A Hedonic Regression on Wines

730427792

ECON 490: Wine Economics

May 30, 2023

Introduction

We are interested in what factors contribute to the price of wine and interested in following up on a 2020 study that also performed a hedonic analysis on wine.[1] So, we collected data on 219 wines from 13 different stores near Chapel Hill, NC, and the following is the statistical summary of the data collected.

	Vintage	Red	White	Rosé	Orange	Sparkling	Winery?	Vineyard?	Oaked?	AVA	Price/750mL	Rating	ABV %
count	219.00	219.00	219.00	219.00	219.00	219.00	219.00	219.00	219.00	219.00	219.00	219.00	219.00
mean	2020.01	0.56	0.40	0.06	0.00	0.04	0.90	0.51	0.68	0.81	27.88	87.92	13.62
std	1.44	0.50	0.49	0.24	0.07	0.19	0.30	0.50	0.47	0.39	29.89	3.06	1.29
min	2014.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.99	72.00	6.80
25%	2019.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	12.99	86.00	13.00
50%	2021.00	1.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	18.99	88.00	13.80
75%	2021.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	33.50	90.00	14.50
max	2022.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	260.99	96.00	16.00

Noteworthy, across all wines, the average vintage is 2020, the oldest is 2014, and the newest is 2022. 56% of wines are red, 40% are white, and 6% are rosé. Additionally, 4% of the wines are sparkling. Wineries produced 90% of the wine, and 51% was produced by vineyards. 68% of the wine was aged in an oak barrel, and 81% of the sample was produced in an AVA.

Across all wines, the average gross price is $27.88, the cheapest wine is $3.99, and the most expensive is $260.99; the mean price is greater than the median price ($18.99), which signifies a skewed right distribution.

Using wine-searcher.com, we collected the rating (out of 100) for each wine. Across wines in the sample, the average rating is 87.92, the lowest is 72, and the highest is 96. Lastly, the average alcohol by volume (ABV) is 13.62%, the lowest is 6.8%, and the highest is 16%.

[1] Keating, "A Hedonic Study of Napa Valley"

Data Cleaning

Similar to the wine data from the 2020 study, our price data is skewed.[2] In order to overcome this problem, Ln(price) will serve as the dependent variable instead of price; this is the same technique used in the 2020 study.[3] In addition, since there was only one orange wine in the data, I removed it since one bottle is not an adequate sample to represent orange wine. Lastly, I created a new category of data named "OTHER" for varieties and states that have less than 3 entries. This is because I didn't want to remove valuable data, but I also didn't want to add variables to the model that have less than 3 samples to represent them. Below are the statistics for the cleaned data.

[2] Keating, "A Hedonic Study of Napa Valley," 320
[3] Keating, "A Hedonic Study of Napa Valley," 320

	count	mean	std	min	25%	50%	75%	max
Red	219.00	0.56	0.50	0.00	0.00	1.00	1.00	1.00
White	219.00	0.40	0.49	0.00	0.00	0.00	1.00	1.00
Rosé	219.00	0.06	0.24	0.00	0.00	0.00	0.00	1.00
Sparkling	219.00	0.04	0.19	0.00	0.00	0.00	0.00	1.00
Winery?	219.00	0.90	0.30	0.00	1.00	1.00	1.00	1.00
Vineyard?	219.00	0.51	0.50	0.00	0.00	1.00	1.00	1.00
Oaked?	219.00	0.68	0.47	0.00	0.00	1.00	1.00	1.00
AVA	219.00	0.81	0.39	0.00	1.00	1.00	1.00	1.00
Rating	219.00	87.92	3.06	72.00	86.00	88.00	90.00	96.00
ABV %	219.00	13.62	1.29	6.80	13.00	13.80	14.50	16.00
Log_Price	219.00	3.04	0.69	1.38	2.56	2.94	3.51	5.56
Age	219.00	2.99	1.44	1.00	2.00	2.00	4.00	9.00
Variety/Vine_BLEND	219.00	0.17	0.38	0.00	0.00	0.00	0.00	1.00
Variety/Vine_CABERNET SAUVIGNON	219.00	0.18	0.38	0.00	0.00	0.00	0.00	1.00
Variety/Vine_CHARDONNAY	219.00	0.17	0.38	0.00	0.00	0.00	0.00	1.00
Variety/Vine_MERLOT	219.00	0.05	0.22	0.00	0.00	0.00	0.00	1.00
Variety/Vine_OTHER	219.00	0.05	0.23	0.00	0.00	0.00	0.00	1.00
Variety/Vine_PETITE SIRAH	219.00	0.01	0.12	0.00	0.00	0.00	0.00	1.00
Variety/Vine_PINOT GRIGIO	219.00	0.03	0.18	0.00	0.00	0.00	0.00	1.00
Variety/Vine_PINOT GRIS	219.00	0.04	0.19	0.00	0.00	0.00	0.00	1.00
Variety/Vine_PINOT NOIR	219.00	0.16	0.37	0.00	0.00	0.00	0.00	1.00
Variety/Vine_RIESLING	219.00	0.03	0.16	0.00	0.00	0.00	0.00	1.00
Variety/Vine_ROSÉ	219.00	0.02	0.15	0.00	0.00	0.00	0.00	1.00
Variety/Vine_SAUVIGNON BLANC	219.00	0.05	0.23	0.00	0.00	0.00	0.00	1.00
Variety/Vine_ZINFANDEL	219.00	0.03	0.16	0.00	0.00	0.00	0.00	1.00
State_CALIFORNIA	219.00	0.72	0.45	0.00	0.00	1.00	1.00	1.00
State_NEW YORK	219.00	0.02	0.13	0.00	0.00	0.00	0.00	1.00
State_NORTH CAROLINA	219.00	0.05	0.22	0.00	0.00	0.00	0.00	1.00
State_OREGON	219.00	0.12	0.32	0.00	0.00	0.00	0.00	1.00
State_OTHER	219.00	0.02	0.15	0.00	0.00	0.00	0.00	1.00
State_WASHINGTON	219.00	0.07	0.25	0.00	0.00	0.00	0.00	1.00

Break Down for Variety/Vine_OTHER

	Variety/Vine	Count
0	MOSCATO	2
1	SYRAH	2
2	GRENACHE	2
3	GREEN APPLE WINE	1
4	NIAGARA	1
5	MUSCADINE	1
6	VIOGNEIR	1
7	SCUPPERNONG	1
8	CARIGNAN	1

Break Down for State_OTHER

	State	Count
0	WASHINGOTN	2
1	INDIANA	1
2	NEVADA	1
3	WEST VIRGINIA	1

Regression

OLS Regression Results

Dep. Variable:	Log_Price	**R-squared:**	0.585
Model:	OLS	**Adj. R-squared:**	0.526
Method:	Least Squares	**F-statistic:**	9.974
Date:	Mon, 29 May 2023	**Prob (F-statistic):**	3.42e-24
Time:	19:54:12	**Log-Likelihood:**	-133.18
No. Observations:	219	**AIC:**	322.4
Df Residuals:	191	**BIC:**	417.2
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.6522	1.089	-7.946	0.000	-10.800	-6.505
Red	-0.2901	0.201	-1.443	0.151	-0.687	0.106
Rosé	0.1372	0.242	0.567	0.572	-0.340	0.615
Sparkling	0.3088	0.205	1.504	0.134	-0.096	0.714
Winery?	-0.0777	0.116	-0.668	0.505	-0.307	0.152
Vineyard?	0.1196	0.073	1.644	0.102	-0.024	0.263
Oaked?	0.1411	0.081	1.741	0.083	-0.019	0.301
AVA	0.0992	0.099	1.004	0.317	-0.096	0.294
Rating	0.0968	0.012	8.041	0.000	0.073	0.121
ABV %	0.1745	0.038	4.558	0.000	0.099	0.250
Age	0.0791	0.026	3.059	0.003	0.028	0.130
Variety/Vine_BLEND	0.3388	0.240	1.410	0.160	-0.135	0.813
Variety/Vine_CABERNET SAUVIGNON	0.4815	0.252	1.913	0.057	-0.015	0.978
Variety/Vine_CHARDONNAY	-0.1244	0.165	-0.754	0.452	-0.450	0.201
Variety/Vine_MERLOT	0.3931	0.280	1.405	0.162	-0.159	0.945
Variety/Vine_OTHER	0.1987	0.227	0.874	0.383	-0.250	0.647
Variety/Vine_PETITE SIRAH	0.6206	0.367	1.691	0.092	-0.103	1.344
Variety/Vine_PINOT GRIGIO	-0.0328	0.230	-0.143	0.887	-0.486	0.421
Variety/Vine_PINOT GRIS	-0.2604	0.231	-1.127	0.261	-0.716	0.195
Variety/Vine_PINOT NOIR	0.3100	0.259	1.196	0.233	-0.201	0.821
Variety/Vine_RIESLING	0.1266	0.253	0.501	0.617	-0.372	0.625
Variety/Vine_ROSÉ	0.3692	0.359	1.028	0.305	-0.339	1.077
Variety/Vine_ZINFANDEL	0.2171	0.311	0.697	0.486	-0.397	0.831
State_CALIFORNIA	0.3177	0.134	2.364	0.019	0.053	0.583
State_NEW YORK	0.4928	0.284	1.737	0.084	-0.067	1.052
State_NORTH CAROLINA	0.5009	0.213	2.355	0.020	0.081	0.920
State_OREGON	0.5079	0.174	2.918	0.004	0.165	0.851
State_OTHER	0.1352	0.266	0.508	0.612	-0.390	0.660

Omnibus:	6.751	**Durbin-Watson:**	1.892
Prob(Omnibus):	0.034	**Jarque-Bera (JB):**	7.993
Skew:	0.253	**Prob(JB):**	0.0184
Kurtosis:	3.787	**Cond. No.**	3.07e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.07e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Interpretation

I am going to address the coefficients that are statistically significant (>90% confidence), because in cases where coefficients do not meet this criterion, additional data is required to have a satisfactory level of certainty about their impact on the price of wine. Furthermore, regarding the second note generated by the model, there is multicollinearity between the data collected. This means that variables in the regression may have hidden dependence on other variables even though a linear regression is supposed to be in terms of independent variables.

The first statistically significant coefficient is for "Oaked?" That is, if a wine is oaked, one should expect a 14.11% increase in price. This makes sense because oaked barrels are typically more expensive than the alternatives, which should reflect in the price.

The coefficient for rating is significant, and it means that for every one point increase in rating, the price of wine increases by 9.68%; however, the 2020 study found the increase to be 7.2%, which is below the lower bound (by ~1%) of the regression above.[4] ABV is also significant, and for every one percent increase in alcohol, there is an expected 17.45% increase in price. Age is significant, and for every one year a wine ages, the price is expected to increase by 7.91%; while the 2020 study found the increase to be 5.6%, which is still within the lower bounds of the regression above.[5] The reason for collinearity is most likely due to the inclusion of rating, ABV, and age. This is because when wine ages, yeast has more time to break down sugar and produce alcohol, which increases ABV. Moreover, if sommeliers tend to give aged bottles a higher rating, then there are 3 variables that are interdependent even though these variables are assumed to be independent in a linear regression. This is most likely the reason why there is a

[4] Keating, "A Hedonic Study of Napa Valley," 324
[5] Keating, "A Hedonic Study of Napa Valley," 324

collinear warning. For this reason, a follow-up study should choose either rating, ABV, or age and see how the results compare.

Using sauvignon blanc as the base wine, if the wine is cabernet sauvignon, there is an expected price increase of 48.15%, and in the 2020 study, the price increase is 77.7%, which is within the upper bounds of the regression above.[6] For petite syrah, there is an expected price increase of 62%, and in the 2020 study, there is a price increase of 32%, which is within the lower bounds of the regression above.[7] Out of the 4 remaining wines that were looked at in both my regression and in the 2020 study, 3 shared similar coefficients except mine are not statistically significant.[8] Chardonnay is the anomaly where I found there to be an expected price decrease and the 2020 study found an expected price increase.[9]

Using Washington state as a base, the price of wines produced in California, New York, North Carolina, and Oregon are expected to be 32% more expensive on the low end and 51% on the high end. Granted, all other states (besides California with 72% and Oregon with 12%) accounted for 5% of the data or less, which means the results should be taken with a grain of salt since the data may not fairly represent all wines from Washington, New York, and North Carolina.

Lastly, the data shows that if wine comes from an AVA, the price is expected to increase by 9.9% and in the 2020 study, AVA status accounts for a 12.2% increase in wine price, which is within the bounds.[10] However, the short fall is that the coefficient for AVA is not statically

---

[6] Keating, "A Hedonic Study of Napa Valley," 324
[7] Keating, "A Hedonic Study of Napa Valley," 324
[8] Keating, "A Hedonic Study of Napa Valley," 324
[9] Keating, "A Hedonic Study of Napa Valley," 324
[10] Keating, "A Hedonic Study of Napa Valley," 324

significant in the regression above, so more data is needed to confirm that AVA increases price. Nevertheless, with the data that is present, preliminary results are similar to the 2020 study.

Conclusion

Overall, the 2020 study and this study have similar conclusions. In both, age, rating, and having AVA status positively affect the price of wine.[11] Both studies also agree that compared to sauvignon blanc, all other varietals used in both studies (except chardonnay), positively impact the price of wine.[12] I also made findings that are not in the 2020 study. That is, aging in oak increases the expected price of wine and so does having a higher ABV (though there could be collinearity with age).

Example Prediction

Example Prediction using

Producer	Variety/Vine	Vintage	Red	White	Rosé	Orange	Sparkling	State	Winery?	Vineyard?	Oak aging (in months)	AVA	Price/750mL	Rating	ABV %
Turnbull	Cabernet Sauvignon	2020	1	0	0	0	0	California	1	1	YES	1	60.99	91	14.7

```
# predict using the model
log_actual_price = np.log(60.99)
data = np.array([1, # constant
                 1,0,0, # red
                 1, 1, # winery, vineyard
                 1, # oaked
                 1, # ava status
                 91, # Rating
                 14.7, #ABV
                 3, # Age
                 0, 1, 0,0,0,0,0,0,0,0,0,0, # Variety/Vine
                 1,0,0,0,0]) # State

print(f"Estimated log(price)= {data.dot(model.params.values)}")
print("Actual log(price)= " + str(log_actual_price))
```

✓ 0.0s Python

```
Estimated log(price)= 3.7546425842764846
Actual log(price)= 4.110709916308365
```

[11] Keating, "A Hedonic Study of Napa Valley," 324
[12] Keating, "A Hedonic Study of Napa Valley," 324

Bibliography

Keating, Grant Bartlett. “An Empirical Analysis of the Effect of Sub-Divisions of American Viticultural Areas on Wine Prices: A Hedonic Study of Napa Valley.” *Journal of Wine Economics* 15, no. 3 (2020): 312–29. doi:10.1017/jwe.2020.29.