Christopher Gagné
Economics 272
April 31st, 2003

A Hedonic Model of Computer Prices

Introduction

Computers have become a virtual commodity. Most Windows-based personal computer (PC) manufacturers produce nearly homogeneous products, sometimes even in the same factories. It is impossible to ignore the rapid advances in technology over the past thirty years that have brought personal computers with exponentially more power than the mainframes of yore into the hands of typical consumers. Despite these advances in technology, the inflation-adjusted price of a new PC has fallen only slightly over the past twenty years. For example, a high-end 16-megahertz (MHz) 80386 computer with a small monochrome monitor and an 80-megabyte hard drive cost $3792 (1996 dollars) in January 1988. Today, a high-end 3,060 MHz Pentium IV with a 19" color display and 200-gigabyte hard drive costs $2695 (1996 dollars), a reduction in price of only 29%.

However, this price index (in the loosest sense of the term) does not consider the quality of the computer; if we were to divide the price of the computers by any significant metric (such as CPU speed, hard drive capacity, or RAM quantity), one would see that there has been a drastic reduction in the price of computers. Suppose one were to divide the real price of the computer by its total MHz. The computer from 1988 was $237 per MHz compared to the current $0.88 per MHz. Using such a metric, computer prices have fallen by 99.63%—quite a remarkable change, and quite different from the 29% figure. What could be discovered, then, if one were to calculate the change in price for a number of relevant factors?

My model attempts to quantify the value of a computer based on some metric of its quality and can be used to analyze the prices of computers nearly twenty-five years old. It is an important component of an overall quality-adjusted computer price model (which is well beyond the scope of this assignment). However, it is not the intention of this model to be a canonical metric of the value of PCs for any length of time. Instead, the reader should consider this model as a snapshot of time, a model whose value is greatest when compared to other similar models. Indeed, it is not the current price of computers that is so interesting, but rather the trend. A hedonic model of computer prices is useless: unlike a house, it is trivial to purchase every single different component of a computer separately. We can easily determine the expected value of a computer by summing the value of its components (whose true ß is trivial to discover independently), adding a premium for assembled systems, and discounting somewhat based on age if necessary.

The drastic reduction of computer prices over the past twenty-five years has created an entirely new culture in America; who could imagine bringing much of the computing power of a $10,000 mainframe to the pocket of a elementary school child in the form of a mere $60 gaming system? Who could envision a new globalization of knowledge services, largely catalyzed by the rapid decreases in computer prices?

Literature review

The economic community has been no stranger to the concept of adjusting computer prices for quality. The need for quality-adjusted computer prices is evident in the Bureau of Labor Statistics (BLS) calculation of the producer price index (PPI) of the computer industry. I have reviewed an excellent paper by Michael Holdway entitled

"Quality-Adjusting Computer Prices in the Producer Price Index: An Overview,"[1] which acted as a sounding board for many of my theories. Holdway described how the BLS calculates the PPI for computer prices. One component of that calculation is a hedonic model, much like my own, that is re-calculated with new data sets each quarter as the BLS generates new PPI statistics. The BLS obtains its computer pricing data from the webpages of several major computer manufacturers, and so is able to obtain over 600 observations with a minimum of effort.

Holdway presented an example regression from June 1999. The BLS measured twenty-three independent variables, all but five of which were dummy variables. Twenty variables measured a variety of computer specifications (from hard drive capacity to monitor quality). Three other dummy variables were used to capture whether the computer was sold by "Company A," "Company B," or "Company C," apparently companies with a reputation for statistically higher or lower prices (presumably resulting from brand name recognition or quality).

685 unique observations were used for his regression. All functional forms were linear. All independent variables except "Sound card and 2 Speakers" were significant at the 95% level, $\bar{R}^2$ was 0.963, the standard error of the dependent variable was 85.2, and the F statistic was 773.6. No other statistical information was available. Based on my cursory analysis of the available data, the BLS equation was a success.

[1] 1. Holdway, Michael. 2003. *Quality-Adjusting Computer Prices in the Producer Price Index: An Overview* [HTML Document]. Bureau of Labor Statistics, October 16 2001 [cited April 24 2003]. Available from http://www.bls.gov/ppi/ppicomqa.htm.

Data source and characteristics

The BLS collected data from the web pages of several major computer manufacturers. However, I was interested in a model that might have some applicability to older computers as well. Based on my relatively accurate hypothesis that the sole fact that a computer is in new condition does not make it particularly more expensive, I felt it appropriate to consider both new and used computers. Furthermore, many interesting and older computers are rarely sold in new condition. It would be extremely difficult to obtain a reliable coefficient for a "new" dummy variable because the value of "new" for older computer could be either quite high (as a collectors' item) or quite low (as for totally banal, useless computers).

Based on the theory that the a reasonable estimate of the value of an object is the last price at which it was sold, I manually obtained my data from recently completed computer auctions on eBay.com. eBay.com is the world's largest Internet-based auction service, and has a sufficient balance of buyers and sellers to imply relatively accurate prices devoid of the shortage or surplus quantity effects of pricing.

Variable specification and functional form

In total, I obtained 156 observations for price and seven independent variables as follows:

P_i Final auction price, not including shipping charges
LAP_i Laptop dummy
MAC_i Macintosh dummy
CPU_i Central Processing Unit (CPU) speed in MHz
RAM_i Random Access Memory (RAM) quantity in MB
HD_i Total hard drive capacity in GB
BC_i Total number of basic included components, including

- CRT monitor < 19", including built-in displays
- Inkjet printer
- CD-R/CD-RW drive

- DVD-ROM drive
- Wireless networking capability
- Legal copy of standard operating system (assumed unless stated otherwise)

PC_i Total number of substantial included components, including

- CRT monitor ≥ 19", including built-in displays
- LCD monitor, including built-in displays on desktops but not on laptops
- Laser printer
- DVD-R/DVD-RW drive
- Substantial legally included software, per package (such as Photoshop, Microsoft Office, etc)

P_i is the final value of the winning bid of the i^{th} eBay computer auction in US$, not including shipping. I did not include uncompleted auctions (i.e. when the reserve bid price had not been met) in the data set.

LAP_i is a dummy variable equal to 1 if the i^{th} computer was a laptop. From this dummy variable, three additional slope dummies were included: $LAP_i \cdot CPU_i$, $LAP_i \cdot HD_i$, $LAP_i \cdot RAM_i$. I chose to add these four variables because laptops are typically more expensive than a desktop computer, all other independent variables held equal. This occurs for two reasons: first, laptops typically include expensive components such as a built-in LCD display, a large battery, and a motherboard with highly specialized components. Furthermore, there is some evidence that the price of CPUs, hard drives, and RAM is greater for laptops due to the level miniaturization necessary. This effect is best seen in the price of hard drives and is somewhat difficult to observe in RAM and CPU prices.

MAC_i is a dummy variable equal to 1 if the i^{th} computer was a Macintosh. Macintosh computers are well known in the industry for their innovative operating system, quality industrial design, and legendary ease of use. For these reasons alone, Macintosh users are willing to pay a premium for the computer, all other independent

variables held equal. Furthermore, the architecture of Macintosh computers is quite different from their PC counterparts; for any given CPU rating, the Macintosh will be a faster and more expensive system. As a result, it is not possible to directly compare the MHz rating of a Macintosh to that of a PC. It is for this reason that I included the independent variable $MAC_i \cdot CPU_i$, which permits a higher valuation of the Macintosh's CPU speed.

CPU_i is a measurement of the stated CPU speed in MHz in the i^{th} computer. A machine with an 800MHz CPU is *roughly* twice as fast as a machine with a 400MHz CPU. I used a total MHz rating for machines with more than one CPU (quite common for Macintoshes and rare for PCs) because most modern operating systems are capable of using more than one processor very efficiently. While CPU speeds have generally been growing at an exponential rate over the past twenty years. The two samples I mentioned exhibited a 19,125%, or nearly cubic, increase in speed in just 15 years. The change in $\frac{\$}{MHz}$ has followed a very similar trend. The most recent (and fastest) chips would exhibit a price premium due to their novelty, however I do not believe this is a justification for using the semi-log left or semi-log right functional form. Furthermore, I expect the effects of the new-chip premium to be somewhat offset by a slight penalty exacted against computers with the slowest CPUs (because they are relatively the least useful).

RAM_i is the total quantity of RAM in the i^{th} computer, measured in megabytes. I am using the linear functional form with the same reasoning found in the CPU variable.

HD_i is the total capacity of all hard drives in the i^{th} computer, measured in gigabytes. I am using the linear functional form with the same reasoning found in the CPU variable.

BC_i is an index that measures the number of *basic components*, or minor upgrades beyond a basic system, found in the i^{th} computer. This is a combination variable that serves to capture the presence of six dummy variables. It would certainly be possible to run a regression with a separate independent variable for each of the measured basic components as there is very little multicollinearity between these components, but doing so would be beyond the scope of this assignment. Because we are measuring the number of independent components in a particular system, and because none of the components are more or less valuable in a 'fully-loaded' configuration in comparison to a basic configuration, the linear form seems the most appropriate.

PC_i is an index that measures the number of *premium components*, or minor upgrades beyond a basic system, found in the i^{th} computer. This is a combination variable that serves to capture the presence of five dummy variables, and bears the same features, functional form, and logic as BC_i.

Intentionally omitted variables

I have intentionally omitted a large number of variables for at least one of the following reasons:

1. Lack of reliable data availability or unreasonable difficulty in measurement/quantification (such the physical condition of the computer)
2. Perfect or very severe imperfect theoretical correlation with an existing variable (such as bus speed vs. CPU speed)

3. Low value of individual component, therefore relatively irrelevant in purchasing decision (such as a basic keyboard).

Related intentionally omitted variables include: power supply wattage, manufacturer, age, operating system, hard drive type, RAM type, included keyboard/mouse/speakers, CPU type, bus speed, feedback rating of the seller, shipping price, auction characteristics (such as quality of photos), warranty policy, physical condition of the computer, and seller feedback rating.

Shipping price was a particularly interesting variable to consider. Some auctions listed the price of shipping in the item description, others required that the buyer contact the seller after the auction after the auction and provide a zip code. I chose to not include the shipping price because it would be impossible to obtain shipping price data in the latter scenario, because only including auctions that listed a shipping price would cause bias, and because I expect the shipping price to be irrelevant unless it was significantly outside of the buyer's expected range. Furthermore, I noticed that the quoted shipping prices were quite similar across auctions. For this reason, I expected much of the shipping price to be reflected in the constant.

Equation "0"

On April 18th 2003, I submitted a model with expected signs and variables as follows:

$$P_i = \beta_0 + \beta_1 LAP_i + \beta_2 (LAP_i \cdot CPU_i) + \beta_3 (LAP_i \cdot HD_i) + \beta_4 (LAP_i \cdot RAM_i) + \beta_5 MAC_i + \beta_6 (MAC_i \cdot CPU_i) + \beta_7 CPU_i + \beta_8 RAM_i + \beta_9 HD_i + \beta_{10} BC_i + \beta_{11} PC_i + \varepsilon$$

I ran an initial regression on this model (attached) using an preliminary data set of 73 observations, and included some computer outputs in the section entitled Equation "0." At an initial consultation with Professor Studenmund, I was advised that although the data for seven independent variables were collected, the eleven functional independent variables in the model specification exceeded the seven independent variable limit imposed by the assignment. I therefore elected to remove all laptops and Macintosh computers from the data sample, which brings the equation down to a slightly more manageable five independent variables. I had hoped that my model would explain the prices of both laptops and Macintosh computers, providing information about four distinct types of computers. Recall that there are tremendous differences between laptops and desktops, Macintosh computers and PCs. I therefore believe that instead of attempting to use the same data set and dealing with the repercussions of a few significant, critical, and omitted variables, it was best to remove the laptop- and Macintosh-positive observations from the data set, and start anew. Because removing all laptops and Macintosh caused a sharp reduction in the number of observations, (from an initial 103 to 46), I collected additional data to bring the total number of observations (of desktop PCs) to a useful but manageable 100.

Equation 1

As mentioned earlier, I elected to remove the MAC_i and LAP_i observations and variables, which reduced the number of independent variables in the equation from 11 to 5.

Model specification (independent variables and functional form):

$$P_i = \beta_0 + \beta_1 CPU_i + \beta_2 RAM_i + \beta_3 HD_i + \beta_4 BC_i + \beta_5 PC_i + \varepsilon_i$$

Expected signs of coefficients:

$$P = f(\overset{+}{CPU}, \overset{+}{RAM}, \overset{+}{HD}, \overset{+}{BC}, \overset{+}{PC}) + \varepsilon_i$$

Estimated equation:

$$\hat{P}_i = \underset{}{13.0602} + \underset{(0.0411)}{0.2926 CPU_i} - \underset{(0.1059)}{0.0549 RAM_i} - \underset{(1.1980)}{1.4505 HD_i} + \underset{(17.6137)}{47.3140 BC_i} + \underset{(89.0196)}{547.8555 PC_i}$$

$$t = 7.1079 \quad t = -0.5181 \quad t = -1.2108 \quad t = 2.6861 \quad t = 6.1543$$

$n = 100 \quad t_c = 1.662 \quad \bar{R}^2 = .764 \quad F = 65.451 \quad d = 1.1045$

F-test (1%):

$H_0 : \beta_{CPU} = \beta_{RAM} = \beta_{HD} = \beta_{BC} = \beta_{PC} = 0$ $\qquad F = 65.451$

$H_A : H_0$ is not true $\qquad F_c = 3.22$

$F > F_c \Rightarrow$ Reject H_0

t-tests (5%):

$H_0 : \beta_{CPU} = 0$ $\qquad t_{CPU} = \dfrac{0.2926 - 0}{0.0411} = 7.1079$

$H_A : \beta_{CPU} > 0$ $\qquad t_c = 1.662$

$|t_{CPU}| > t_c \Rightarrow t_{CPU} > 0 \Rightarrow$ Reject H_0

$H_0 : \beta_{RAM} = 0$ $\qquad t_{RAM} = \dfrac{0.0549 - 0}{0.1059} = -0.5181$

$H_A : \beta_{RAM} > 0$ $\qquad t_c = 1.662$

$|t_{RAM}| \not> t_c \Rightarrow t_{RAM} \not> 0 \Rightarrow$ Cannot reject H_0

$H_0 : \beta_{HD} = 0$ $\qquad t_{HD} = \dfrac{-1.4505 - 0}{1.1980} = -1.2108$

$H_A : \beta_{HD} > 0$ $\qquad t_c = 1.662$

$|t_{HD}| \not> t_c \Rightarrow t_{HD} \not> 0 \Rightarrow$ Cannot reject H_0

$H_0 : \beta_{BC} \leq 0$
$H_A : \beta_{BC} > 0$
$t_{BC} = \frac{47.3140 - 0}{17.6137} = 2.6861$
$t_c = 1.662$
$|t_{BC}| > t_c \therefore t_{BC} > 0 \therefore$ Reject H_0

$H_0 : \beta_{PC} \leq 0$
$H_A : \beta_{PC} > 0$
$t_{PC} = \frac{547.8555 - 0}{89.0196} = 6.1543$
$t_c = 1.662$
$|t_{PC}| > t_c \therefore t_{PC} > 0 \therefore$ Reject H_0

VIF tests:

$VIF_{CPU} = \frac{1}{1 - 0.7904} = 4.7709$
$VIF_c = 5$
$VIF_{CPU} < VIF_c \therefore$ No severe multicollinearity for CPU_i.

$VIF_{RAM} = \frac{1}{1 - 0.5328} = 2.1404$
$VIF_c = 5$
$VIF_{RAM} < VIF_c \therefore$ No severe multicollinearity for RAM_i.

$VIF_{HD} = \frac{1}{1 - 0.8172} = 5.4704$
$VIF_c = 5$
$VIF_{HD} > VIF_c \therefore$ Severe multicollinearity exists for HD_i.

$VIF_{BC} = \frac{1}{1 - 0.4206} = 1.7259$
$VIF_c = 5$
$VIF_{BC} < VIF_c \therefore$ No severe multicollinearity for BC_i.

$VIF_{PC} = \frac{1}{1 - 0.2099} = 1.2656$
$VIF_c = 5$
$VIF_{PC} < VIF_c \therefore$ No severe multicollinearity for PC_i.

Durbin Watson test:

$H_0 : \rho \leq 0$
$H_A : \rho > 0$
$n = 100$
$k' = 5$
$\Rightarrow$ $d_L = 1.57$
$d_U = 1.78$
$d = 1.1045$

$d < d_L \therefore$ Reject H_0, Impure positive serial correlation exists.

Park test:

Testing for: $\ln(e_i^2) = \beta_0 + \beta_1 \ln CPU_i + u_i$

$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

Result: $\ln(\hat{e}_i^2) = 3.6021 + 0.7461 \ln CPU_i$
(0.2114)
$t = 3.5291$

$n = 100 \quad t_c = 2.63 \quad \bar{R}^2 = .1037$

$|t_{CPU}| > t_c \Rightarrow$ Reject H_0, Heteroskedasticity is present.

Diagnostics:

Omitted variables:

Two somewhat surprising results—the presence of serial correlation and heteroskedasticity (described in greater detail later)—point to the reasonable hypothesis that there are omitted variables. There is some evidence of positive bias in BC_i and PC_i. However, BC_i, which measures the presence of a basic component such as a CD-RW or DVD drive, is quite close to the expected value of one of those used components separately. The coefficient of PC_i is somewhat higher than would be expected, however not to the extent to encourage the belief that an omitted variable is having a significant effect. It is difficult to measure whether there is omitted variable bias in the coefficients of CPU_i, RAM_i, or HD_i, because there is significant multicollinearity between those variables that is resulting in a unexpected statistically significant high coefficient for CPU_i, and statistically insignificant negatively-biased coefficients for RAM_i and HD_i.

Nonetheless, it is worth considering the potential impact of an omitted variable. With the obvious exception of the shipping price, most of the variables I mentioned in the "intentionally omitted variables" section would have a positive coefficient and be positively correlated with all other independent variables. Let us consider the two

possible effects of two different types of omitted variables: feature variables, and a shipping price variables.

I have intentionally left out a number of feature variables, ignoring nearly 17 features that the BLS found significant with a large data set. The bias is as follows:

$$\text{Bias} = \beta_2\beta_1 = \beta_{FEATURE} \cdot \rho(r_{IN,FEATURE}) = + \cdot + = +$$

As mentioned earlier, there was some evidence of positive bias in the estimated coefficients. This was certainly expected. Lets analyze the effect of the omitted shipping price. The expected sign of $ß_{SHIPPING}$ is negative—we would expect a high shipping price to drive down the maximum bid given a fixed willingness to pay. The relationship between the shipping price is minor but positive. There should be very little difference in the weight between cheaper and more expensive desktops. However, more expensive equipment will require more insurance, which increases the shipping price.

$$\text{Bias} = \beta_2\beta_1 = \beta_{SHIPPING} \cdot \rho(r_{IN,SHIPPING}) = - \cdot + = -$$

Therefore, we would expect the omission of the shipping price to have a negative bias on other coefficients, which may be overshadowed by the omission of statistically significant feature variables. Every variable that measures a feature present in a computer will be significant with a sufficiently large data set. The omission of a multitude of these variables is not particularly distressing as I believe I have captured the few variables that are the strongest proxy for the quality of a computer.

Irrelevant variables:

All variables, except for RAM_i and HD_i (which I believe are affected by multicollinearity), are statistically significant. However, RAM_i is much less significant than would be expected. It is true that multicollinearity can cause a strong variable to take

much of the weight of less relevant variables, however, I would have expected a greater significance from RAM_i.

Incorrect functional form:

There is some evidence of incorrect functional form given the presence of unexpected serial correlation. The BLS equations and my previous price theory supported the use of the linear form. However, it is quite possible that as the "quality" of the computer goes down, people are going to value it far less than expected because it is less functional in comparison to other, newer computers. Furthermore, the "latest and greatest" of computers will command a price premium. It is for this reason that a semi-log left functional form would seem to be a better theoretical fit.

Multicollinearity:

There is very strong evidence of multicollinearity in the estimated equation. The standard errors (and therefore variances) of CPU_i, RAM_i, and HD_i, are unexpectedly high, and the t-scores of these variables are unexpectedly low. Furthermore, the value of $ß_{CPU}$ is far too high, while the values of $ß_{RAM}$ and $ß_{HD}$ are negative when their expected signs are positive. Finally, the VIF test for CPU was close to significant (VIF_{CPU}=4.7709), and the VIF test for HD was significant (VIF_{HD}=5.4704).

Serial correlation:

This model uses a cross-sectional data set, so serial correlation is unlikely. However, the Durbin-Watson test confirms with 95% (and 99%) confidence that impure positive serial correlation is present. I had previously assumed that the order of my data set was random, which would inhibit the presence of a significant Durbin-Watson test even in the face of material omitted variables or incorrect functional form. To

determine the randomness of the ordering of data set against P_i, I elected to run $P_i = \beta_0 + \beta_1 P_{i-1} + \varepsilon_i$ to determine if there was any serial correlation in the dependent variable. I received the following results:

$$\hat{P}_i = 191.3965 + 0.1920 P_{i-1}$$
$$(0.0995)$$
$$t = 1.9294$$

$$n = 99 \quad t_c \approx 1.29 \quad \bar{R}^2 = .027$$

When I randomized the order of the data set, I received:

$$\hat{P}_i = 211.7163 + 0.1135 P_{i-1}$$
$$(0.1014)$$
$$t = 1.119$$

$$n = 99 \quad t_c \approx 1.29 \quad \bar{R}^2 = .002$$

There is some evidence, then, that there is some order in the data set. A visual analysis shows that P_i appears to decrease as the observations progress. It is for this reason that I suspect the presence of either a significant omitted variable or an incorrect functional form, diagnosed by the unexpectedly low value of the d-statistic.

The two top contenders for omitted variable are feature variables and the shipping price variable. Feature variables are positively correlated with P_i, while the shipping variable is negatively correlated with P_i. Because I am intentionally omitting a wide variety of feature variables, I expect these to have the greatest effect on serial correlation. The only variables that could plausibly have an incorrect functional form are CPU_i and HD_i. RAM_i seems too insignificant to have a strong effect on the d-statistic.

Heteroskedasticity:

I ran the Park test on CPU_i, which is one measure of scale or quality of the machine. Typically, as the CPU speed increases, the price of the machine also increases.

As a result, there is greater possibility for variance in the price of the machine. The expected presence of heteroskedasticity does not particularly concern me, but it is an indication that there may be a significant omitted variable or, less likely, an incorrect functional form.

Equation conclusion:

It is clear that there are a number of material problems with this equation. I am particularly concerned with omitted variables, an irrelevant variable, multicollinearity, impure positive serial correlation, and heteroskedasticity. However, I believe that a final diagnosis of this equation is severely hampered by the presence of multicollinearity. It is for this reason that I believe that it is best to remedy the multicollinearity problem.

There are three solutions for multicollinearity: drop the redundant variable(s), create a composite variable, or do nothing. There are two possible courses of action. If one were to build this equation to attempt to predict the final selling price on eBay of a given computer's components, it would make sense to build a composite variable. However, as my model is designed to act as one component in a greater model, the specific details about a computer are less important than its overall quality. For this reason, my next change will be to drop the variable RAM_i and test for its irrelevance.

What if I did decide to create a composite variable? What would that variable look like? First, it seems foolhardy to simply sum CPU_i, RAM_i, and HD_i. The scale of each of these variables are quite different. CPU_i typically ranges from 16 to 3,060, RAM_i typically ranges from 4 to 1500, and HD_i typically ranges from 0.04 to 200. For this reason, a weighted approach seems best.

What is the best method of weighting these variables to create a composite? Let us look to the prices of the individual components for guidance. I obtained the prices

and specifications of a number of CPUs, RAM chips, and hard drives. I then obtained and averaged the per MHz, megabyte, and gigabyte price for CPUs, RAM, and hard drives respectively. On average a CPU is ~$0.0669 per MHz. A RAM chip is ~$0.2130 per megabyte. A hard drive is ~$1.4843 per gigabyte. Therefore, it makes sense to create a Q_i variable such that $Q_i=0.0669CPU_i+0.2130RAM_i+1.4843HD_i$. Because there is a value difference between new individual components and components included in a machine, I do not expect the value of ßQ_i to be equal to 1. However, this is not troublesome as long as their relative value is accurate.

Equation 2

Model specification (independent variables and functional form):

$$P_i = \beta_0 + \beta_1 CPU_i + \beta_2 HD_i + \beta_3 BC_i + \beta_4 PC_i + \varepsilon_i$$

Expected signs of coefficients:

$$P = f(\overset{+}{CPU}, \overset{+}{HD}, \overset{+}{BC}, \overset{+}{PC}) + \varepsilon_i$$

Estimated equation:

$$\hat{P}_i = 10.5279 + 0.2862 CPU_i - 1.4910 HD_i + 45.5668 BC_i + 538.1649 PC_i$$
$$(0.0391) \quad (1.1908) \quad (17.2212) \quad (86.6967)$$
$$t = 7.3186 \quad t = -1.2520 \quad t = 2.6459 \quad t = 6.2074$$

$n = 100 \quad t_c = 1.659 \quad \bar{R}^2 = .766 \quad F = 82.381 \quad d = 1.1065$

F-test (1%):

$H_0 : \beta_{CPU} = \beta_{HD} = \beta_{BC} = \beta_{PC} = 0$ $\quad F = 82.381$

$H_A : H_0$ is not true $\quad F_c = 3.52$

$F > F_c \Rightarrow$ Reject H_0

t-tests (5%):

$H_0 : \beta_{CPU} \leq 0$ $\quad t_{CPU} = \dfrac{0.2862 - 0}{0.0391} = 7.3186$

$H_A : \beta_{CPU} > 0$ $\quad t_c = 1.659$

$|t_{CPU}| > t_c \Rightarrow t_{CPU} > 0 \Rightarrow$ Reject H_0

$H_0 : \beta_{HD} \geq 0$ $\quad t_{HD} = \dfrac{-1.4910 - 0}{1.1980} = -1.2520$

$H_A : \beta_{HD} < 0$ $\quad t_c = -1.659$

$|t_{HD}| \not> t_c \Rightarrow t_{HD} \not< 0 \Rightarrow$ Cannot reject H_0

$H_0 : \beta_{BC} \leq 0$ $\quad t_{BC} = \dfrac{47.3140 - 0}{17.6137} = 2.6861$

$H_A : \beta_{BC} > 0$ $\quad t_c = 1.659$

$|t_{BC}| > t_c \Rightarrow t_{BC} > 0 \Rightarrow$ Reject H_0

$H_0 : \beta_{PC} \leq 0$ $\quad t_{PC} = \dfrac{547.8555 - 0}{89.0196} = 6.1543$

$H_A : \beta_{PC} > 0$ $\quad t_c = 1.659$

$|t_{PC}| > t_c \Rightarrow t_{PC} > 0 \Rightarrow$ Reject H_0

VIF tests:

$VIF_{CPU} = \frac{1}{1 - 0.7695} = 4.3383$ $\quad VIF_{CPU} < VIF_c$ ∴ No severe multicollinearity for CPU_i.

$VIF_c = 5$

$VIF_{HD} = \frac{1}{1 - 0.8164} = 5.4466$ $\quad VIF_{HD} > VIF_c$ ∴ Severe multicollinearity exists for HD_i.

$VIF_c = 5$

$VIF_{BC} = \frac{1}{1 - 0.3985} = 1.6625$ $\quad VIF_{BC} < VIF_c$ ∴ No severe multicollinearity for BC_i.

$VIF_c = 5$

$VIF_{PC} = \frac{1}{1 - 0.1734} = 1.2097$ $\quad VIF_{PC} < VIF_c$ ∴ No severe multicollinearity for PC_i.

$VIF_c = 5$

Durbin Watson test:

$H_0 : \rho = 0$ $\quad n = 100$ $\quad d_L = 1.59$ $\quad d = 1.1065$

$H_A : \rho > 0$ $\quad k' = 5$ $\quad d_U = 1.76$

$d < d_L$ ∴ Reject H_O, Impure positive serial correlation exists.

Park test:

Testing for: $\ln(e_i^2) = \beta_0 + \beta_1 \ln CPU_i + u_i$

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

Result: $\ln(\hat{e}_i^2) = 3.3692 + 0.7777 \ln CPU_i$

(0.2167)

$t = 3.588$

$n = 100$ $\quad t_c = 2.63$ $\quad \overline{R}^2 = .1071$

$|t_{CPU}| > t_c$ ∴ Reject H_0, Heteroskedasticity is present.

Diagnostics:

Omitted variables:

Unsurprisingly, there is still evidence of omitted feature variables.

Irrelevant variables:

Let us run the four specification criteria on RAM_i:

1. Theory: RAM_i is a measurement of computational power and speed, and is fairly correlated (r=0.687) with CPU_i. Typically, there is a balance between CPU speed and the expected amount of RAM. Its inclusion is *weakly* theoretically valid.
2. t-Test: The t-statistic for RAM_i was -0.5181—hardly significant. Even considering the potential effects of multicollinearity in this equation, the t-statistic suggests its removal.
3. $\overline{R}^2$: $\overline{R}^2$improves slightly when RAM_i is *removed* from the equation, weakly suggesting that removing RAM_i was the correct action.
4. Bias: None of the other coefficients changed in any significant way.

For these reasons, the removal of RAM_i from the equation was the correct thing to do.

Incorrect functional form:

Evidence of incorrect functional form is still present.

Multicollinearity:

There is still strong evidence of multicollinearity in this equation. VIF_{HD} is 5.4466.

Serial correlation:

Evidence of serial correlation is still present.

Heteroskedasticity:

Evidence of heteroskedasticity is still present.

Equation conclusion:

It is clear that there are still a number of material problems with this equation. I still concerned with multicollinearity, impure positive serial correlation, and heteroskedasticity. I am only somewhat surprised that the removal of RAM_i only marginally lowered the effects of multicollinearity.

At this time, it makes the greatest sense to attempt to remedy the multicollinearity issue by creating a composite variable such that Q_i=0.0669CPU_i+1.4843HD_i and running a new regression. HD_i seems just too relevant to simply drop.

Equation 3

Model specification (independent variables and functional form):

$$P_i = \beta_0 + \beta_1 Q_i + \beta_2 BC_i + \beta_3 PC_i + \varepsilon_i$$

Expected signs of coefficients:

$$P = f(\overset{+}{Q}, \overset{+}{BC}, \overset{+}{PC}) + \varepsilon_i$$

Estimated equation:

$$\hat{P}_i = \underset{}{45.5279} + \underset{(0.2241)}{2.0941 Q_i} + \underset{(17.8953)}{27.8629 BC_i} + \underset{(91.7962)}{477.7399 PC_i}$$
$$t = 9.3412 \quad t = 1.5569 \quad t = 5.2043$$

$n = 100 \quad t_c = 1.657 \quad \bar{R}^2 = .730 \quad F = 90.389 \quad d = 0.9104$

F-test (1%):

$H_0 : \beta_{CPU} = \beta_{HD} = \beta_{BC} = \beta_{PC} = 0$ $\quad F = 90.389$
$H_A : H_0$ is not true $\quad F_c = 4.02$ $\quad F > F_c \Rightarrow$ Reject H_O

t-tests (5%):

$H_0 : \beta_Q \leq 0$ $\quad t_Q = \dfrac{166.7074 - 0}{16.7308} = 9.9640$ $\quad |t_Q| > t_c \Rightarrow t_Q > 0 \Rightarrow$ Reject H_0
$H_A : \beta_Q > 0$ $\quad t_c = 1.657$

$H_0 : \beta_{BC} \leq 0$ $\quad t_{BC} = \dfrac{40.4660 - 0}{16.5029} = 2.4520$ $\quad |t_{BC}| > t_c \Rightarrow t_{BC} > 0 \Rightarrow$ Reject H_0
$H_A : \beta_{BC} > 0$ $\quad t_c = 1.657$

$H_0 : \beta_{PC} \leq 0$ $\quad t_{PC} = \dfrac{557.4210 - 0}{86.6458} = 6.4333$ $\quad |t_{PC}| > t_c \Rightarrow t_{PC} > 0 \Rightarrow$ Reject H_0
$H_A : \beta_{PC} > 0$ $\quad t_c = 1.657$

VIF tests:

$VIF_Q = \dfrac{1}{1 - 0.3126} = 1.4547$ $\quad VIF_Q < VIF_c \Rightarrow$ No severe multicollinearity for Q_i.
$VIF_c = 5$

$VIF_{BC} = \frac{1}{1 - 0.2893} = 1.4070$ $\quad VIF_{BC} < VIF_c$ ⇒ No severe multicollinearity for BC_i.

$VIF_c = 5$

$VIF_{PC} = \frac{1}{1 - 0.1020} = 1.1135$ $\quad VIF_{PC} < VIF_c$ ⇒ No severe multicollinearity for PC_i.

$VIF_c = 5$

Durbin Watson test:

$H_O : \rho = 0 \quad n = 100 \quad d_L = 1.61 \quad d = 1.237$

$H_A : \rho > 0 \quad k' = 5 \quad d_U = 1.74$

$d < d_L$ ⇒ Reject H_O, Impure positive serial correlation exists.

Park test:

Testing for: $\ln(e_i^2) = \beta_0 + \beta_1 \ln(\ln Q_i) + u_i$

$H_O : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

Result: $\widehat{\ln(e_i^2)} = 3.722 + 3.4432 \ln(\ln Q_i)$

(0.8399)

$t = 4.099$

$n = 100 \quad t_c = 2.63 \quad \overline{R}^2 = .0593$

$|t_{CPU}| > t_c$ ⇒ Reject H_0, Heteroskedasticity is present.

Diagnostics:

Omitted variables:

Unsurprisingly, there is still evidence of omitted feature variables.

Irrelevant variables:

There is no evidence of irrelevant variables.

Incorrect functional form:

Evidence of incorrect functional form is still present.

Multicollinearity:

There is no evidence of severe imperfect multicollinearity in this equation.

Serial correlation:

Evidence of serial correlation is still present.

Heteroskedasticity:

Evidence of heteroskedasticity is still present.

Equation conclusion:

Given the fact that the Durbin Watson and Park tests are still suggesting problems, I feel that it would be valuable to fix the functional form of Q and consider the next step from there. To fix the functional form problem, I intend to run a semi-log right on Q_i. This seems to be the best theoretical fit because the quality of the computer is less important as long as it is recent. However, the lower in quality the computer gets, the more rapidly its value should drop off as it is becomes less useful to a greater number of people. At some Q_i, we can even expect low quality to have a negative effect on the value of the machine. With luck, this should reduce serial correlation and heteroskedasticity.

Equation 4

Model specification (independent variables and functional form):

$$P_i = \beta_0 + \beta_1 \ln Q_i + \beta_2 BC_i + \beta_3 PC_i + \varepsilon_i$$

Expected signs of coefficients:

$$P = \beta(\ln \overset{+}{Q}, \overset{+}{BC}, \overset{+}{PC}) + \varepsilon_i$$

Estimated equation:

$$\hat{P}_i = \underset{}{-441.6925} + \underset{(16.7308)}{166.7074 \ln Q_i} + \underset{(16.50295)}{40.46608 BC_i} + \underset{(86.6458)}{557.4210 PC_i}$$
$$t = 9.9640 \qquad t = 2.4520 \qquad t = 6.433325$$

$n = 100 \quad t_c = 1.657 \quad \bar{R}^2 = .746 \quad F = 98.419 \quad d = 1.237$

F-test (1%):

$H_0 : \beta_{CPU} = \beta_{HD} = \beta_{BC} = \beta_{PC} = 0$ $\qquad F = 90.389$

$H_A : H_0$ is not true $\qquad F_c = 4.02$

$F > F_c \Rightarrow$ Reject H_O

t-tests (5%):

$H_0 : \beta_Q = 0$
$H_A : \beta_Q > 0$

$$t_Q = \frac{2.0941 - 0}{0.2241} = 9.3412$$
$t_c = 1.657$

$|t_Q| > t_c \Rightarrow t_Q > 0 \Rightarrow$ Reject H_0

$H_0 : \beta_{BC} = 0$
$H_A : \beta_{BC} > 0$

$$t_{BC} = \frac{27.8629 - 0}{17.8953} = 1.5569$$
$t_c = 1.657$

$|t_{BC}| \not> t_c \Rightarrow t_{BC} > 0 \Rightarrow$ Do not reject H_0

$H_0 : \beta_{PC} = 0$
$H_A : \beta_{PC} > 0$

$$t_{PC} = \frac{477.7399 - 0}{91.7962} = 5.2043$$
$t_c = 1.657$

$|t_{PC}| > t_c \Rightarrow t_{PC} > 0 \Rightarrow$ Reject H_0

VIF tests:

$$VIF_Q = \frac{1}{1 - 0.4150} = 1.7094$$
$VIF_c = 5$

$VIF_Q < VIF_c \Rightarrow$ No severe multicollinearity for Q_i.

$VIF_{BC} = \frac{1}{1 - 0.3560} = 1.5528$ $\quad VIF_{BC} < VIF_c \Rightarrow$ No severe multicollinearity for BC_i.

$VIF_c = 5$

$VIF_{PC} = \frac{1}{1 - 0.1475} = 1.1730$ $\quad VIF_{PC} < VIF_c \Rightarrow$ No severe multicollinearity for PC_i.

$VIF_c = 5$

Durbin Watson test:

$H_O : \rho = 0 \quad n = 100 \quad d_L = 1.61$
$H_A : \rho > 0 \quad k' = 5 \quad d_U = 1.74 \quad d = 0.9104$

$d < d_L \Rightarrow$ Reject H_O, Impure positive serial correlation exists.

Park test:

Testing for: $\ln(e_i^2) = \alpha_0 + \alpha_1 \ln Q_i + u_i$

$H_O : \alpha_1 = 0$
$H_A : \alpha_1 \neq 0$

Result: $\ln(\hat{e}_i^2) = 4.334 + 1.015 \ln Q_i$
(0.2263)
$t = 4.489$

$n = 100 \quad t_c = 2.63 \quad \bar{R}^2 = .1621$

$|t_{CPU}| > t_c \Rightarrow$ Reject H_0, Heteroskedasticity is present.

Diagnostics:

Omitted variables:

Unsurprisingly, there is still evidence of omitted feature variables.

Irrelevant variables:

While the t-score for BC_i implies that it is not significant, I believe that the theory is strong enough for its retention. Furthermore, previous equations described it as being significant. I do not see a point to running a separate equation without BC_i at this time.

Incorrect functional form:

I do not believe that there are problems with the functional form of this equation.

Multicollinearity:

There is no evidence of severe imperfect multicollinearity in this equation.

Serial correlation:

Evidence of serial correlation is still present, although the Durbin-Watson statistic has improved with the change in functional form for Q_i. This supports the hypothesis that the change in functional form was correct.

Heteroskedasticity:

Evidence of heteroskedasticity is still present.

Equation conclusion:

At this time, it makes the most sense to run weighted least squares with Q as the weighting series. It no longer seems necessary to continue to run VIF-tests on the remainder of the regressions, as I have already established that there is very little possibility of severe imperfect correlation with the current set of independent variables.

Equation 5

Model specification (independent variables and functional form):

$$\frac{P_i}{Q} = \square_1 + \frac{\square_0}{Q} + \square_1 \frac{\ln Q_i}{Q} + \square_2 \frac{BC_i}{Q} + \square_3 \frac{PC_i}{Q} + u_i$$

Estimated equation:

$$\hat{P}_i = \square 664.2899 + \underset{(35.065)}{221.9477 \ln Q_i} + \underset{(16.8196)}{14.8901 BC_i} + \underset{(86.6458)}{636.3127 PC_i}$$
$$t = 6.3295 \qquad t = .8852 \qquad t = 13.4242$$

$n = 100 \quad t_c \square 1.657 \quad \overline{\mathrm{R}}^2 = .939 \quad \mathrm{F} = 109.04 \quad \mathrm{d} = 1.58$

Durbin Watson test:

$$H_O : \square \square 0 \quad n = 100 \quad \square \quad d_L = 1.61 \quad d = 1.585$$
$$H_A : \square > 0 \quad k' = 5 \quad \square \quad d_U = 1.74$$

$d < d_L \square$ Reject H_O, Impure positive serial correlation exists.

Diagnostics:

Omitted variables:

Unsurprisingly, there is still evidence of omitted feature variables.

Irrelevant variables:

While the t-score for BC_i implies that it is not significant, I believe that the theory is strong enough for its retention. Furthermore, previous equations described it as being significant. I do not see a point to running a separate equation without BC_i at this time.

Incorrect functional form:

I do not believe that there are problems with the functional form of this equation.

Multicollinearity:

There is no evidence of severe imperfect multicollinearity in this equation.

Serial correlation:

Evidence of serial correlation is still present, although the Durbin-Watson statistic has improved with the switch to WLS.

Heteroskedasticity:

There should be no evidence of heteroskedasticity in this equation.

Equation conclusion:

It does not make sense to run GLS, because I believe that the serial correlation of the error term is minor. It does not make sense to make a drastic change to the functional form of this equation to determine whether or not the included variables are *truly* significant when their relevance has already been proven.

It is my belief that equation 5 is the best possible equation given the available data and the framework of this assignment. Because of the variables limitations imposed by this assignment, I have omitted over a dozen feature-related variables that have been shown to be significant in other data sets. However, I believe that the variables I have chosen are the most theoretically valid of the 23 variables presented by the BLS. I do not believe that the shipping price variable is significant enough to warrant its attempted inclusion, which may actually cause further bias due to the limitations imposed by its collection as mentioned earlier.

In conclusion, I believe that an econometric model of computer prices is extremely uninteresting when used to analyze current computers and prices. Such a model attempts to quantify the value of coefficients whose true values are easy to discover in the real world—in the components aisle at the local Fry's electronics.

However, historical trends in computer prices would be quite interesting to analyze, and I believe that this equation would be an extremely useful starting point.