Empirical Economics with R

1 Predicting Prices for Bordeaux Red Wines

Uni Ulm

Prof. Dr. Sebastian Kranz

WiSe 20/21

Red Wine from Bordeaux

- Red wines from certain vineyards in Bordeaux are world famous
- A common recommendation by wine experts is to store a bottle of Bordeaux for 8 to 30 years to maximize the pleasure of drinking it.
- There are regular auctions for Bordeaux wines and some people buy younger wines as a financial investment.

Auction prices by vintage

- Auction prices differ substantially between vintages (year of production)
- Here is a table of average auction prices in USD from 1990 and 1991 for 12 Bottles of Bordeaux wines from different vineyards and vintages (year of production)

Vintage	Lafte	Latour	Cheval Blanc	Cos d'Estournel	Montrose	Pichon Lalande	Average
1960	494	464	486				479
1961	4335	5432	3534	1170	1125	1579	4884
1962	889	1064	821	521	456	281	977
1963	340	471		251			406
1964	649	1114	1125	315	350	410	882
1965	190	424				258	307
1966	1274	1537	1260	546	482	734	1406
1967	374	530	441	213	236	243	452
1968	223	365	274				294
1969	251	319		123	84	152	285

Source: Ashenfelter (2008)

Expert Recommendations

- There are many recommendations by wine experts about the taste of different vintages.
- Here are recommendation examples for the vintages 1960-1963 from https://www.thewinecellarinsider.com:

> 1960 Bordeaux Wine – Avoid these wines.
>
> 1961 Bordeaux Wine – A legendary Bordeaux vintage. Big, concentrated and tannic in their youth, numerous great wines were produced in the Medoc, the Right Bank and Pessac Leognan. If well stored, many of these wines are still offering great pleasure today. Latour, Haut Brion and Mouton Rothschild can still improve! Drink or hold.
>
> 1962 Bordeaux Wine– I have not tasted many 1962 wines. They exist in the shadows of the legendary 1961 vintage. Considered better than average from experienced tasters, it’s worth trying a bottle if you see one with decent provenance.
>
> 1963 Bordeaux Wine – Avoid these wines.

Ashenfelter's Wine Formula

- Economist and wine enthusiast Orley Ashenfelter wanted a mathematical method to recommend which vintages of young Bordeaux wines to buy and store. He developed a formula to predict the average "quality" of a vintage:

```
  0.6160  * average temperature during growing season
+ 0.00117 * rainfall in preceeding winter months
- 0.00386 * rainfall in August (harvest month)
```

- Ashenfelter first published the formula in a newsletter and wrote later two more detailed articles Ashenfelter, Ashmore & Lalonde (1995) and Ashenfelter (2008).
    - Here we will mainly follow the analysis in Ashenfelter et al. (1995)
- Wine critics were cited as calling the method "ludicrous and absurd" while Ashenfelter is said to note that "many connoisseurs agree privately with him but refuse to say so publicly" (see here).
- There is a nice Youtube video from 1992 about the development and reaction to this wine formula.

Linear Regression

6 / 30

- Ashenfelter developed his formula by estimating the following *linear regression model*.

$$qual_t = \beta_0 + \beta_1 temp_t + \beta_2 rainwinter_t + \beta_3 rainharvest_t + \beta_4 age_t + u_t$$

- The variable on the left hand side $qual_t$ is called the *dependent variable* (or response). Here it is a numeric measure of the average quality of Bordeaux wines from the vintage of year t. We discuss later how that quality measure is constructed.
- The numerical variables on the right hand side $temp_t$, $rainwinter_t$, $rainharvest_t$ and age_t are called *explanatory variables* (alternative names are *regressors*, *predictors* or *features*).
  - $temp_t$ is the average temperature from April to September in the year t the wine was grown and harvested.
  - $rainwinter_t$ is the average rainfall in the winter months (October-March) before the harvest.
  - $rainharvest_t$ is the average rainfall in the main harvest month August.
  - age_t denotes the age of the vintage at the time its quality is measured. E.g. if we measure the quality of the 1961 vintage in the year 1990 we have $age_t = 29$.

Linear Regression (continued)

7 / 30

$$qual_t = \beta_0 + \beta_1 temp_t + \beta_2 rainwinter_t + \beta_3 rainharvest_t + \beta_4 age_t + u_t$$

- u_t is often called *error term*. It measures the impact of all factors influencing the dependent variable that we have not included as explicit explanatory variables. In our example, we can also more concretely refer to u_t as an unobserved quality shock of vintage t.
- β_0, ..., β_4 are unknown coefficients that we want to estimate. The exact interpretation of these coefficients can differ between applications. E.g. sometimes coefficients can be interpreted in a causal fashion, sometimes not. We will come back to this point later.

Ordinary Least Squares Estimation

8 / 30

- Ashenfelter computed estimates $\hat{\beta}$ of the the true coefficients β using ordinary least squares estimation (OLS).
- Let us briefly recap OLS estimation. Consider a linear regression model of the form:

$$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_K x_{K,t} + u_t$$

- The *predicted values* $\hat{y}_t$ of the dependent variable given an estimate $\hat{\beta}$ are given by

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t} + ... + \hat{\beta}_K x_{K,t}$$

- The *residuals*

$$\hat{u}_t = y_t - \hat{y}_t$$

  are simply the difference between the observed and predicted values of the dependent variable.

OLS estimation (continued)

- The OLS estimator is that value of $\hat{\beta}$ that minimizes the sum of squared residuals in the data set used for estimation:

$$\hat{\beta}^{OLS} = \arg\min_{\hat{\beta}} \sum_{t=1}^{T} \hat{u}_t(\hat{\beta})^2$$

- There is a closed form solution for the OLS estimator. But the formula involves matrix notation, which we don't want to introduce in this course.
- OLS estimators can be computed in R, e.g. with the command `lm` . (`lm` stands for *linear model*).

How to measure wine quality

- To estimate the linear regression, we need a numeric variable that measures the quality of a wine vintage.
- Ashenfelter et al. computed from the 1990/91 auctions a price index of every wine vintage relative to the prices of the famous 1961 vintage, whose price index we normalize to 100.
- We will discuss in the RTutor problem set in more detail how the price index is computed.

First Regression Results

- We first estimate a linear regression of a vintage's price index on all 3 weather variables (but not age). We get the following estimates:

$$\begin{aligned}\hat{p}_t = &\hat{\beta}_0 \quad + \hat{\beta}_1 temp_t \quad + \hat{\beta}_2 rainwinter_t \quad + \hat{\beta}_3 rainharvest_t \\ &= -361 + 22.3 \cdot temp_t - 0.095 \cdot rainwinter_t + 0.060 \cdot rainharvest_t\end{aligned}$$

- The intepretation of $\hat{\beta}_1 = 22.3$ is as follows:
  A by one degree Celcius higher temperature (keeping the other weather variables constant) increases our prediction of a vintage's price index by 22.3 units.
- Much stronger causal interpretations of our estimate would be: We estimate that...
  - *a by one degree Celcius higher temperature (keeping the other weather variables constant) causes a vintage's expected price index to increase by 22.3 units.*
  - *a temperature increase by one degree Celcius (keeping the other weather variables constant), increases a vintage's price index by 22.3 units.*
- It is often not correct to interpret a regression coefficient in causal fashion. Indeed a causal interpretation would not be correct for our regression above, where we don't add age as explanatory variable. We will explain this later.

Adding age as explanatory variable

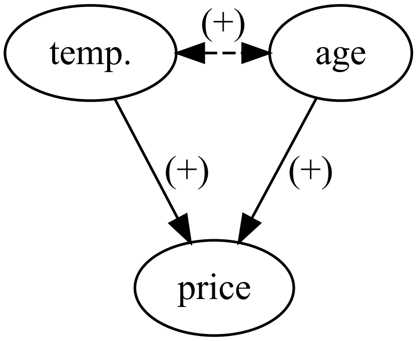
- Let us now also add the `age` of the vintage in the auction year as explanatory variable. We find

$$\begin{aligned}\hat{p}_t = \hat{\beta}_0 &+ \hat{\beta}_1 temp_t &+ \hat{\beta}_2 rainwinter_t &+ \hat{\beta}_3 rainharvest_t &+ \hat{\beta}_4 age_t \\ = -323 &+ 19.1 \cdot temp_t &- 0.1 \cdot rainwinter_t &+ 0.056 \cdot rainharvest_t &+ 0.8 \cdot age_t\end{aligned}$$

- We estimate a positive age coefficient $\hat{\beta}_4 = 0.8$.
  - If we relate it to the temperature coefficient, it seems relatively small. 10 years additional age increases our price prediction by less than an average temperature increase of 0.5 °C.
  - Does this imply that age has relatively little impact on a red wine's quality?
  - No. Even if age can strongly affect quality, it should not so strongly affect auction prices. That is because we can simply increase the age of a young vintage by storing it for some years in a cellar. The age coefficient therefore measures storage costs and other time costs rather than quality differences.
- We also see that the inclusion of `age` changes the estimated coefficients of the weather variables. E.g. the coefficient for temperature becomes smaller. Why?

A small causal model

- The figure on the right illustrates a simple world in which the price of a wine vintage is only systematically affected by the temperature during growing season and its age. Both affect price positively.
- Furthermore, age and temperature are positively correlated. We find such a positive correlation in our data. Earlier years were on average warmer. (Given global warming you might rather expect a negative correlation. But for the sample vintages until 1980, global warming had not yet a measurable effect in Bordeaux.)



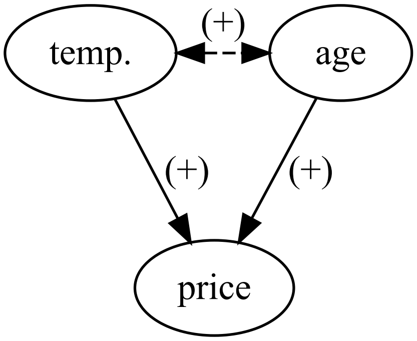
Short regression without age

- Assume we don't include `age` in the regression and estimate just the *short* regression

$$price_t = \beta_0^S + \beta_1^S temp_t + u_t^S$$

Then the OLS estimator $\hat{\beta}_1^S$ estimates the total positive relationship between `temp` and `price`. It is the sum of two sources of positive correlation:

  - The direct causal effect from temperature on price.
  - An indirect correlation because vintages with higher summer temperature are on average older vintages, but a higher age also causes a higher price.

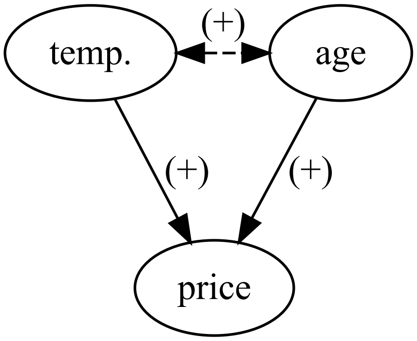


Long regression with age

- Consider the *long* regression model

$$price_t = \beta_0^L + \beta_1^L temp_+ \beta_2^L age_t + u_t^L$$

By adding `age` as additional explanatory variable, we directly *control* for the effect of age on price and it does not affect anymore the coefficient for price. This means the OLS estimator $\hat{\beta}_1^L$ now only estimates the direct causal effect from temperature on price. It is therefore smaller than the estimated coefficient $\hat{\beta}_1^S$ from the short regression.



- Ashenfelter et al. are mainly interested in a predictor for a vintage's quality when it is freshly harvested, i.e. when age=0. Yet, they still added `age` as a control variable in their regression.
- That is because they wanted to describe how weather factors affect the quality of a red wine ignoring the indirect correlation arising from the correlation between the weather variables and `age`.
- In the regression with `age` as control variable, it seems fairly reasonable to interpret the coefficients for the weather variables in a causal fashion.

More about control variables and estimation of causal effects

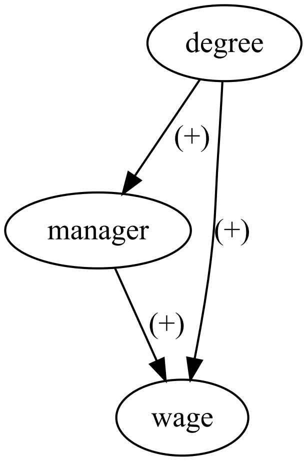
17 / 30

- You might get the wrong impression that if one wants to estimate causal effects, it is always better to add more explanatory variables to control for indirect correlations. But that is not generally true.
  1. Adding too many control variables can lead to problems of *overfitting* if you have too few observations (discussed in next section)
  2. But even with many observations, which control variables to add depends on the question you want to answer.
- Before continuing with our analysis of red wines, we want to discuss control variables and estimation of causal effects, with another example. It is about estimating the effect of education (or more precisely getting a university degree) on wages.

Estimating Returns to Education: Channel Variables

18 / 30

- Assume you want to estimate the causal effect of obtaining a university degree on average wages. In reality that is a very hard problem. But consider the simple world shown on the right.
- Wages shall only be systemtically influenced by whether a person has a university degree and by whether she is a manager. Furthermore, a university degree increases the probability to become a manager.
- Here `manager` is a *channel* variable. One channel by which a university degree positively affects wages is to make it more likely to become a manager.



Estimating Returns to Education: Channel Variables

19 / 30

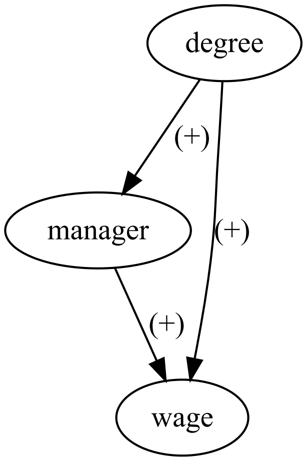
- If we estimate the long regression

$$wage_i = \beta_0^L + \beta_1^L degree_i + \beta_2^L manager_i + u_i^L$$

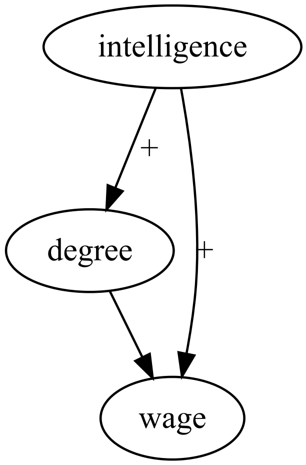
  then $\hat{\beta}_1^L$ only estimates the causal effect of a university degree on wages of the remaining channel, ignoring the channel of making it more likely to become a manager.
- If we estimate the short regression

$$wage_i = \beta_0^S + \beta_1^S degree_i + u_i^S$$

  then $\hat{\beta}_1^S$ estimates the total causal effect of a university degree on wages including all channels.
- In most studies we would be interested in the total causal effect through all channels. We should then estimate the short regression in our example.



Estimating Returns to Education: Unobserved Confounders



- A confounder is, roughly defined, a variable we need to control for in order to estimate a certain causal effect.
- For example, in the figure on the right, `intelligence` is a confounder when estimating the causal effect of `degree` on `wage`. If we don't add `intelligence` as control variable, our estimated coefficient $\hat{\beta}_1$ for `degree` would also contain the indirect correlation (higher intelligence makes it more likely to get a university degree but also directly positively impacts wages) and thus be systematically larger than the causal effect of degree on wage.
- Unfortunately, in many examples there are unobserved confounders for which we don't have data. In our example, also variables like motivation, work ethics, social skills, parent's connections may both directly affect the wages and the probability to obtain a degree. Unobserved confounders can often make it impossible or very hard to estimate certain causal effects.
- Sometimes one can use strategies like randomized experiments, instrumental variable estimation, diff-in-diff estimation, or synthetic control to overcome the problem of unobservable confounders to estimate causal effects. We will explore some of them later in this course.

Interpretation of slope coefficients in a linear regression with logged variables:

In linear regression one often uses variables in logarithms. Here are examples of how to interpret the coefficient for different combinations of variables in log and original levels:

- **level-level**: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  If x increases by one unit our predicted value of y inceases by $\hat{\beta}_1$ units.
- **log-log**: $\widehat{\log y} = \hat{\beta}_0 + \hat{\beta}_1 \log x$
  If x increases by one percent our prediction of y inceases by approximately $\hat{\beta}_1$ percent.
- **log-level**: $\widehat{\log y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  If x increases by one unit our prediction of y inceases by approximately $100 \cdot \hat{\beta}_1$ percent.
- **level-log**: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \log x$
  If x increases by one percent our prediction of y inceases by approximately $0.01 \cdot \hat{\beta}_1$ units.

Back to Bordeaux

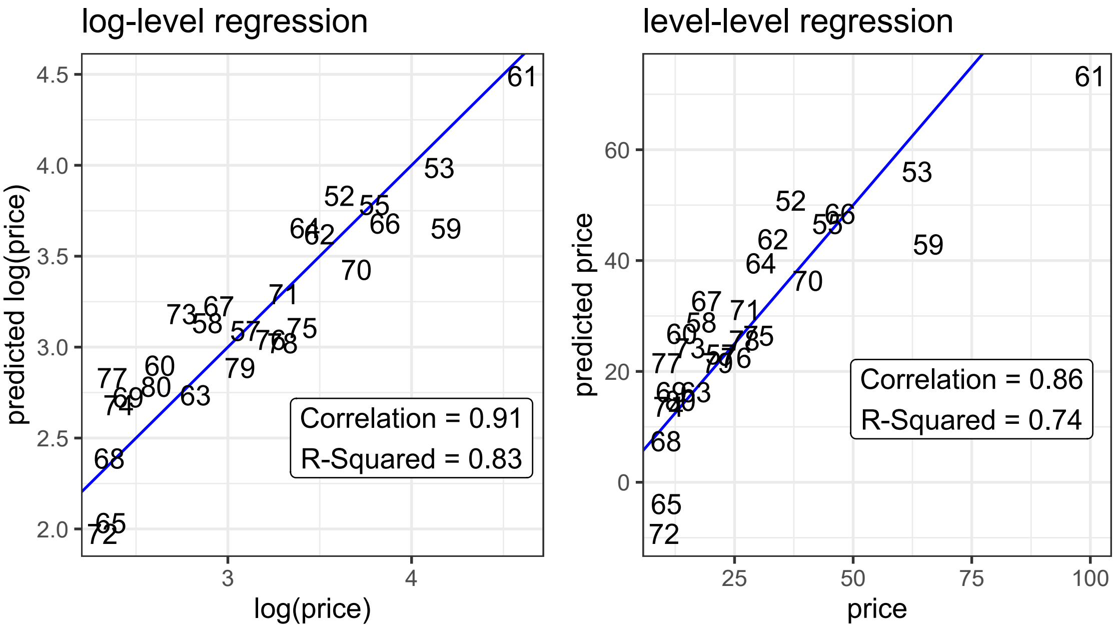
- Let's come back to estimating a formula for the quality of Bordeaux red wine vintages.
- Ashenfelter et al. did use as dependent variable not the price index in levels but the logarithm of the price index.

$$\log p_t = \beta_0 + \beta_1 temp_t + \beta_2 rainwinter_t + \beta_3 rainharvest_t + \beta_4 age_t + u_t$$

- The interpretation of $\hat{\beta}_1$ is now as follows:
  We predict that a vintage's price index increases by $100 \cdot \hat{\beta}_1$ *percent if the temperature during its growing season increases by 1 °C.*
- Think of how the price would change if a good wine and a bad wine both would get better by 1 °C higher temperature.
  - The specification with logged prices implies that the same absolute temperature increase would increase the price of the already good wine by more Euros than that of the initially bad wine.
  - This seems intuitive. Buyers of expensive wines probably are also willing to pay a lot extra if they could get an even better vintage. Instead the extra willingness to pay in absolute Euros to have mediocre wine instead of a bad one is probably not so high.

Graphical comparison

The plots compare the regressions with price index in logs (left) and levels (right). We show the dependent variable against its predicted value.



Discussion of model comparison

- We see that in the log-level regression the predicted values fit the dependent variable much better than in the level-level regression.
  - The residuals for the log-level regression seem nicely randomly distributed and have similar sizes for all price levels.
  - In the level-level regression, high prices seem systematically and strongly under predicted.
  - Also problematic is that the level-level prediction predicts negative prices for two vintages.
- Often regression summaries show the so called R-squared (R^2).
  - It is the square of the correlation between y and $\hat{y}$.
  - The R^2 measures the share of the variance of the dependent variable y that can be explained with the regression within the estimation sample.
  - The log-level model also performs better in the R^2, but the plot is even more convincing that it is the better model.

Out-of-sample prediction accuracy

- While a good in-sample fit is nice, the *much more important* criterion is a model's *out-of-sample* prediction accuracy. I.e. how well it predicts the dependent variable for new data points that have not been looked at when developing and estimating the model.
- Modern machine learning approaches to prediction problems (with larger data sets) have formalized the assessment of the out-of-sample prediction accuracy, by immediately splitting from the sample a *test data set* that won't be looked at when developing the model.
- For our estimated wine formula, we ideally would assess the the out-of-sample prediction accuracy by looking at newer vintages from 1980 onwards (which have not been used in developing the model), collect their weather data, predict the log price index and compare it to the observed price index.

Out-of-sample prediction accuracy

- The data by Ashenfelter et al. also contains weather data for the vintages from 1981 to 1989 that were not used to estimate the model.
- Applying the estimated formula for these vintages, one finds e.g. that the 1986 vintage is predicted to perform quite poorly: its predicted quality ranks at the 10% lowest place for all vintages from 1951 to 1988.
- In the Youtube video Ashenfelter argues that this vintage may be a nice test because wine critics laudatet it as a very good vintage which sharply contrasts the model formula.

Qualitative out-of-sample prediction

If one searches the internet for Bordeaux 1986, one finds different verbal assessments. A first assessment from thewinecellarinsider.com:

> Critics at the time were enamored with 1986 Bordeaux wine, when they first tasted them. But time has not been kind to most 1986 Bordeaux wine. The fruit has fled over the past few decades and with few exceptions, only the brutal, hard tannins remain. 1986 Bordeaux wine has power, structure and concentration, but most lack charm, elegance or softness. 1986 Bordeaux wine is a stern, old school Bordeaux vintage that fans of what is known as “traditional Bordeaux” enjoy.

www.falstaff.at has a description in German from 2011:

> Die Preise für die Weine aus 1986 sind mittlerweile auf ein sehr hohes Niveau geklettert und liegen für empfehlenswerte, trinkreife Weine mit etwas Potenzial (wie Gruaud-Larose oder Rauzan-Ségla) jenseits der € 150,–, angesichts der hohen En-primeur-Preise hat sich das allerdings relativiert. Eine Einzelflasche Lafite dürfte aktuell kaum unter € 1.600,– zu bekommen sein. Von den Weinen des rechten Ufers sollte man eher die Finger lassen, es sei denn, man hat das Glück, den wunderbaren Lafleur im Keller zu haben.

Note: En-primeur prices are the prices of wines that are still young.

The reviews seem mixed. While the vintage does not seem as great as initially praised by critics, it also does not seem as bad as the regression model would predict.

Out-of-sample prediction accuracy

- Unfortunately, I could nowhere find data for price indeces for the later vintages and also found no article that systematically assessed the out-of-sample prediction accuracy of the Ashenfelter formula with such a data set.
- Interestingly, even in his article from 2008 Ashenfelter does not perform such a quantitative asessement of the out-of-sample prediction accuracy of his original formula.
    - One reason could be that it might be relatively difficult to collect again the auction data for a price index.
    - Perhaps there are also lower incentives to perform the required data collection if one suspects that the out-of-sample prediction performance might not be as good as hoped.
- Ashenfelter (2008) discusses, however, qualitatively the out-of-sample prediction accuracy for some other vintages like 2000 and 2003 by comparing model predictions and critics assessment. He does not discuss there the 1986 vintage anymore, however.

Assessment of the wine formula

- Looking just at the in-sample fit, the model looks very nice and intuitive for predicting the quality of young wines. It is not often that one has such a good fit by economic models with so few explanatory variables.
- However, a systematic assessment of the out-of-sample prediction accuracy is missing (or at least I have not found it). But out-of-sample prediction accuracy is probably the most important criterion for a prediction model. In this sense, it is hard to assess the quality of the model.
- We will come back to out-of-sample prediction accuracy in Chapter 2 where we discuss machine learning.

References

- Ashenfelter, O., Ashmore, D., & Lalonde, R. (1995). Bordeaux wine vintage quality and the weather. Chance, 8(4), 7-14.
- Ashenfelter, O. (2008). Predicting the quality and prices of Bordeaux wine. The Economic Journal, 118(529).