An Analysis of 90s Computer Prices & Specifications

Given Sandamela and Victor Ma

Statistics 2: Research Project - Final Report

Data Key: `https://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/computers.html`

Data Set Download: `https://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/computers.csv`

Related Articles:

`https://www.jstor.org/stable/25146432`

`https://conference.nber.org/confer/2000/si2000/berndt.pdf`

Introduction

This research project aims to perform an in-depth analysis of the various specifications and prices of computers that were being sold from during the early 1990s. We wanted to base our research in this topic, because the world of tech continues to be a very interesting topic of discussion. Advancements in computer processing, graphics, and data storage have allowed the average person to get their hands on extremely powerful computers at decent prices. We are interested in looking at the difference the average price of computers over a certain time period. Furthermore, we want to find specifications such as manufacturers, hard drive size, clock speed, RAM, etc. that would help accurately predict the price of a computer. We believe that the advances in computing and memory storage during the 90s would give way to some rather interesting findings. However, it may potentially be difficult to generalize the results and interpret them for future utilization. In fact, the research article by Ernst R. Berndt and others state that "In the context of computers, due in part to rapid technological progress, it has proven to be somewhat difficult to obtain system performance measures that are valid across models and over time" (Berndt et al. 12). For our study, we decided to use a data set from the Vincent Arel-Bundock repository that also has a research article associated with it. We intend to utilize this article sparingly to help us better understand our own analyses and make comparisons to their conclusions when needed. We are also looking to utilize other articles of similar research that would help us understand our findings.

Overall, our goal is to utilize several statistical strategies in order to formulate the most accurate models for predicting prices of computers in the early 90s. Our main objective would be to fit the best multiple linear regression model using methods such as step-wise variable selection.

Before formal analysis is done, we hypothesize that clock speed, hard-drive space, and screen size are the most statistically significant predictors of increasing the price of a computer. However, due to advancements in computer science during the 1990s, we predict that as time increases, overall price of a computer decreases.

Materials and Methods

Our data came from a collection of R Datasets from vincentarelbundock.github.io. The `computer` dataset contains information about the price of computers sold between the years 1993 and 1995 based on software

and hardware design properties. These properties are defined in Table 1 below. The data is a cross-section from 1993 to 1995 with 11 variables and 6259 observations where each observation or row represents a personal computer being sold in the United States.

Table 1: Definitions of Variables

Name	Variable	Role	Definition	Values
rownames	NA	NA	the observational unit id	NA
price	Response	Quantitative	price of the PC in USD	USD
speed	Explanatory	Categorical	clock speed in MHz	MHz
hd	Explanatory	Categorical	size of hard drive in MB	Megabytes
ram	Explanatory	Categorical	size of Ram in in MB	Megabytes
screen	Explanatory	Categorical	size of screen in inches	Inches
cd	Explanatory	Categorical	is CD-ROM present?	yes/no
multi	Explanatory	Categorical	is a multimedia kit (speakers, sound card) included?	yes/no
premium	Explanatory	Categorical	is the manufacturer was a 'premium' firm (IBM, COMPAQ)?	yes/no
ads	Explanatory	Categorical	number of 486 price listings for each month	Number value
trend	Explanatory	Quantitative	time trend from January 1993 to November 1995	year

The `computers` dataset did not have any missing values, so we did not to remove any observations or columns from the dataset. We kept every row from the original dataset of 6259 by 11. We used all 11 variables for our EDA and analysis.

We are interested in investigating the most optimal model for predicting the price of a personal computer. We decided to perform variable selection to determine which variable were optimal predictors of the price of personal computers. We used both forward and backward variable selection using best subset regression. We will elaborate on our findings in the 'Discussion' section. To understand our data, we calculated the summary statistics for computers to analyze the summaries for the variable `premium`, `price`, `speed` etc.

Furthermore, we performed EDA on our `computers` dataset. We plotted a scatterplot for `price` and `trend`. Similarly, we plotted a scatterplot for `price` and `premium` to identify whether premium or trend is a better predictor of computer price. We then decided to plot joint scatterplot of `price` and `trend` with a 'ggplot' color aesthetic set to `premium`. This led us to explore applying a logistic regression model on `premium` to predict

To build our logistic regression model we had to create an indicator variable for `premium` since on the original dataset it is a categorical yes or no variable. We set yes for `premium` as 1 and no for `premium` as 0.

Results

When it came to analyze the numerical summaries of each of the variables, we found some interesting results. For example, from the 6259 observations, the average price of a computer was 2219.57 USD. The minimum price was 949, while the maximum price was 5399. Quite a large difference! Furthermore, the average clock speed in MHz was 52, with the maximum being 100 and the minimum being 25. The average hard drive size was 416 megabytes, with the maximum being 2100 and the minimum being 80. However, the median was 340 megabytes. The average median ram was 8 megabytes, with the minimum being 2 and maximum being 32. Finally, the average screen size was 14.6 inches, with the minimum being 14, and maximum being 17. It is important to note that the median was also 14.

The proportion of computers that came with a CD-ROM drive was 46.46%. The proportion of computers that came with a multimedia kit was 13.94%. The proportion of computers that were manufactured by a 'premium' firm (IBM or COMPAQ) was 90.22%. This will be important to remember for later. It seems like the market (at least seen on PC Magazine) was dominated by these premium manufacturers.

First, we decided to fit a linear regression model on price with simply just trend. We found it to be statistically significant with a p-value $< 2.2e\text{-}16$. However, the R-squared value is a measly 0.039, as expected. We moved

onto the model with all variables predicting price. We also found it to be statistically significant with a p-value $<$ 2.2e-16. Note that CD, Multi, and Premium are all predicting if they do indeed have these features. From the computed output, we are able to create the model:

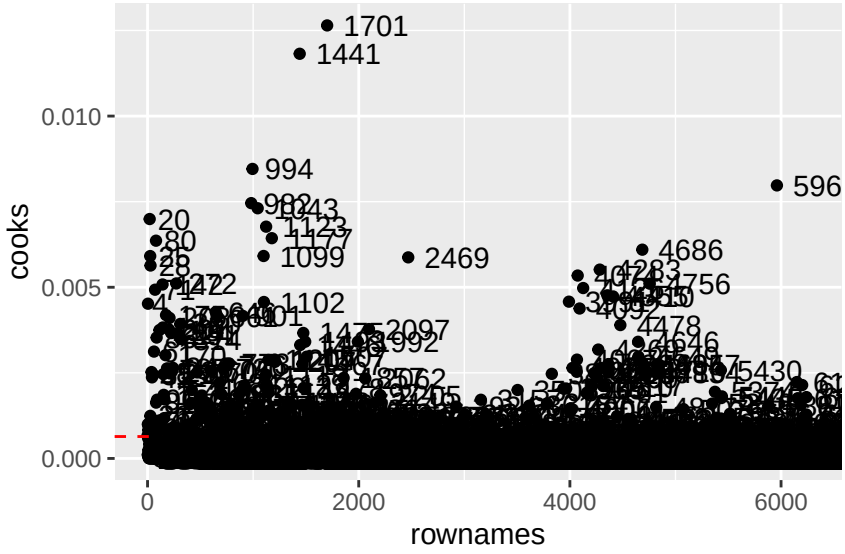
$$\widehat{Price} = 307.98 + 9.32 \times Speed + 0.78...$$
$$... \times HD + 48.25 \times RAM + 123.08 \times Screen + 60.91 \times CD + ...$$
$$...104.32 \times Multi - 509.22 \times Premium + 0.65 \times Ads - 51.84 \times Trend$$

We moved onto using BIC, Mallow's Cp, and Adjusted R^2 criterion combined with step-wise selection to try and find the best fit multiple linear regression model. With BIC, we saw that speed, ram, and trend were signifcant, with premium being questionable. Mallow's Cp cemented the variables ram and trend being significant, while speed and premium were quite close to being solid predictors. Finally, the adjusted R^2 criteria showed us the same story. Therefore, we were able to move on and create a model with the four aforementioned variables. We computed the VIF values, and each variable saw a value of 1.26 or less, which indicates no problematic multicollinearity. The model we curated had a p-value of $<$ 2.2e-16 and an adjusted R-squared value of 0.6995. Here is the model:

$$\widehat{Price} = 2127.24 + 10.38 \times Speed + 75.65 \times RAM - 480.04 \times Premium - 40.27 \times Trend$$

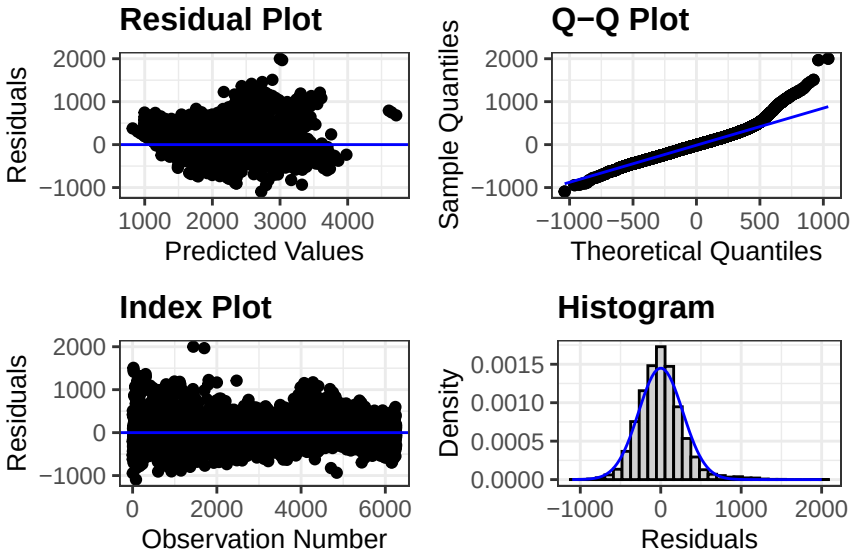
```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```



One important thing to note is that roughly 5.54% of observations are above Cook's distance. Fortunately, most of these observations seem quite normal, and wouldn't be any problem in our future analyses. Interestingly, there top two points have values much higher than the other observations at 0.0126 and 0.0118, while the 3rd and 4th highest observations have values of 0.00846 and 0.00798. These are data entries 1701 and 1441, and they are premium computers both with prices of 4999 USD. Their prices seem to be indicative of their high values, perhaps indicating corporate price gouging of low-spec computers. We note that both of these computers have 66 MHz, 525 Megabytes of HD, 8 Megabytes of RAM, and 17 inch screens. However, we cannot be sure of the reasoning behind these higher prices. Therefore, we continue moving onto our main objectives and analyses.

```
##      ram    trend    speed  premium   screen       hd      ads    multi
## 2.974628 2.022790 1.265364 1.109388 1.081644 4.207395 1.217218 1.290568
##       cd
## 1.859370
```



Now we have essentially have two models. A nine variable model and a four variable model. The most significant predictors being Speed, RAM, Premium, and Trend. We decide to explore further and look at the interaction term Premium:Trend. Premium indicating that the computer was indeed manufactured by a 'premium' firm. This is the model we calculated:

$$\widehat{Price} = 1853.27 + 10.31 \times Speed + 75.93 \times RAM - 191.08 \times Premium - 21.80 \times Trend - 19.33 \times Premium : Trend$$

We recognized that the model has a p-value of 2.2e-16 and an Adjusted R-squared value of 0.7024. Each predictor seemed to be significant. Furthermore, we noticed that the Premium:Trend interaction term indicated an additional -19.33 decrease in price for each increase in month. We believe that this model could be comparable to our nine variable model, a simpler version perhaps.

Discussion

Our goal was to find a model with few variables that could accurately predict the price of computers in the 90s while also accounting for a high amount of variability. Furthermore, we wanted to see if there were any interesting conclusions that we could make in regards to personal computer prices during the 90s.

We tried to fit various different models, whether it was with interaction terms, four variables, nine variables, just trend, and just about everything. Furthermore, for simplifying the final model, we used a step-wise method for selecting predictors, checked VIF values, looked at Cook's distance values, and used BICs, Mallow's Cp, Adjusted R^2 criterion. However, we still believe that all nine predictors are critical to the most accurate model. We observed that a significant amount of the variability in personal computer prices could be explained by our model with the nine variables.

It is important to note that we tried seeing if we could replicate this result using subset regularization. We selected for the best two subsets on the `computer` dataset. Interestingly, all the subsets showed that the best model had the same variables from the step function we performed earlier. We concluded that the best model to predict price had, at the very least, the variables Speed, RAM, Premium, and Trend.

An important thing to note is that we tested the significance of a model with the interaction term Premium:Trend, as well as Speed, RAM, Premium, and Trend. It was indicative that the model was very good for predicting price, with a comparable R^2 value compared to our model with nine predictors. This does not mean that other statistically significant predictors no not exist. Our tests indicate that the nine previously mentioned variables in the data are crucial to predicting price. However, we believe that the five variables that we decided to include in this interaction model are the most indicative of a change in price. Notably, we saw that the interaction term Premium:Trend showed that for each increase in month, a premium firm's computer listing would cost 41.1408 less, compared to a non-premium firm's listing costing just 21.80 less. We believe that this model could be crucial if were specifically looking at the differences between premium and non-premium firms. Ultimately, if we wanted a much simpler model yet comparably accurate model for predicting the price of a computer in the early 90s, we would have:

$$\widehat{Price} = 1853.27 + 10.31 \times Speed + 75.93 \times RAM - 191.08 \times Premium - 21.80 \times Trend - 19.33 \times Premium : Trend$$

In the analysis by T. Stengos and E. Zacharias, they showed that there exist "nonlinearities in the intertemporal pricing of PC components" (Stengos and Zacharias 16). Furthermore, they did not find evidence that "PC manufacturers charge higher component prices for their top performance PCs during the early 90s. Continuing our analysis, we saw that for every additional month in `trend` the average price of a personal computer on PC Magazine decreases by 51.85 USD given that other variables are kept constant. This statistic is supported by the research article by Ernst R. Berndt et al., as they concluded,"Although mobile PC price declines were rather modest in the 1980s (around -8%), they picked up from 1989 to 1994 (-23%), and even more so in the late 1990s (-32%)" (Berndt et al. 21). This could potentially help explain our relatively low R^2 value, meaning that there are other potential variables that could help better explain the pricing of personal computers in the 90s. Ultimately, when it comes to applying our results, it would perhaps be best to generalize to personal computers being sold during the early 90s in the United States. We cannot apply these findings to other time periods or decades, because of the nature of our topic. Shown by the supporting articles, the rapid development in computing allow for better or similar computer hardware for a decreased price every year. Furthermore, we should only generalize our results to the United States; it has been widely known that it is a leader in the technology sector. Interestingly, an article by the New York Times in 1998 indicate that only 23% of households had a computer in Europe while 45% of households in the U.S. did (Reier). With places such as Silicon Valley continually innovating and creating newer and better products, we must recognize the uniqueness of our dataset when we generalize our analyses.

Future research:

Our best model had nine predictors with only $R^2 = 0.76$. Our model has quite a few predictors, yet 24% of the variability in the prices of computers is still unexplained. In future research it would be interesting to see if we could simplify our model further. We could explore the model using `VIP` to analyze the most important variables in the model. To assess any multicollinearity from our best model, we used `vif` and found that all our variables had a VIF value of less than 5. However, `hd` had a value of 4.21, which means, we could simplify our best model further.

Limitations

A challenge with this dataset was how balanced it was. We calculated that only 9.8% of the computers were non-premium, and 90.2% of the computers were classified as premium. We do not know exactly the metric in which these computers were classified as premium or non-premium other than if they were from IBM and COMPAQ. In future research, it would be helpful to have cross validation to create training and testing splits. We do not know whether this would help balance the data, however, we know that it was an issue in our EDA.

Annotated Appendix

Our data did not have any missing values, and we did not have to filter the data. We used the entire original dataset for our analysis. Our Q-Q Plot showed a few outliers, but this was not as significant. Furthermore, our histogram of residuals was mostly normal, hence we chose to not remove any outliers. Most of our residuals are centered at zero.

To build our models, we used step-wise variable selection in both directions. Furthermore, we also used R^2, Mallow's Cp, and BIC criterion to find the best model. All of these models gave us the same variables for the most optimal model. Our best model was the one we specified earlier with nine variables. We used this model to access our conditions.

Conditions:

- *Linearity.* Residual plot is mostly patternless around 0. Points are about evenly spread above and below 0. There are a couple outliers on the right and a couple above 0.
- *Independence of residuals.* Unknown, since we do not know exactly how the sample of personal computers was collected.
- *Normality of residuals.* The Normal QQ plot shows extreme outliers on the right end (standardized residuals well beyond 2).
- *Equal variance in residuals.* The trend in the Index plot is not constant. There is greater variability for smaller fitted values compared to higher observations.
- *Sample.* We do not know exactly how the sample for personal computers was collected, so we cannot make claims about a population for generalization. Since the explanatory variable (price) was not randomly assigned to a computer, we cannot make claims of causation.

Citations:

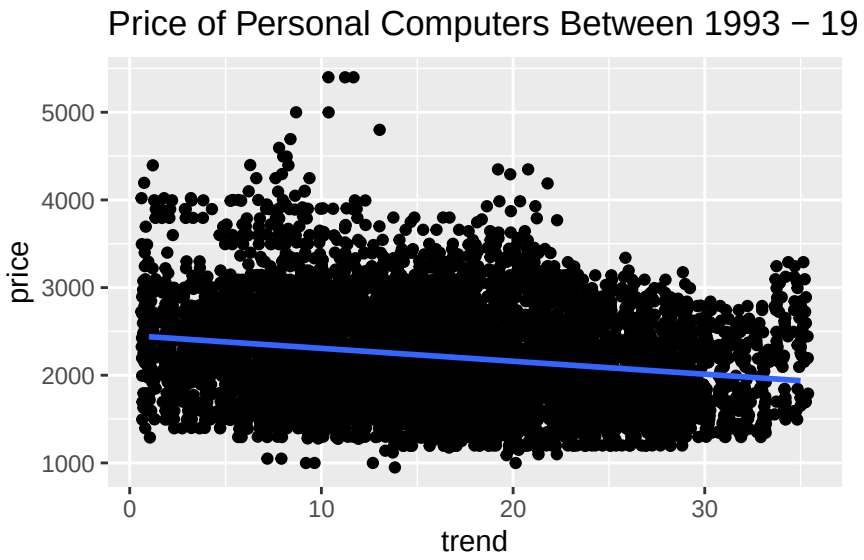
Stengos, T., and E. Zacharias. "Intertemporal Pricing and Price Discrimination: A Semiparametric Hedonic Analysis of the Personal Computer Market." Journal of Applied Econometrics, vol. 21, no. 3, 2006, pp. 371–386, https://doi.org/10.1002/jae.828.

Berndt, Ernst R, and Neal J Rappaport. "Price and Quality of Desktop and Mobile Personal Computers: A Quarter-Century Historical Overview." American Economic Review, vol. 91, no. 2, May 2001, pp. 268–273, https://doi.org/10.1257/aer.91.2.268.

1998 New York Times Article:

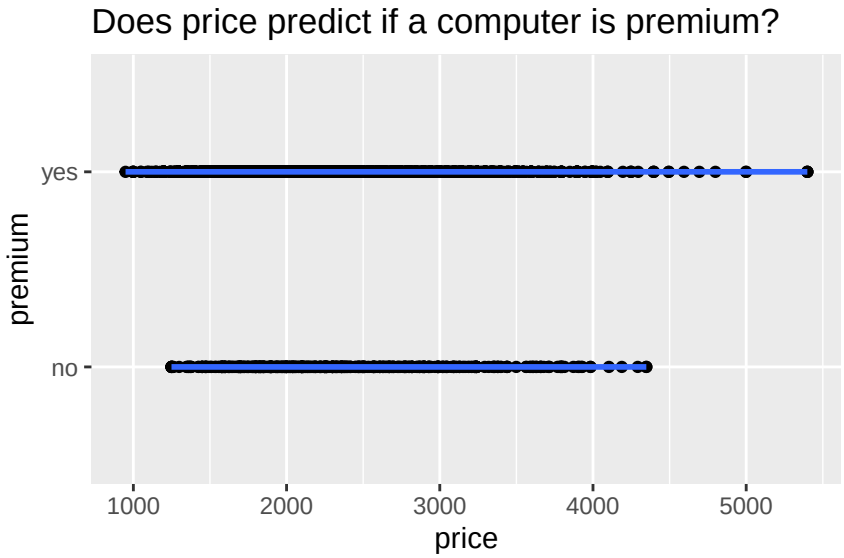
nytimes.com/1998/03/19/news/computers-in-us-do-far-more-and-are-cheaper-why-europeans-lag-in-using.html

```
## `geom_smooth()` using formula = 'y ~ x'
```



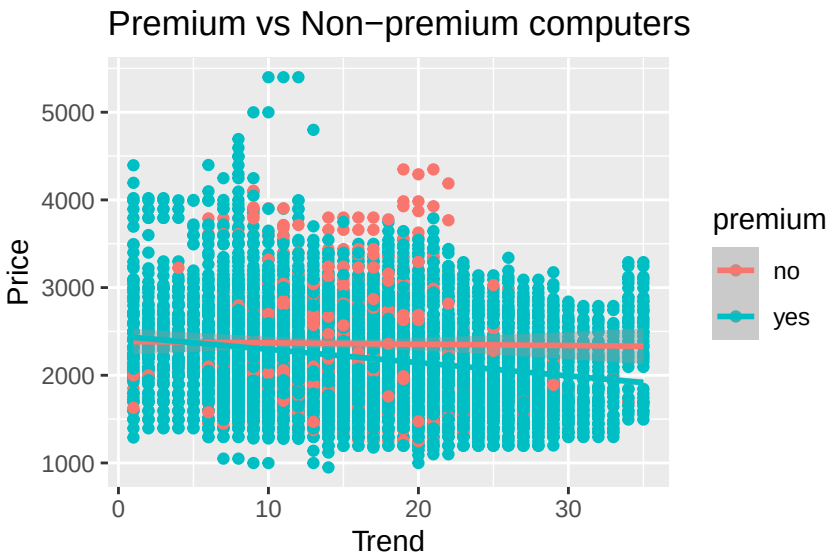
Earlier, we explored the effect of trend (increase in time, specifically a month) on the price of a computer from 1993 to 1995. We recognized a moderate negative relationship.

```
## `geom_smooth()` using formula = 'y ~ x'
```

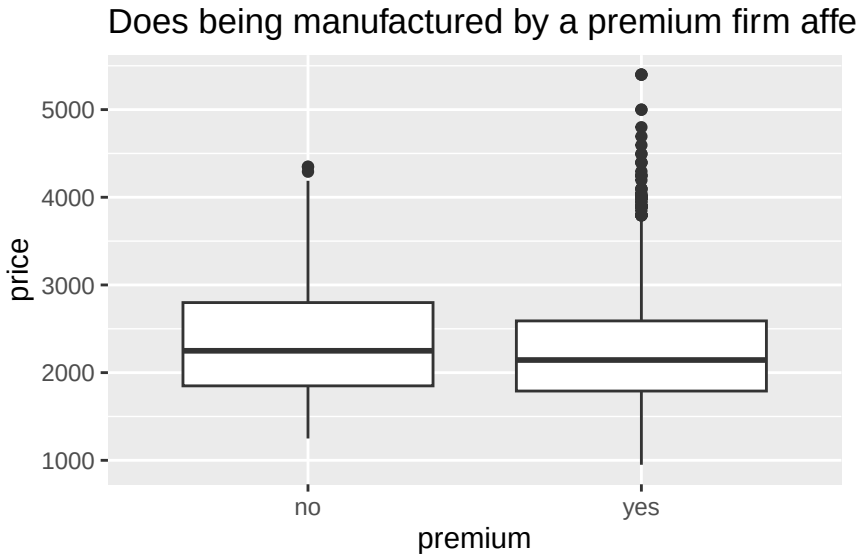


Later on, we explored the effect of premium-ness on price (premium-ness indicating that the manufacturer was IBM or COMPAQ, according to the dataset resources). We recognized that there are a decent amount of outliers in the premium group, possibly indicating that premium companies tend to price their computers at a . . . premium

```
## `geom_smooth()` using formula = 'y ~ x'
```



We were led to explore a coded scatterplot of price predicted by premium, which raised even more questions. We recognized that non-premium has a less steep line compared to premium. Therefore, there is a possibility that the price decrease over time in premium products is greater than the non-premium products.



Lastly, we explored the boxplot that cleared up many of our previous questions. The median price of non-premium products is technically higher than premium products. However, the premium group shows multiple outliers in the box plot that influences the data.

Thank you for taking the time to read our research on computers during the early 1990s. We hope our analyses were insightful for you.

- Given Sandamela and Victor Ma