

APS 425 – Winter 2009

Multiple Regression Review

Instructor: G. William Schwert

275-2470

schwert@schwert.simon.rochester.edu

Multiple Regression Model

- We have studied the multiple regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_n X_{ni} + e_i$$

- We don't know the true values for $\beta_0, \beta_1, \dots, \beta_n$

Multiple Regression Model

- Given a sample, we find, as before, estimators b_0, b_1, \dots, b_n , by minimizing the sum of squared prediction errors:

$$\sum \hat{e}_i^2 = \sum (Y_i - \hat{Y}_i)^2, \text{ where } \hat{Y}_i = b_0 + b_1X_{1i} + \dots + b_nX_{ni}$$

- The estimators b_0, b_1, \dots, b_n are unbiased, consistent, and efficient estimators of the population parameters $\beta_0, \beta_1, \dots, \beta_n$ if the following six assumptions are satisfied

Multiple Regression Model

- Six Assumptions
 - $E(e_i) = 0$
 - The model is correctly specified, i.e.,

$$Y_i = \beta_0 + \beta_1X_{1i} + \dots + \beta_nX_{ni} + e_i$$
 - $\text{Corr}(X_{ki}, e_i) = 0$ for all i, k
 - e_i Has a normal distribution
 - $\text{Var}(e_i) = \sigma = \text{a constant}$
 - $\text{Corr}(e_i, e_j) = 0$ for all i, j
- Hence, if these assumptions are satisfied, the estimators b_0, b_1, \dots, b_n provide accurate information about the values of the population parameters $\beta_0, \beta_1, \dots, \beta_n$

Example: Wine Prices & Weather

- The Excel spreadsheet A425_WINE.XLSX contains market prices for a collection of 25 high quality Bordeaux wines (not including Château Petrus or Château Mouton Rothschild, both of which have prices that are often out of line with their “quality”) from different vintages (years). All prices (PRICE) are expressed relative to the prices of the 1961 vintage, which is renowned for being the best during this period. So, for example, the portfolio of 25 1989 vintage Bordeaux wines costs 23% as much as the same wines from the 1961 vintage.

Example: Wine Prices & Weather

- The data were provided by Professor Orley Ashenfelter of Princeton University, publisher of *Liquid Assets*, a wine newsletter that provides current auction prices for wines and forecasts quality of new wine vintages [<http://www.liquidasset.com>]. There are no prices for wines after 1989 because these wines were not mature at the time these data were prepared. One of the goals of this exercise is to construct a method of forecasting the prices (or values) of these wines.

Example: Wine Prices & Weather

- The weather variables for the Bordeaux region of France are some of the main determinants of the quality of wine. Harvest rainfall (HARVRAIN in mm) is important because if it rains too much during the harvest season then the wines will be too watery or too diluted. The better vintages have dry harvest periods and are said to be more concentrated. Summer temperature (SUMTEMP in average degrees centigrade) is also important because the hotter weather is necessary for the grapes to fully ripen. Riper, sweeter fruit produces a better quality wine.

Example: Wine Prices & Weather

- Riper, sweeter fruit produces a better quality wine. Winter rainfall (WINTRAIN in mm) is important because wetter weather is good for the grape vines early in the growing season. The average temperature during the harvest season (SEPTEMP) is also included because some people suspect that wines that are “soft and easy drinking” are made when it was hot during the September when the grapes were being picked.

Example: Wine Prices & Weather

- Age is also an important determinant of the price of wine. The reason for this is largely because the quality of wines improves with age. A typical wine might take 10 years to mature and continues to improve in quality beyond that point. Of course, it is also true that the price must be increasing with age, otherwise consumers would not buy wines when they were young (they could put their money in the bank instead and buy the wines when they were older).
- A quick glance at the data reveals that 1961, 1953, and 1959 are among the hottest and driest years for Bordeaux wines, and also have the highest relative prices. Of course, these are also some of the older wines in our data.

Wine Prices & Weather: Questions

- Are the theoretical predictions about the effect of weather on wine quality supported by these data?
- If you think about wine as an investment, is there any evidence that it pays to buy wine when it is young and store it, or should you spend your money on wine after it has matured?
- Prof. Ashenfelter originally analyzed these data using the 1952-80 sample period and became so famous in wine circles that the *New York Times* wrote an extensive story about his equation in their weekend edition (see abstract below). Is there any evidence that the model for wine prices changes when you include the additional data from 1981-89?

Wine Prices & Weather: Questions

- Often wine connoisseurs do tastings of Bordeaux wines when they are still developing in large oak barrels and try to forecast what the wine will be like when it is drinkable. For example, Robert Parker has become famous because people have come to trust his skill at evaluating wines in this way.

I have included Parker's ratings of the major Bordeaux regions for each year from 1975-98 from his web page [<http://www.winetech.com/html/vintchrt.html>] and then averaged them to create a vintage quality measure called "PARKER" in the spreadsheet. Do Parker's quality rankings help explain prices?

- How would you create an index of quality for different vintages using only weather information? How does it compare with Parker's ratings?
- How would you forecast prices from 1990-98?

Results, 1952-89

- Start with simple regression that tries to explain price as a function of rain during the harvest (HARVRAIN) and during the prior winter (WINTRAIN), and temperature during the growing season (SUMTEMP) and during the harvest season (SEPTEMP)

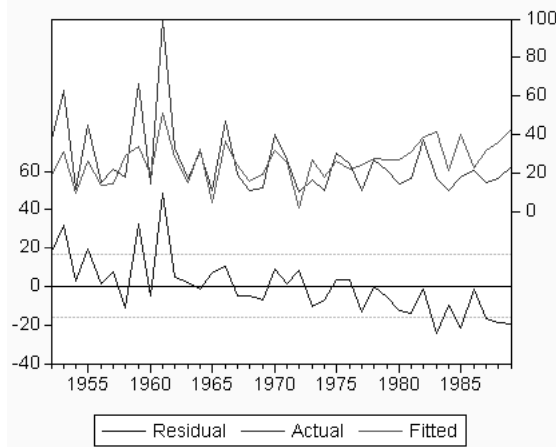
Dependent Variable: PRICE
Method: Least Squares
Sample: 1952 1989
Included observations: 38

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-173.8848	68.98331	-2.520679	0.0167
WINTRAIN	0.026542	0.020784	1.277028	0.2105
HARVRAIN	-0.028712	0.047755	-0.601224	0.5518
SUMTEMP	7.660968	4.183120	1.831400	0.0761
SEPTEMP	3.514580	2.625908	1.338425	0.1899

R-squared	0.316510	Mean dependent var	25.80876
Adjusted R-squared	0.233663	S.D. dependent var	18.57255
S.E. of regression	16.25854	Akaike info criterion	8.537193
Sum squared resid	8723.220	Schwarz criterion	8.752664
Log likelihood	-157.2067	F-statistic	3.820407
Durbin-Watson stat	1.556011	Prob(F-statistic)	0.011669

Results, 1952-89

- It looks like the residuals (blue line on the bottom) have higher mean and variance in the early years
 - They seem to be trending down and their amplitude is larger in the early data
- => Try adding the time variable to reflect that fact that older wines cost more (otherwise, why would anyone store them for drinking later?)



Results, 1952-89

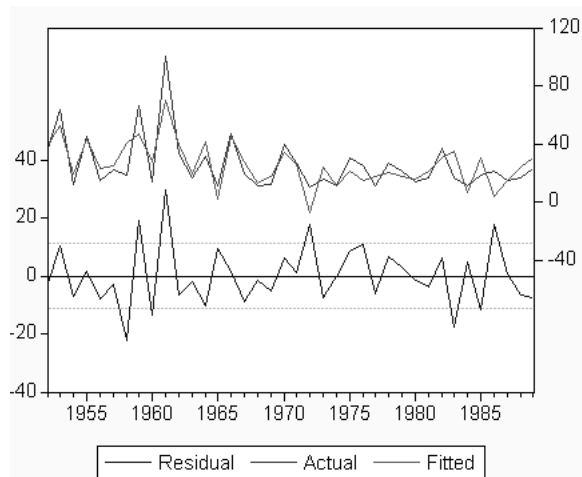
- It looks like adding TIME to reflect to different age of the vintages was important (t-stat of -5.91)
- R^2 increases from 31.7% to 67.3%
- The weather variables seem to make sense: higher temperatures are associated with better (higher priced) wine; rain before the growing season is good, but during harvest is bad

Dependent Variable: PRICE
 Method: Least Squares
 Sample: 1952 1989
 Included observations: 38

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-226.7766	49.24225	-4.605325	0.0001
WINTRAIN	0.039187	0.014745	2.657598	0.0122
HARVRAIN	-0.055092	0.033818	-1.629104	0.1131
SUMTEMP	11.62203	3.011789	3.858846	0.0005
SEPTEMP	3.687754	1.843502	2.000406	0.0540
TIME	-1.069174	0.180795	-5.913737	0.0000
R-squared	0.673422	Mean dependent var	25.80876	
Adjusted R-squared	0.622394	S.D. dependent var	18.57255	
S.E. of regression	11.41277	Akaike info criterion	7.851281	
Sum squared resid	4168.039	Schwarz criterion	8.109847	
Log likelihood	-143.1743	F-statistic	13.19716	
Durbin-Watson stat	2.905511	Prob(F-statistic)	0.000001	

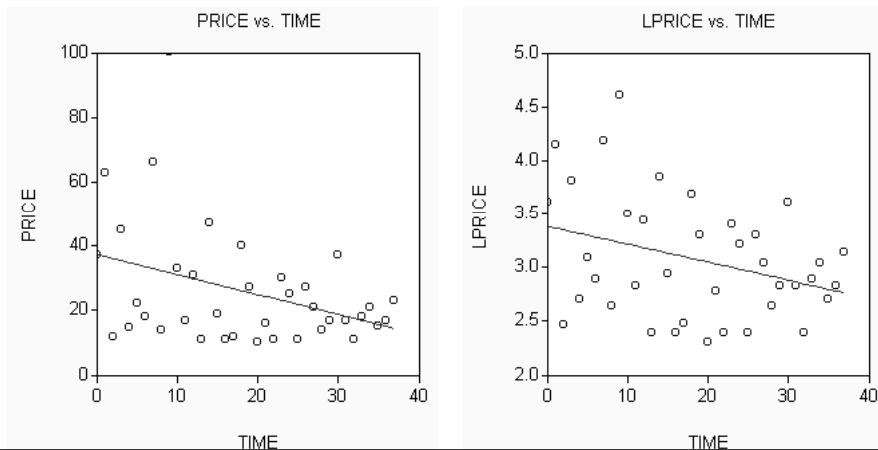
Results, 1952-89

- We have fixed the trend, but it still looks like the residuals (blue line on the bottom) have higher variance in the early years
- => Try log transformation for price



Scatterplots of Price or Log(price) vs Time

- The log(price) plot looks like it will have less heteroskedasticity



Log(Price) Results, 1952-89

- R^2 in the log model is a little higher than in the “raw” model (71.2% vs. 67.3%)

=> Try log transformation for price

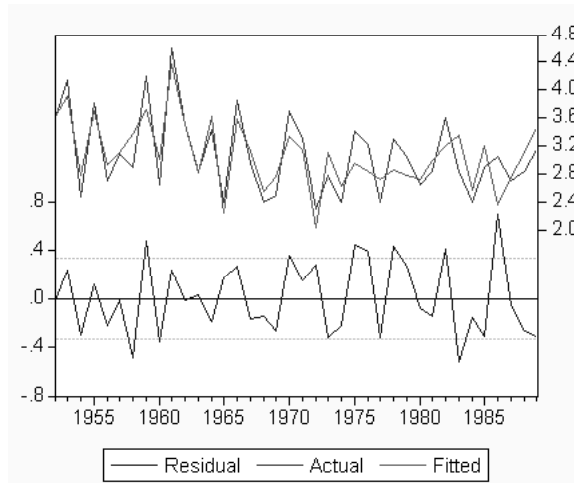
Dependent Variable: LPRICE
 Method: Least Squares
 Sample: 1952 1989
 Included observations: 38

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5.139407	1.418973	-3.621920	0.0010
WINTRAIN	0.000932	0.000425	2.194256	0.0366
HARVRAIN	-0.002469	0.000974	-2.533755	0.0164
SUMTEMP	0.445627	0.086788	5.134653	0.0000
SEPTEMP	0.068776	0.053123	1.294664	0.2047
TIME	-0.031682	0.005210	-6.081244	0.0000

R-squared	0.712009	Mean dependent var	3.071806
Adjusted R-squared	0.667010	S.D. dependent var	0.569917
S.E. of regression	0.328872	Akaike info criterion	0.757645
Sum squared resid	3.461023	Schwarz criterion	1.016211
Log likelihood	-8.395247	F-statistic	15.82292
Durbin-Watson stat	2.543196	Prob(F-statistic)	0.000000

Log(Price) Results, 1952-89

- These plots look much better: amplitude of the residuals is similar throughout 1952-89
- This is because using log(price) is essentially like looking at percentage changes, rather than absolute changes, in wine prices
- % changes are more likely to have the same distribution across long time periods



Log(Price) Results, 1952-89

- Even though there is no particular reason to suspect that there is a heteroskedasticity problem, as a check I also use “heteroskedasticity consistent standard errors”
- As expected, the t-stats do not really change very much
- Note: the only things that change are the std errors of the coefficients and t-stats

Dependent Variable: LPRICE

Method: Least Squares

Sample: 1952 1989

Included observations: 38

White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5.139407	1.521039	-3.378878	0.0019
WINTRAIN	0.000932	0.000431	2.164137	0.0380
HARVRAIN	-0.002469	0.000983	-2.511663	0.0173
SUMTEMP	0.445627	0.087140	5.113952	0.0000
SEPTEMP	0.068776	0.050180	1.370596	0.1800
TIME	-0.031682	0.005343	-5.929260	0.0000
R-squared	0.712009	Mean dependent var	3.071806	
Adjusted R-squared	0.667010	S.D. dependent var	0.569917	
S.E. of regression	0.328872	Akaike info criterion	0.757645	
Sum squared resid	3.461023	Schwarz criterion	1.016211	
Log likelihood	-8.395247	F-statistic	15.82292	
Durbin-Watson stat	2.543196	Prob(F-statistic)	0.000000	

(1) Are the theoretical predictions about the effect of weather on wine quality supported by these data?

- Yes, common sense seems to accord with the apparent effects of weather conditions on average wine prices:
 - higher temperatures are associated with better (higher priced) wine
 - Although the effect during the harvest season is weak (t-stat = 1.37)
 - rain before the growing season is good, but during harvest is bad

(2) If you think about wine as an investment, is there any evidence that it pays to buy wine when it is young and store it, or should you spend your money on wine after it has matured?

- The coefficient of time is the effect of one more year of aging on the $\log(\text{price})$, $\partial \log(\text{price}) / \partial t$, which is like the (continuously compounded) interest rate
- The regression implies that, HAOVC, wine prices increase 3.17% for each additional year of aging

(2) If you think about wine as an investment, is there any evidence that it pays to buy wine when it is young and store it, or should you spend your money on wine after it has matured?

- Whether it makes sense to buy wine when it is young and store it for drinking later depends on whether the “real return” on your alternative investments, adjusted for the risk, is more or less than 3.17%
 - These are “real returns” since the prices are all measured relative to the current price of the 1961 vintage in 1989
 - Overall inflation from year to year should presumably affect the prices of all vintages proportionally
 - You will learn more about risk adjustment and reasonable expected real and nominal rates of return in your Capital Markets course, FIN 411

(3) Is there any evidence that the model for wine prices changes when you include the additional data from 1981-89?
Log(Price) Results, 1952-80

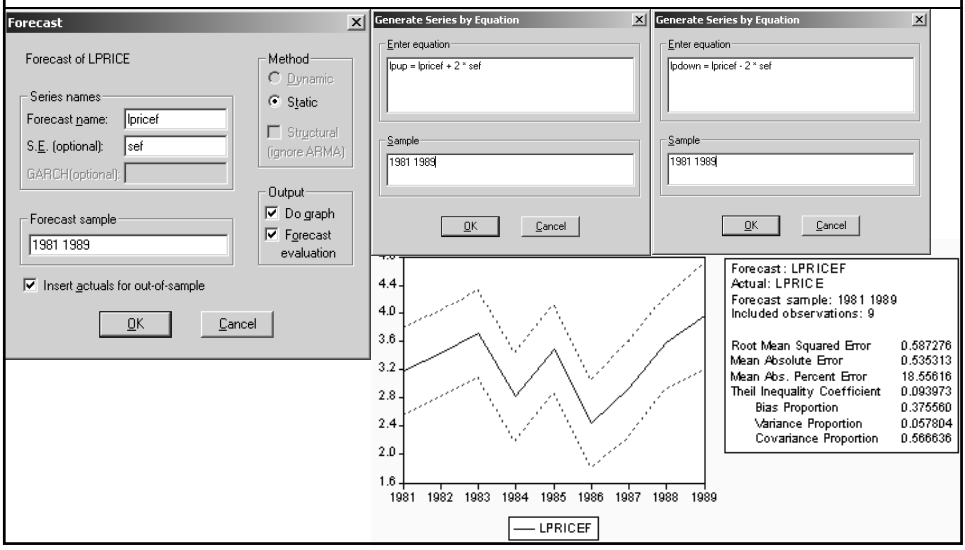
- Results for 1952-80 look even better
- R² is over 84% (adj R² is over 80%) vs. 71.2% (66.7%) for longer sample
- Effect of harvest temperature seems unimportant (t-stat = 0.19)
- Other weather effects are stronger
- Estimate of real interest rate is 2.45%

Dependent Variable: LPRICE
 Method: Least Squares
 Sample: 1952 1980
 Included observations: 29
 White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-6.882528	1.435070	-4.795954	0.0001
WINTRAIN	0.001188	0.000433	2.742142	0.0116
HARVRAIN	-0.003668	0.000853	-4.301199	0.0003
SUMTEMP	0.610360	0.076174	8.012709	0.0000
SEPTEMP	0.009277	0.047614	0.194834	0.8472
TIME	-0.024451	0.005724	-4.271351	0.0003

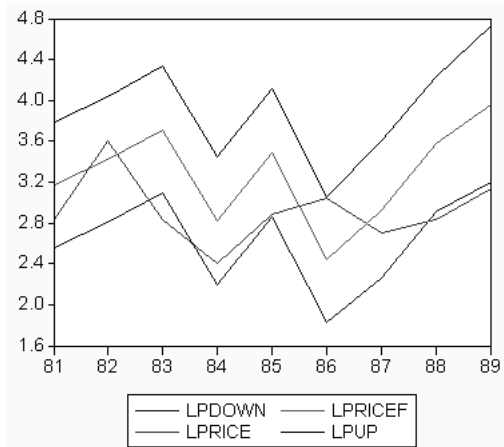
R-squared	0.840989	Mean dependent var	3.118681
Adjusted R-squared	0.806422	S.D. dependent var	0.623025
S.E. of regression	0.274115	Akaike info criterion	0.431456
Sum squared resid	1.728203	Schwarz criterion	0.714345
Log likelihood	-0.256118	F-statistic	24.32890
Durbin-Watson stat	2.848488	Prob(F-statistic)	0.000000

**Forecast Log(price) from 1981-89
 Using Model from 1952-80**



Forecast Log(price) from 1981-89 Using Model from 1952-80

- It looks like the actual price falls close to or outside the 95% prediction interval several times in this 9 year period
- 1983, 1985, 1988 and 1989 are low
- Thus, relative to the 1952-80 experience, these seem to be years when the market is undervaluing the wines of these vintages
- 1986 is high (over-valued?)



(4) Can Robert Parker Improve on Weather Forecasts?

- Since Parker's ratings are only available for the 1975-89 period when we also have price data, this sample size is much smaller
- Nevertheless, HAOVC, the Parker coefficient is about .06 (implying a price that is 6% higher for each Parker rating point)
- Parker's ratings seem to subsume the information in the weather
 - Which is not surprising since Parker should know what the weather was like, as well as frequently taste these wines

Dependent Variable: LOG(PRICE)
Method: Least Squares
Sample (adjusted): 1970 1989
Included observations: 15 after adjustments
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.433955	2.083727	-0.208259	0.8402
WINTRAIN	-0.000674	0.000560	-1.204022	0.2630
HARVRAIN	-0.000229	0.001187	-0.192630	0.8520
SUMTEMP	0.024611	0.081148	0.303281	0.7694
SEPTEMP	-0.010853	0.070562	-0.153814	0.8816
TIME	-0.043671	0.007409	-5.894252	0.0004
PARKER	0.058095	0.012246	4.744072	0.0015

R-squared	0.856712	Mean dependent var	3.098214
Adjusted R-squared	0.749246	S.D. dependent var	0.318512
S.E. of regression	0.159496	Akaike info criterion	-0.528869
Sum squared resid	0.203512	Schwarz criterion	-0.198446
Log likelihood	10.96652	F-statistic	7.971934
Durbin-Watson stat	2.839029	Prob(F-statistic)	0.004960

(5) How would you create an index of quality for different vintages using only weather information?

- Presumably “quality” abstracts from age, so a forecast of the price of the wine setting time equal to its average value in the sample (or any other constant number) would give a market-based measure of quality

Quality Results, 1952-98

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5.139407	1.521039	-3.378878	0.0019
WINTRAIN	0.000932	0.000431	2.164137	0.0380
HARVRAIN	-0.002469	0.000983	-2.511663	0.0173
SUMTEMP	0.445627	0.087140	5.113952	0.0000
SEPTEMP	0.068776	0.050180	1.370596	0.1800
TIME	-0.031682	0.005343	-5.929260	0.0000

R-squared	0.712009	Mean dependent var	3.071806
Adjusted R-squared	0.667010	S. D. dependent var	0.569917
S. E. of regression	0.328872	Akaike info criterion	0.757645
Sum squared resid	3.461023	Schwarz criterion	1.016211
Log likelihood	-8.395247	F-statistic	15.82292
Durbin-Watson stat	2.543196	Prob(F-statistic)	0.000000

Forecast

Forecast of LPRICE

Series names:

Forecast name:

S.E. (optional):

GARCH (optional):

Method:

Dynamic

Static

Structural (ignore ARMA)

Forecast sample:

Insert actuals for out-of-sample

Output:

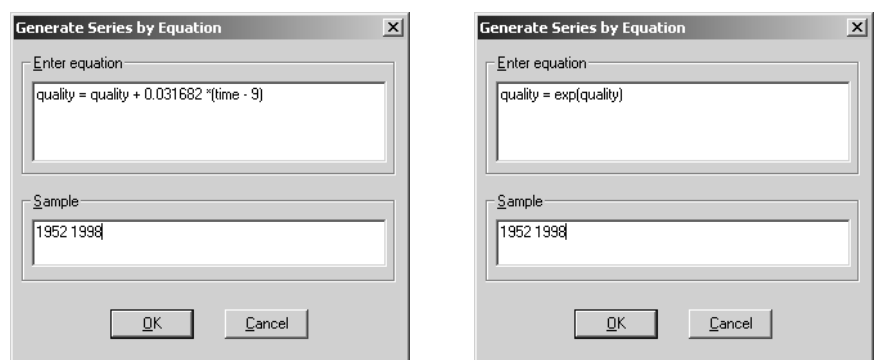
Do graph

Forecast evaluation

• Create forecasts of log(price) including TIME, call it “QUALITY”

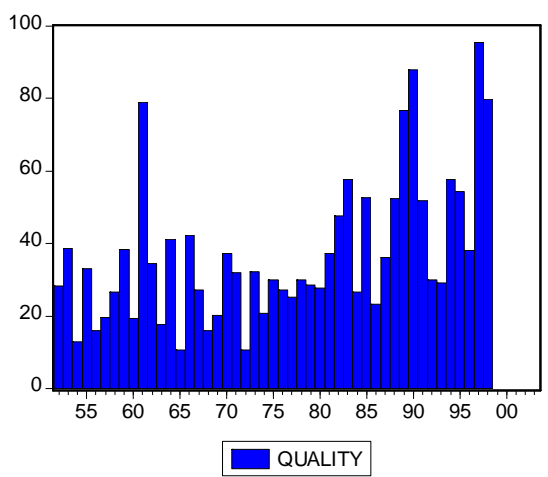
Quality Results, 1952-98

- Since TIME = 9 in 1961, we adjust the forecast of log(price) that we have been calling "QUALITY" by adding $0.031682 * (TIME - 9)$ so 1961 is our base year
- We then transform "QUALITY" back to units of price by using the exponential transformation



Quality Results, 1952-98

- For the period 1952-80, 1961 stands out as the highest quality year (78.9)
- Since 1980, however, there have been many extremely good years:
 - 1983 = 57.7
 - 1989 = 76.7
 - 1990 = 87.8
 - 1994 = 57.7
 - 1995 = 54.4
 - 1997 = 95.3
 - 1998 = 79.9



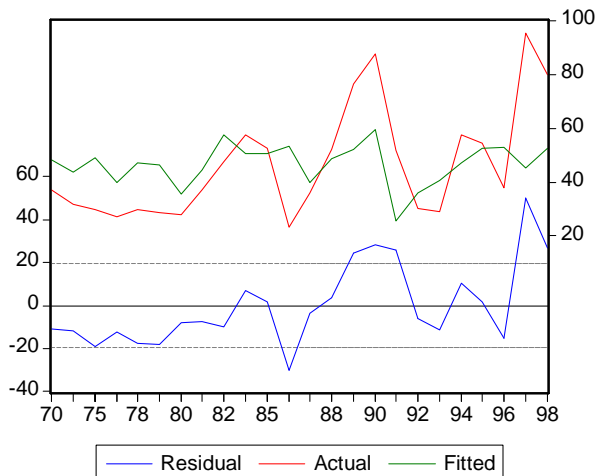
Quality Results, 1952-98

Equation Specification Equation specification: Dependent variable followed by list of regressors including AR and PDL terms. OR an explicit equation like $Y=c(1)+c(2)X$. quality c parker		Dependent Variable: QUALITY Method: Least Squares Sample (adjusted): 1970 1998 Included observations: 24 after adjustments White Heteroskedasticity-Consistent Standard Errors & Covariance				
Estimation settings Method: LS - Least Squares (NLS and ARMA) Sample: 1952 1998		Variable	Coefficient	Std. Error	t-Statistic	Prob.
		C	-67.46418	66.71339	-1.011254	0.3229
		PARKER	1.330433	0.780902	1.703714	0.1025
		R-squared	0.137607	Mean dependent var	46.67033	
		Adjusted R-squared	0.098407	S.D. dependent var	20.63884	
		S.E. of regression	19.59704	Akaike info criterion	8.868290	
		Sum squared resid	8448.971	Schwarz criterion	8.966461	
		Log likelihood	-104.4195	F-statistic	3.510412	
		Durbin-Watson stat	1.268058	Prob(F-statistic)	0.074330	

•Regression of quality on Parker shows they are positively correlated, but the relation is not that strong
 -R² is 9.8% and t-stat for Parker is 1.70

Quality Results, 1952-98

- According to this analysis, Parker over-rates 1986, and vastly under-rates 1989, 1990, 1997, and 1998
- If the wine market follows Parker, the implication of this analysis is that you should purchase the 89, 90, 97, and 98 vintages, but avoid 86



(6) How would you forecast prices from 1990-98?

- Eviews makes an unbiased prediction of price

Forecast

Forecast equation
EQ03H

Series to forecast
 PRICE LOG(PRICE)

Series names
Forecast name: pricet
S.E. (optional):
GARCH(optional):

Method
Static forecast (no dynamics in equation)
 Structural (ignore ARMA)
 Coef uncertainty in S.E. calc

Forecast sample
1952 2003

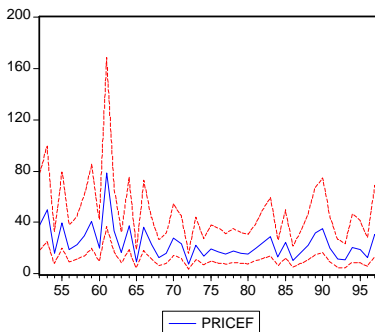
Output
 Forecast graph
 Forecast evaluation

Insert actuals for out-of-sample observations

OK Cancel

Forecast Prices, 1952-98

- You can see that forecasted prices for vintages after 1982 are small compared with the benchmark year of 1961
- This reflects the high quality of the 61 vintage and the time value of money
- The 89 and 90 vintages should be more expensive than the famous 82 vintage (which is not true)



Forecast: PRICEF
Actual: PRICE
Forecast sample: 1952 2003
Adjusted sample: 1952 1998
Included observations: 38

Root Mean Squared Error	8.493032
Mean Absolute Error	6.472997
Mean Abs. Percent Error	26.10137
Theil Inequality Coefficient	0.143227
Bias Proportion	0.031966
Variance Proportion	0.364700
Covariance Proportion	0.603334

Conclusions

- Simple regression methods seem to give very useful forecasts of wine quality based on publicly available data
- The implied real rate of interest from buying and storing wine is around 2.5% to 3%
 - Buy & store if this is an adequate return for you, otherwise, invest your money and buy these wines at auction after they have matured

Conclusions

- Since weather is known long before vintages are available for tasting, you could use these regression methods to tell you whether to buy a particular vintage's "futures" contracts (e.g., through Century Liquor)
- Parker's quality ratings do not correlate strongly with weather factors
 - If current retail prices of wines are strongly influenced by Parker's ratings, buy the vintages that he under-rates and avoid the ones he over-rates

Conclusions

Interesting Questions:

- (1) Do you think regressions like these would work as well for the prices of one particular Chateau (as opposed to the average prices across 25 Chateaux)?
Why, or why not?

- (2) Do you think regressions like these would work as well for the prices of a group of 25 high quality California Cabernets? *Why, or why not?*

Links

Excel spreadsheet

http://schwert.simon.rochester.edu/a425/a425_wine.xlsx

Eviews worksheet

http://schwert.simon.rochester.edu/a425/a425_wine.wf1

APS 425 Home Page

<http://schwert.simon.rochester.edu/a425/a425main.htm>