant from competitors will, all else equal, have higher markups. Finally the cost of producing products at different points on the tradeoff curve is important. Varying these determinants can produce any combination of positive and negative slopes for a hedonic regression, even though all consumers agree that more of each product characteristic is preferred to less.

Acknowledgments

I thank Patricia Langohr for many valuable comments.

References

Berry, S., Levinsohn, J., and A. Pakes, 1995, Automobile prices in market equilibrium, Econometrica 63, 841-890.

Erickson, T., and A. Pakes, 2011, An experimental component index for the CPI: from annual computer data to monthly data on other goods, American Economic Review 101, 1707-1738.

Pakes, A., 2003, A reconsideration of hedonic price indices, with an application to PC's, American Economic Review 93, 1578-1596.

FIGURE 1