Predicting wine prices before they hit the market

DANIEL ADIELE

2024-02-22

Learning Outcomes

In this project I will replicate the results from a famous paper that tried to predict wine prices before they hit the market.

Skills and learning outcomes I gained from the project.

- Perform Multiple Regression in R
- Use quadratic terms in multiple regression equations
- Perform F-test to compare models
- Correct Standard Errors for Heteroscedacity
- Use ggplot2 and matplot to graph scatter plots and regression lines in R
- Use `modelsummary` to display regression tables
- Use `margins` command to calculate marginal effects
- Use `aggregate` to calculate group means

Wine Prices Futures

One economist was a wine connoisseur, but he hated paying too much for good wine. One day he decided that there needs to be better way to predict the quality of wine without having to rely on the opinion of "experts". For his own amusement and frugality, he developed an econometric method to predict the price and quality of Bordeaux wines.

"Predicting the quality and prices of Bordeaux wines" Orley Ashenfelter (Princeton University) No 37297, Working Papers from American Association of Wine Economists

Abstract: Bordeaux wines have been made in much the same way for centuries. This article shows that the variability in the quality and prices of Bordeaux vintages is predicted by the weather that created the grapes. The price equation provides a measure of the real rate of return to holding wines (about 2–3% per annum) and implies far greater variability in the early or 'en primeur' wine prices than is observed. The analysis provides a useful basis for assessing market inefficiency, the effect of climate change on the wine industry and the role of expert opinion in determining wine prices.

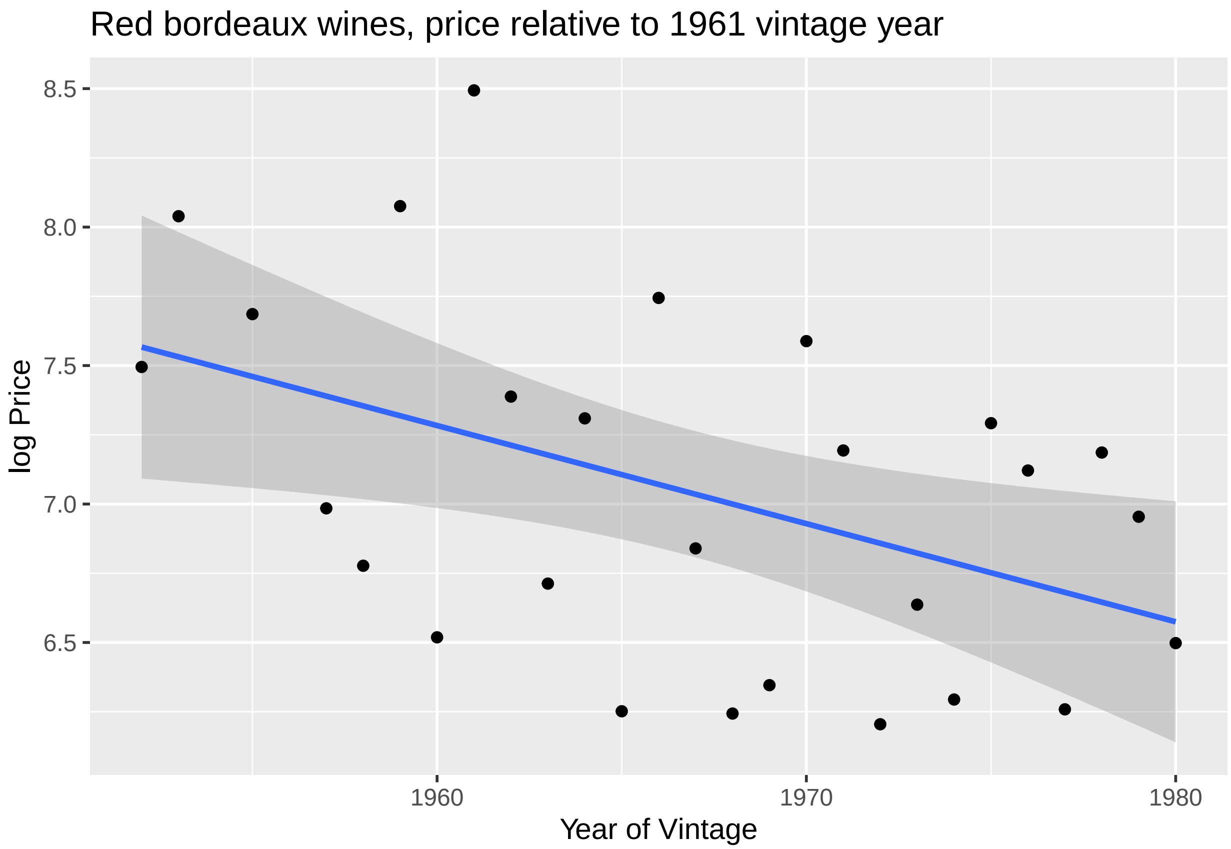
The wine.csv file contains 27 red Bordeaux vintages. The data is the same data15 originally employed by Ashenfelter, Ashmore, and Lalonde (1995), except for the inclusion of the variable Year, the exclusion of NAs and the reference price used for the wine. Each row has the following variables:

Year: year in which grapes were harvested to make wine. Price: logarithm of the average market price for Bordeaux vintages according to a series of auctions. The price is relative to the price of the 1961 vintage, regarded as the best one ever recorded. WinterRain: winter rainfall (in mm). AGST: Average Growing Season Temperature (in Celsius degrees). HarvestRain: harvest rainfall (in mm). Age: age of the wine, measured in 1983 as the number of years stored in a cask. FrancePop: population of France at Year (in thousands).

1. **Recreate Figure 1:** Use ggplot2 to recreate Figure 1.

```
wine <- read.csv("https://raw.githubusercontent.com/egarpor/handy/master/datasets/wine.csv")
# Create a scatter plot with a simple regression line
# going through it of the log price and year of vintage.
#
wine %>% ggplot(aes(x=Year ,y=Price )) + geom_smooth(method =lm ) +
  geom_point()+labs(title = "Red bordeaux wines, price relative to 1961 vintage year") + xlab ("Year of
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



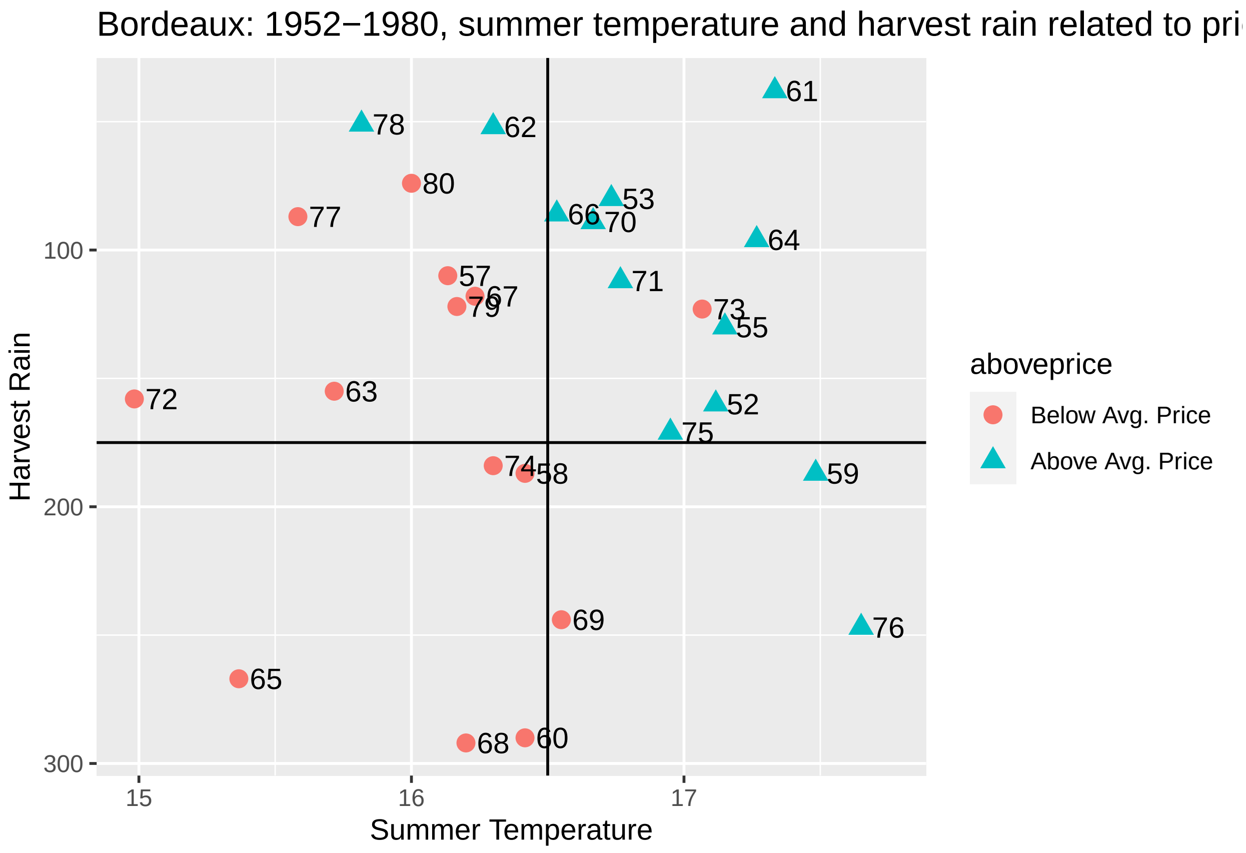
##tidyverse allows you think from left to right.

Be sure to label your Y and X axis. Be sure to include the regression line.

2. **Recreate Figure 2** Use ggplot2 to recreate Figure 2. This is a scatter plot of harvest rain and summer temperature. There are some additional pieces of information that are included. Use different colors for label wine that are above the average price and below the average price. Also include a data point label using the year of vintage.

```
avgprice <- mean(wine$Price)
wine$aboveprice <- 0
wine$aboveprice[wine$Price > avgprice] <- 1
wine$aboveprice <- factor(wine$aboveprice, labels = c("Below Avg. Price", "Above Avg. Price"))
```

```
wine %>% ggplot(aes(x=AGST ,y=HarvestRain ,label =Year - 1900)) + geom_point(aes(shape = aboveprice, col
```



3. **Recreate Table 3:** Estimate the first two regressions in Table3. You can assume homoscedastic standard errors for the first two regressions. Next, re-estimate the second regression, but use robust standard errors.

```
reg1 <- lm(Price~Age, data = wine)
reg2 <- lm(Price~Age+AGST+HarvestRain+WinterRain, data = wine)
modelsummary(list(reg1,reg2,reg2), vcov = c("iid", "iid", "robust"), coef_rename = c("Age"="Age of vinta
```

Manufacturing Costs

You are a consultant for a manufacturing company. The manufacturing company is interested in estimating its cost function.

$$Cost = F(Quantity)$$

.

4. First, run a simple linear regression. Next, run another regression with a quadratic term for quantity. Use the `I()` to add the quadratic term. Finally, add a third regression that includes a cubic term of quantity along with the quadratic term. Report these three regessions using `modelsummary`.

	(1)	(2)	(3)
(Intercept)	6.469	−3.652	−3.652
	(0.247)	(1.688)	(2.137)
Age of vintage	0.035	0.024	0.024
	(0.014)	(0.007)	(0.007)
Average temperature over growing season (April-september)		0.616	0.616
		(0.095)	(0.124)
Rain in september and August		−0.004	−0.004
		(0.001)	(0.001)
Rain in the months preceeding the vintage (October-March)		0.001	0.001
		(0.000)	(0.001)
Num.Obs.	27	27	27
R2	0.212	0.828	0.828
R2 Adj.	0.180	0.796	0.796
AIC	50.6	15.6	15.6
BIC	54.5	23.4	23.4
Log.Lik.	−22.307	−1.796	−1.796
F	6.725	26.390	23.503
RMSE	0.55	0.26	0.26
Std.Errors	IID	IID	HC3

```r
EconomiesOfScale <- read.csv("~/Downloads/MSBA 650/MSBA 650 WEEK 1/EconomiesOfScale.csv")
reg3 <- lm(TotalCost ~ Quantity, data = EconomiesOfScale)
reg4 <- lm(TotalCost ~ Quantity + I(Quantity^2), data = EconomiesOfScale)
reg5 <- lm(TotalCost ~ Quantity + I(Quantity^2)+I(Quantity^3), data = EconomiesOfScale)
modelsummary(list(reg3, reg4, reg5), vcov = c("robust", "robust", "robust"), coef_rename = (coef_rename
```

THE ONE WITH THE SQUARE (Qunatity with the square) IS BETTER ONE, WHY? BECAUSE IT IS SMALLER AND USES LESS VARIABLES

WHENEVER YOU ARE STOCK WITH TWO MODELS YOU WANT TO PICK THE VARIABLE THAT IS SMALLER AND STILL EXPLAINS YOU VARIABLES VERY WELL.

5. Perform an F-test to see if adding the cubic term significantly improves the regression compared to the quadratic regression.

```r
# We can do a variety of F-test in R
# Since we are just comparing two regessions, we will use anova
anova(reg4,reg5)
```

```
## Analysis of Variance Table
##
## Model 1: TotalCost ~ Quantity + I(Quantity^2)
## Model 2: TotalCost ~ Quantity + I(Quantity^2) + I(Quantity^3)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    997 512889
## 2    996 512873  1    15.796 0.0307  0.861
```

	(1)	(2)	(3)
(Intercept)	96.089***	73.626***	71.906***
	(2.553)	(5.045)	(9.783)
Quantity	16.164***	26.843***	28.135***
	(0.561)	(2.242)	(6.860)
(Quantity^2)		−1.161***	−1.452
		(0.243)	(1.494)
(Quantity^3)			0.020
			(0.101)
Num.Obs.	1000	1000	1000
R2	0.472	0.481	0.481
R2 Adj.	0.471	0.480	0.480
AIC	9102.2	9085.9	9087.9
BIC	9116.9	9105.6	9112.4
Log.Lik.	−4548.098	−4538.968	−4538.952
F	831.217	497.918	339.666
RMSE	22.85	22.65	22.65
Std.Errors	HC3	HC3	HC3

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

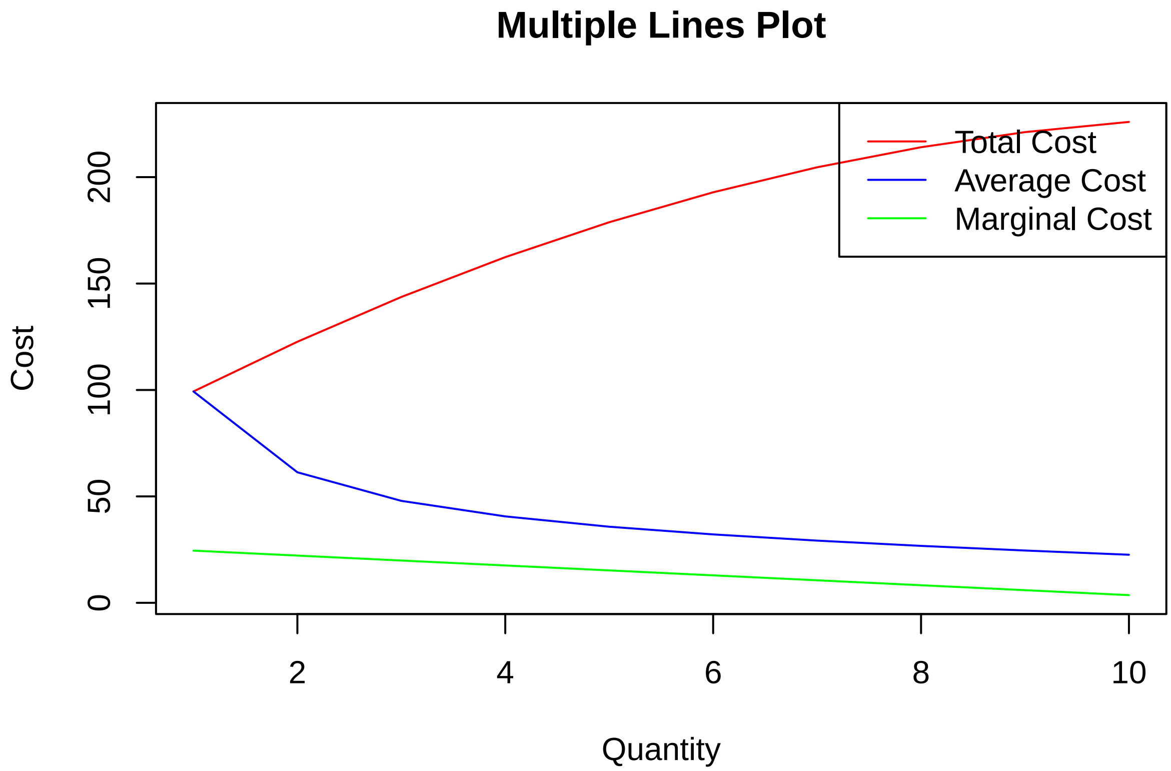
There is no difference between regression 4 and regression 5. We cannot reject the null hypothesis when the p-value is 0.861.

6. Using your preferred regression line, predict total cost, average cost, and marginal cost for the following values of quantity [1,2,3,4,5,6,7,8,9,10].

```r
# We can use the predict() function to get our estimated values of Cost
predicted_cost <- data.frame(Quantity = c(1:10))
predicted_cost$Cost <- predict(reg4, newdata = predicted_cost)
predicted_cost$AC <- predicted_cost$Cost/predicted_cost$Quantity
# We can take derivatives with the margins function
w <- margins(reg4, at = list(Quantity = 1:10))
predicted_cost$MC <- aggregate(w$dydx_Quantity,by=list(w$`_at_number`),FUN=mean)$x
```

7. Graph all three lines on the same graph.

```r
# Plot multiple lines using matplot
matplot(predicted_cost$Quantity, cbind(predicted_cost$Cost, predicted_cost$AC, predicted_cost$MC), type
        col = c("red", "blue", "green"), xlab = "Quantity",
        ylab = "Cost", main = "Multiple Lines Plot")
legend("topright", legend = c("Total Cost", "Average Cost", "Marginal Cost"),
       col = c("red", "blue", "green"),
       lty = 1)
```


Multiple Lines Plot
Total Cost
Average Cost
Marginal Cost
Cost
0
50
100
150
200
2
4
6
8
10
Quantity