L3: Linear Regression

Shan Wang

Lingnan College, Sun Yat-sen University

2020 Data Mining and Machine Learning LN3119

https://wangshan731.github.io/DM-ML/

Last lecture

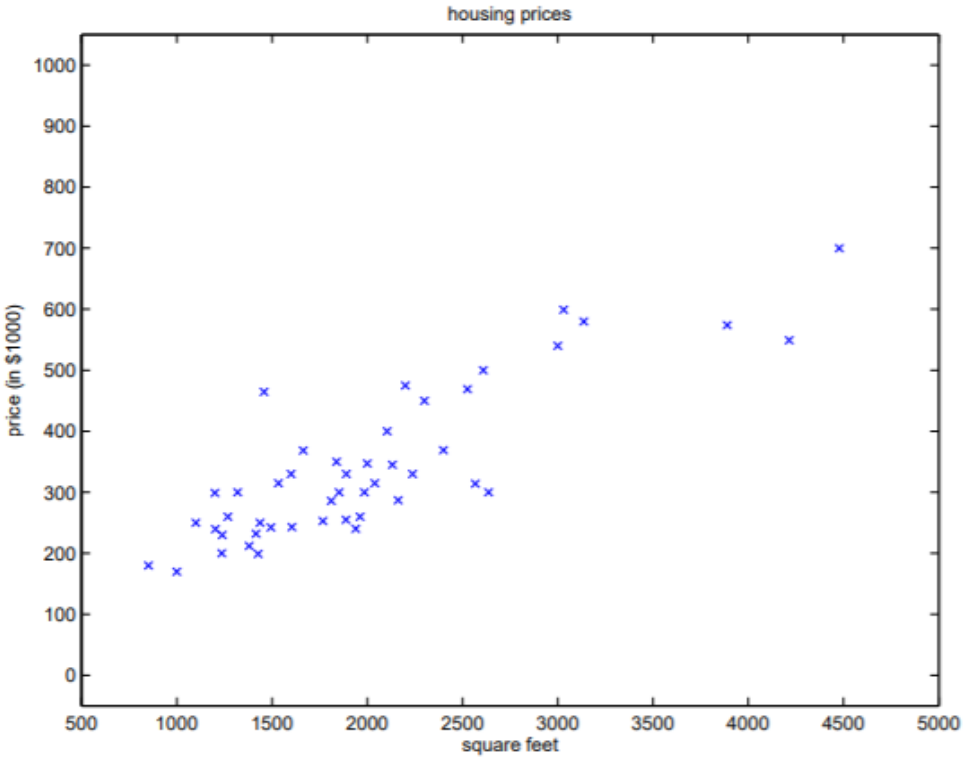
- Basics of supervised learning
  - Learning process: train and prediction
  - Discriminative models and generative models
    - $p_\theta(y|x)$ v.s. $p_\theta(x, y)$
  - Machine learning three elements
    - Model
    - Strategy:
      - loss function
      - generalization error v.s. empirical error
      - MLE v.s. MAP
    - Algorithm
  - Model evaluation
    - For classification and regression
  - Model selection & Regularization
  - Cross validation

Course outline

- Supervised learning
  - Linear regression
  - Logistic regression
  - SVM and kernel
  - Tree models
- Unsupervised learning
  - Clustering
  - PCA
  - EM
- Deep learning
  - Neural networks
  - Convolutional NN
  - Recurrent NN
- Reinforcement learning
  - MDP
  - ADP
  - Deep Q-Network

- Relationship of house living area and price

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



This lecture

- Linear regression method
  - Model
  - Strategy
    - Least squared error
    - Maximum likelihood
  - Algorithm
    - Normal equation
    - Gradient descent method
- Regularization
- Application: quality of wine

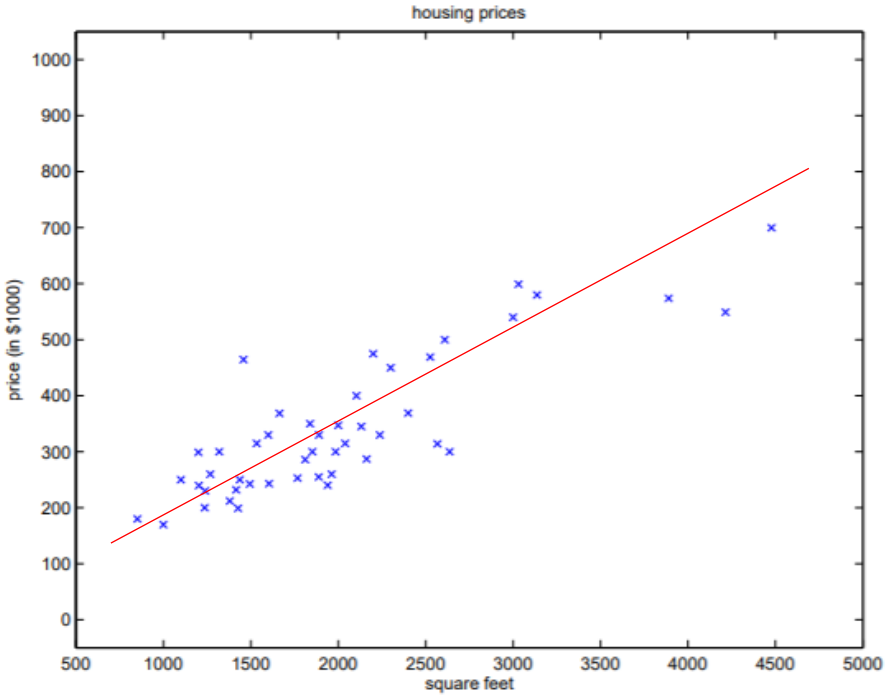
Reference: VE 445, Shuai LI (SJTU); OPIM 326 Daniel Zheng (SMU)

Model

Linear relationship

- Use linear relationship to approximate the function of y and $\boldsymbol{x}$
  - i.e., $y = \boldsymbol{\theta}'\boldsymbol{x}$

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



*Intercept term is omitted in this lecture, you may imagine the first term in $\boldsymbol{x}$ is always 1.

Strategy

Mean squared error

MSE

- How to select the most appropriate linear model?
  - i.e., what is the **strategy**?
- Error: **Mean squared error (MSE)**

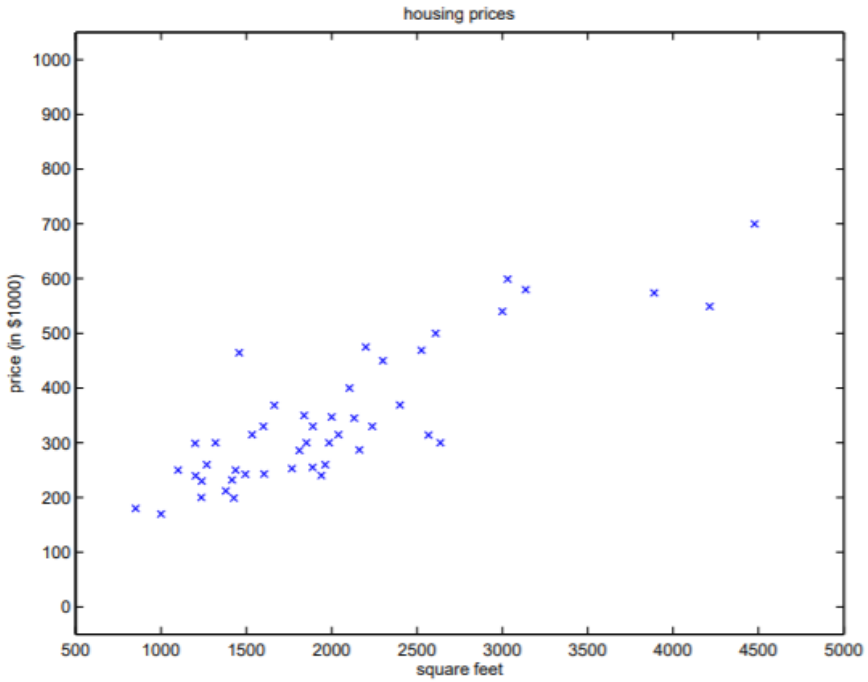
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

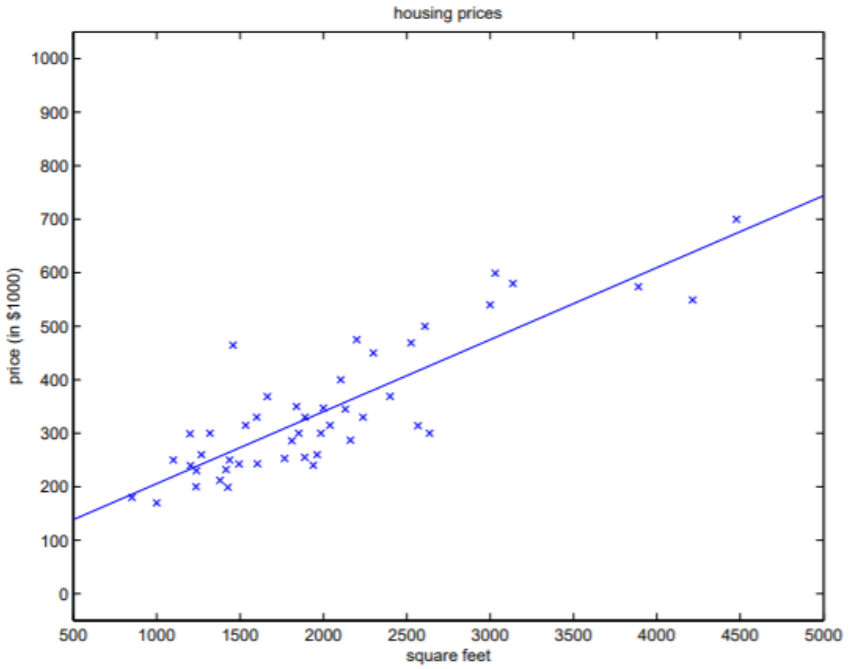
  - Where y and $\hat{y}$ are the true values and predicted values

Minimize MSE

- Find the linear model $y = \boldsymbol{\theta}'\boldsymbol{x} + \varepsilon$ with the smallest MSE

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin} \hat{R}(\boldsymbol{\theta}) = \operatorname{argmin} \frac{1}{N} \sum_{i=1}^{N} (y_i - \boldsymbol{\theta}'\boldsymbol{x}_i)^2$$





Strategy

Maximum likelihoodd

MLE (Probabilistic view)

- Assume for each sampled $(\boldsymbol{x}, y) \sim D$

$$y = \boldsymbol{\theta}'\boldsymbol{x} + \varepsilon$$

- where $\varepsilon \sim N(0, \sigma^2)$, or equivalently, $y \sim N(\boldsymbol{\theta}'\boldsymbol{x}, \sigma^2)$

- The estimator $\widehat{\boldsymbol{\theta}}$ is the maximal likelihood estimator (MLE) of the data $(\boldsymbol{X}, Y)$

$$\widehat{\boldsymbol{\theta}} = \operatorname{argmax} P(Y|\boldsymbol{X}, \boldsymbol{\theta})$$

$$= \operatorname{argmax} \prod P(y_i|\boldsymbol{x}_i, \theta)$$

$$= \operatorname{argmax} \log \prod P(y_i|\boldsymbol{x}_i, \theta)$$

MLE = least squared error

$$\widehat{\boldsymbol{\theta}} = \text{argmax} \sum_{i=1}^{N} \log P(y_i|\boldsymbol{x}_i, \theta)$$

- $P(y_i|x_i, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \boldsymbol{\theta}' \boldsymbol{x}_i)^2}{2\sigma^2}}$
- $\log P(y_i|x_i, \theta) = \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y_i - \boldsymbol{\theta}' \boldsymbol{x}_i)^2}{2\sigma^2}$

$$\widehat{\boldsymbol{\theta}} = \text{argmax} \sum_{i=1}^{N} -(y_i - \boldsymbol{\theta}' \boldsymbol{x}_i)^2$$

$$= \text{argmin} \frac{1}{N} \sum_{i=1}^{N} (y_i - \boldsymbol{\theta}' \boldsymbol{x}_i)^2$$

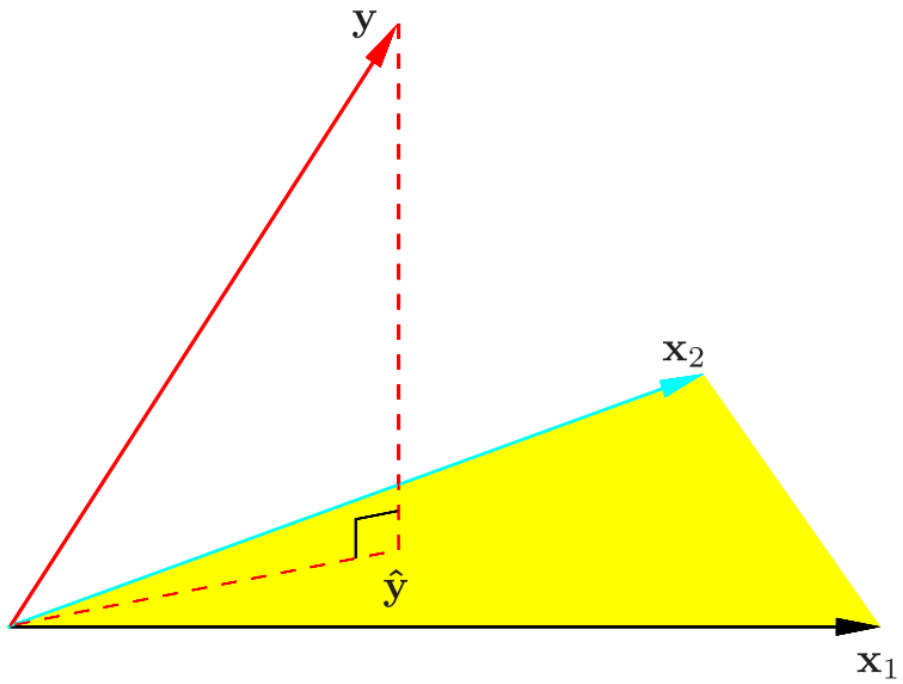
Algorithm

Normal equation

Normal equation

$$\min \frac{1}{N}\sum_{i=1}^{N}(y_i - \boldsymbol{\theta}'\boldsymbol{x}_i)^2$$

- Matrix form: $\min \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\theta})'(\boldsymbol{y} - \boldsymbol{X\theta})$
- FOC:
- $-\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X\theta}) = \boldsymbol{0}$
- $\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X\theta}$
- $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
- $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
  - Hat matrix
  - Projection of y_i

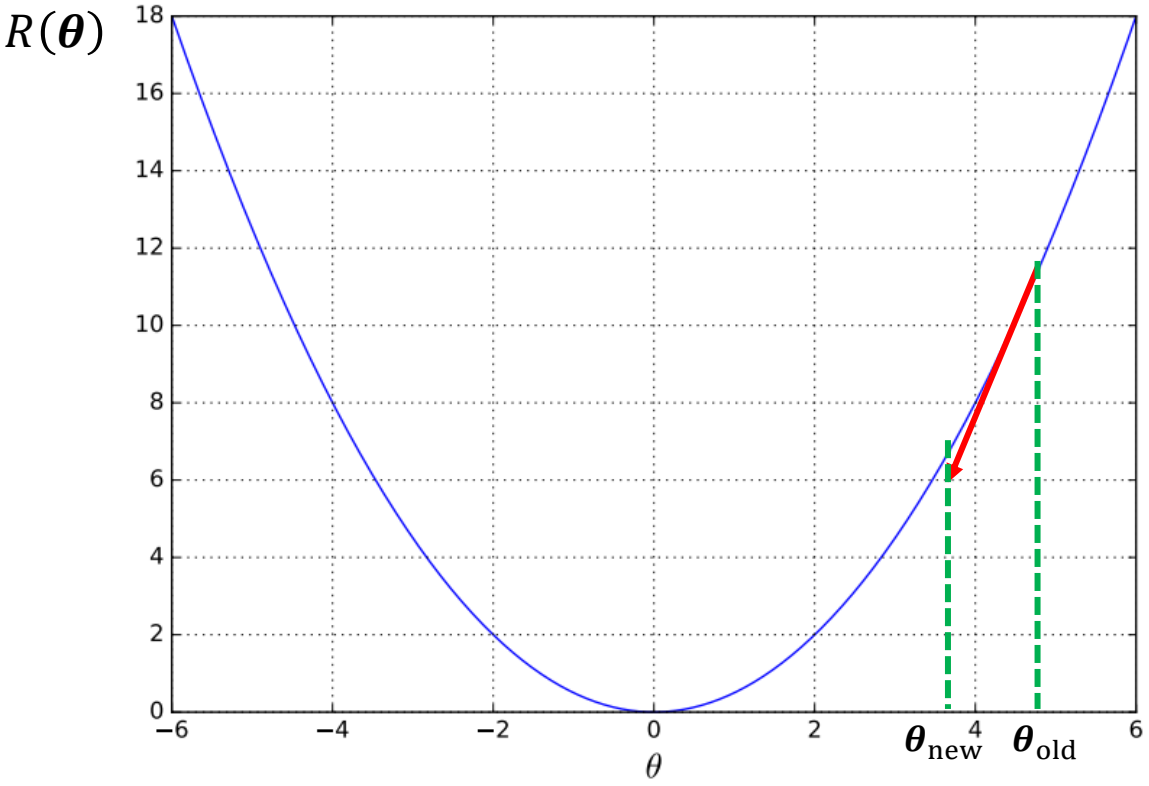


Algorithm

Gradient descent

Gradient descent

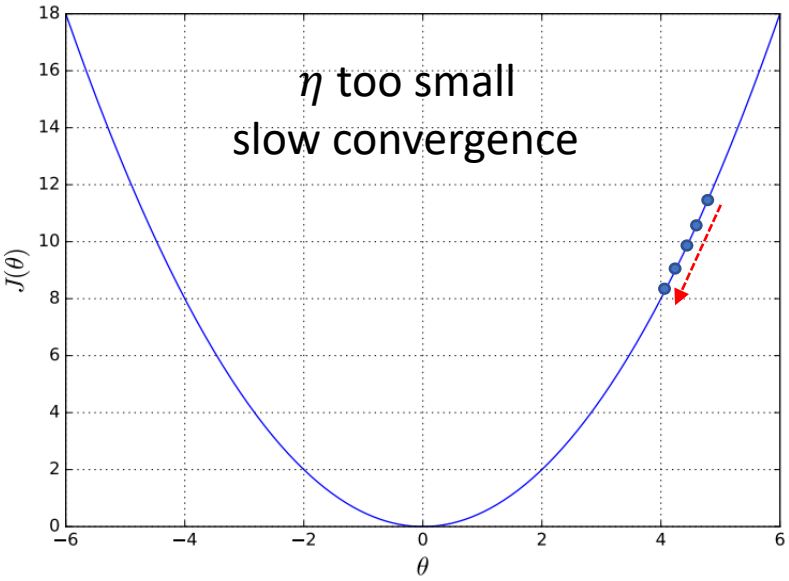
- When the problem size goes large, it is time costly to take inverse matrix
- But the error function $R(\boldsymbol{\theta})$ is convex

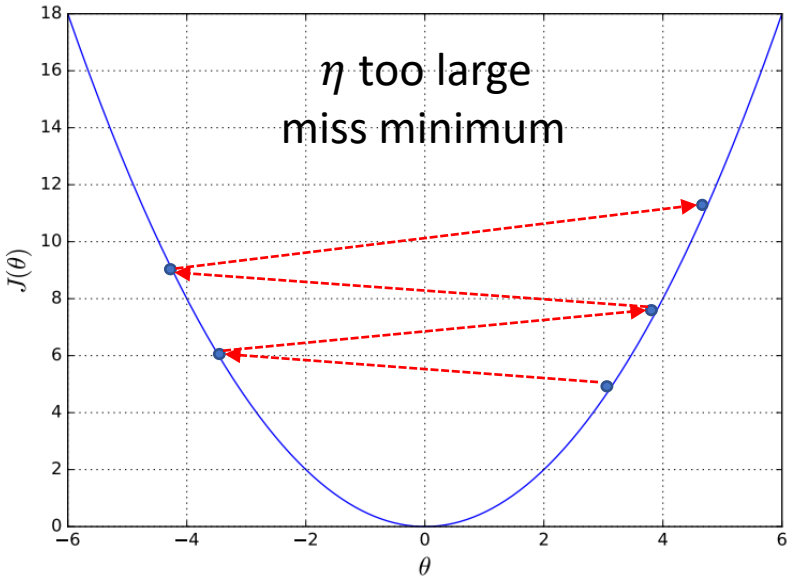


$$\boldsymbol{\theta}_{\text{new}} \leftarrow \boldsymbol{\theta}_{\text{old}} - \eta \frac{\partial R(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Gradient descent – Learning rate

- $\boldsymbol{\theta}_{\text{old}} - \eta \frac{\partial R(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$

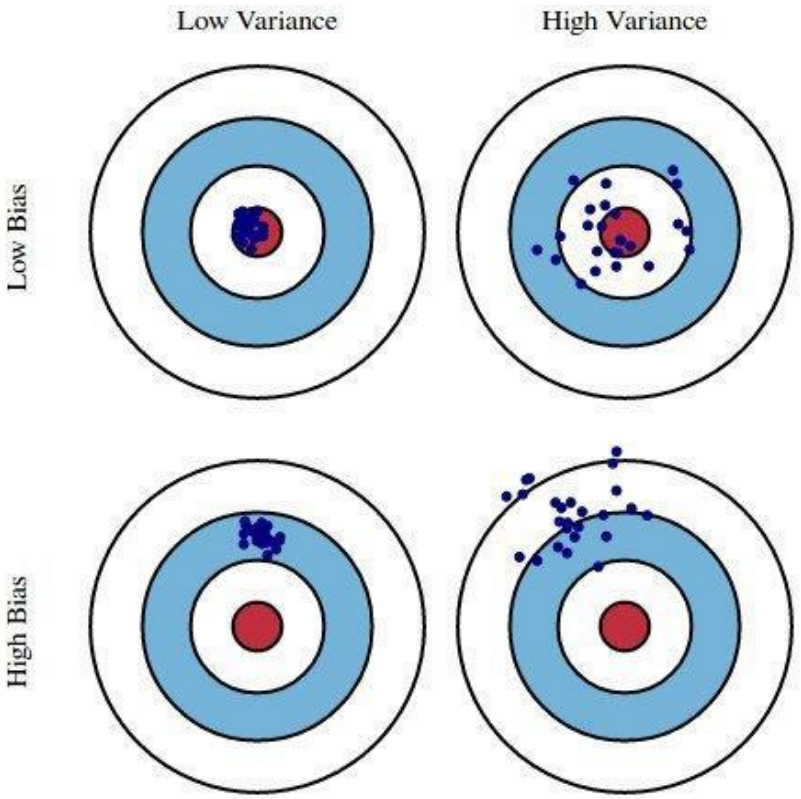




- To see if gradient descent is working, print out $R(\boldsymbol{\theta})$ for each or every several iterations. If $R(\boldsymbol{\theta})$ does not drop properly, adjust η

Regularization

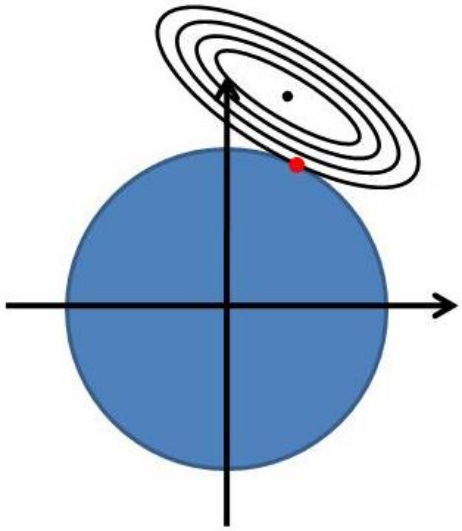
- Bias – Variance trade-off



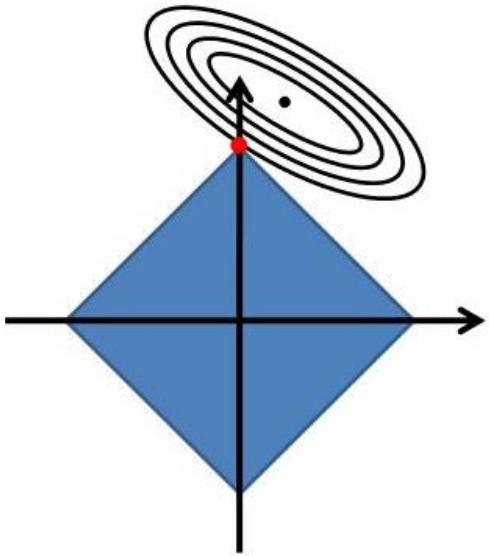
- Ordinary least square (OLS)
  - Unbiased
  - Can have huge variance
    - Multi-collinearity
      - When x_i and x_j are correlated to each other and the y
    - Too many x_i
      - Number of x is close to the number of data
- Solution: reduce variance at the cost of bias
  - Add a penalty term (regularization)

Ridge and Lasso

- L2-norm (Ridge):
  - $\Omega(f = ax + b) = a^2 + b^2$
  - $\widehat{\boldsymbol{\theta}} = \text{argmin} \sum_{i=1}^{N} (y_i - \boldsymbol{\theta}' \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{\theta}\|_2^2$



- L1-norm (Lasso):
  - $\Omega(f = ax + b) = |a| + |b|$
  - $\widehat{\boldsymbol{\theta}} = \text{argmin} \sum_{i=1}^{N} (y_i - \boldsymbol{\theta}' \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{\theta}\|_1$



- λ: trade-off between bias and variance
  - Chosen by cross validation

More on Ridge

- Solution to ridge regression

$$\min \sum_{i=1}^{N} (y_i - \boldsymbol{\theta}' \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{\theta}\|_2^2$$

  - Matrix form: $\min \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X\theta})' (\boldsymbol{y} - \boldsymbol{X\theta}) + \frac{1}{2} \lambda \boldsymbol{\theta}' \boldsymbol{\theta}$
  - FOC:
  - $-\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X\theta}) + \lambda \boldsymbol{I\theta} = \boldsymbol{0}$
  - $\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})\boldsymbol{\theta}$
  - $\widehat{\boldsymbol{\theta}}_{ridge} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$

- Ridge regression and MAP?
  - Bonus question: what is relationship between them?
  - Hint: consider an MAP estimator with Gaussian prior

Application

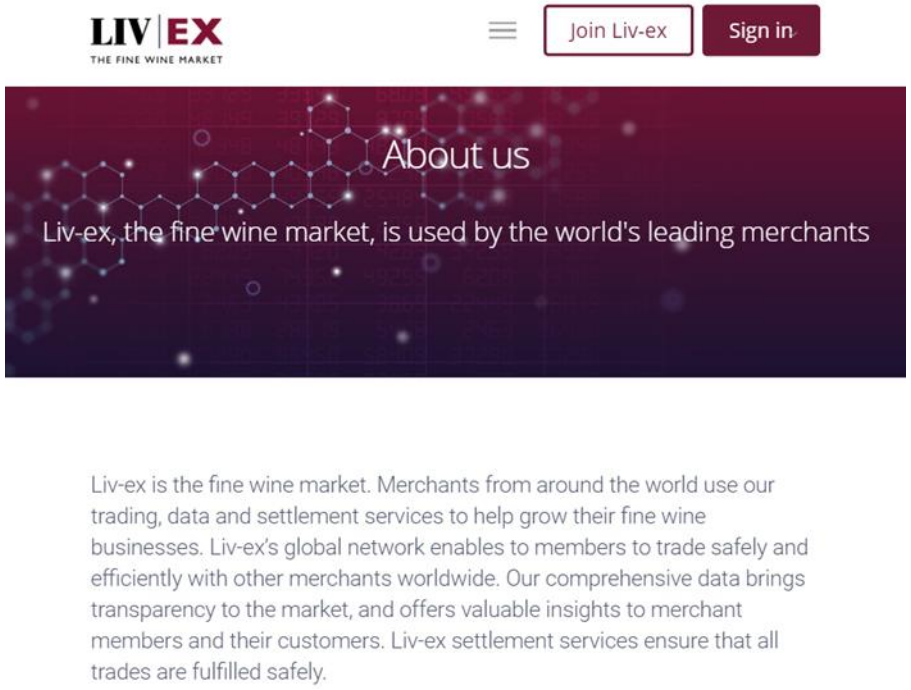
Quality of Wine

Wine Industries

- Large differences in price and quality between years, although wine is produced in a similar way
- Meant to be aged, so hard to tell if wine will be good when it is on the market
- Quality and future price is crucial for planting, harvesting, ageing, bottling decisions
- Considering capacity limit in various process and warehousing

Wine Markets

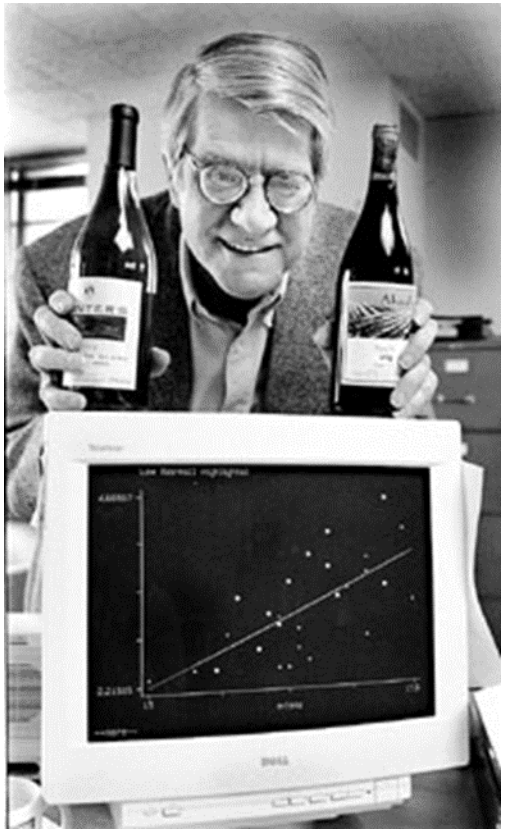
- Various trading markets and platforms
- Wine futures
- Expert tasters predict which ones will be good



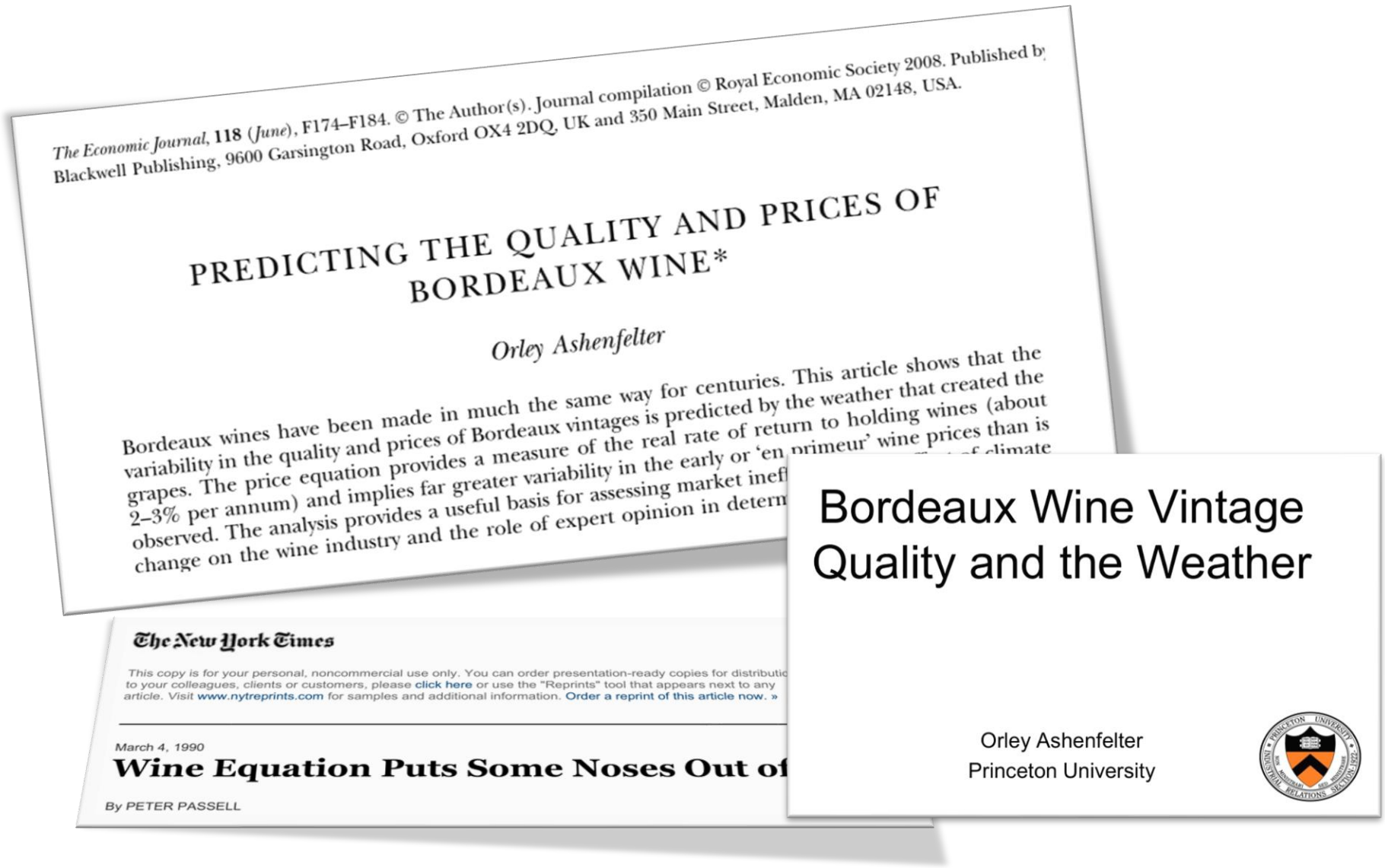
- Can machine learning be used to develop a different system for judging wine?

Predicting the Quality of Wine

- March 1990, Orley Ashenfelter, a Princeton economics professor, claimed that he can predict wine quality without tasting the wine
- Wine quality is measured using price (index)



Application - Quality of Wine



The Economic Journal, **118** (*June*), F174–F184. © The Author(s). Journal compilation © Royal Economic Society 2008. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

PREDICTING THE QUALITY AND PRICES OF BORDEAUX WINE*

Orley Ashenfelter

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 inef... climate change on the wine industry and the role of expert opinion in determ...

The New York Times

March 4, 1990

Wine Equation Puts Some Noses Out of...

By PETER PASSELL

Bordeaux Wine Vintage Quality and the Weather

Orley Ashenfelter
Princeton University

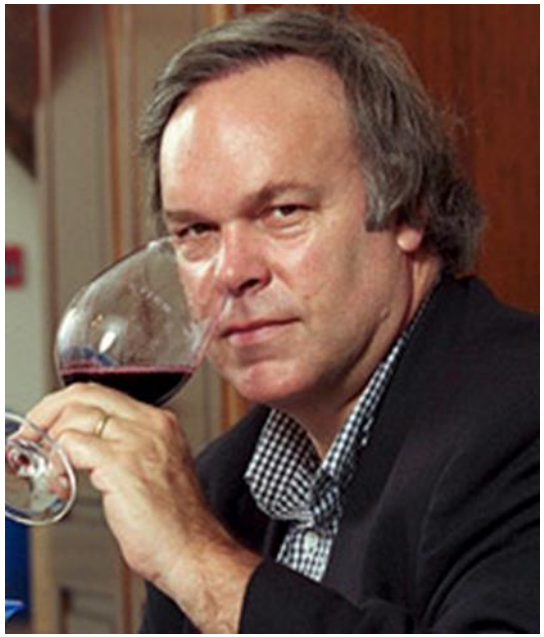
The Expert's Reaction

Robert Parker, the world's most influential wine expert:

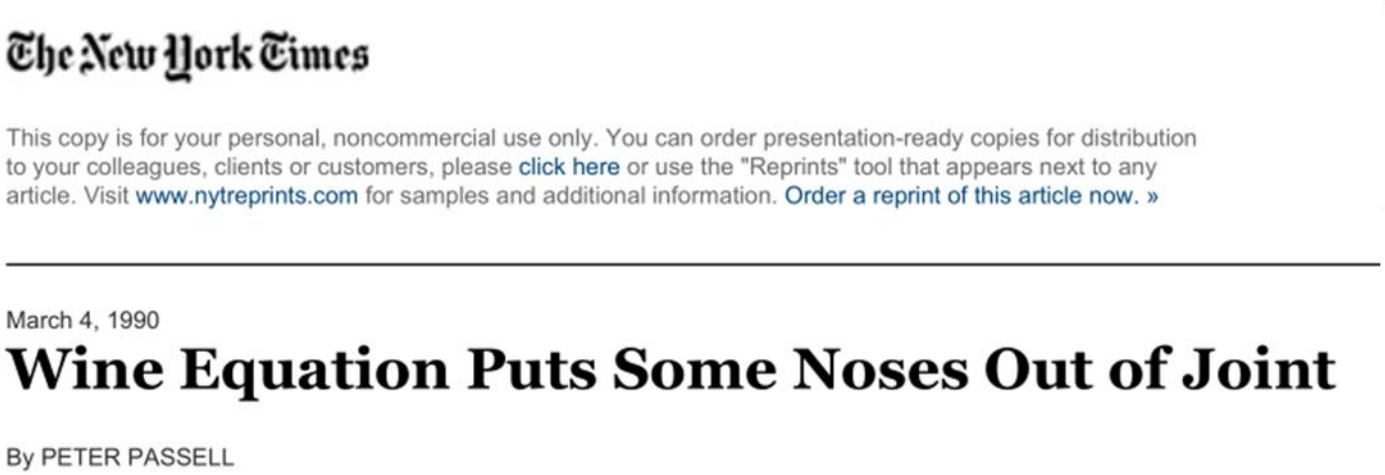
"Ashenfelter is an absolute total sham"

"ludicrous and absurd"

"rather like a movie critic who never goes to see the movie but tells you how good it is based on the actors and the director"



The Industry's Reaction



The New York Times

March 4, 1990

Wine Equation Puts Some Noses Out of Joint

By PETER PASSELL

- William Sokolin, a New York wine merchant, said th e Bordeaux wine industry's view of the work ranges 'somewhere between violent and hysterical'

Orley Ashenfelter in Good Morning America

https://www.youtube.com/watch?v=Ec8hPHLMyzY

A Naïve Prediction

- Take the historical average price
- For any wine, always predict its price using this historical average
- Any reasonable prediction should beat this naïve prediction
- This naïve prediction can be used as a benchmark to measure how well any other prediction performs

Building a Model

- Ashenfelter used **linear regression**
  - Predicts an outcome variable, or *dependent/response variable*
  - Using a set of *independent/explanatory variables*
- Dependent variable: Price (Index)
  - Price Index measures price of many different wineries in thousands of 1990-1991 wine auctions
- Independent variables:
  - Age (older wines are more expensive)
  - Weather
    - Average Growing Season Temperature (AGST)
    - Harvest Rain
    - Winter Rain

Let’s get our hands dirty!

Data Analysis

Let's do some basic data analysis using our WHO data.

Hide

```
WHO$Under15
```

```
  [1] 47.42 21.33 27.42 15.20 47.58 25.96 24.42 20.34 18.95 14.51 22.25 21.62 20.16 30.57 18.99 15.10 16.88 34.4
0 42.95 28.53
 [21] 35.23 16.35 33.75 24.56 25.75 13.53 45.66 44.20 31.23 43.88 16.37 30.17 40.07 48.52 21.38 17.95 28.03 42.1
7 42.37 30.61
 [41] 23.94 41.48 14.98 16.58 17.16 14.56 21.98 45.11 17.66 33.72 25.96 30.53 30.29 31.25 30.62 38.95 43.10 15.6
9 43.29 28.88
 [61] 16.42 18.26 38.49 45.90 17.62 13.17 38.59 14.60 26.96 40.80 42.46 41.55 36.77 35.35 35.72 14.62 20.71 29.4
3 29.27 21.68
 [81] 40.51 21.54 27.53 14.04 27.78 13.12 34.13 25.46 42.37 30.10 24.90 30.21 35.61 14.57 21.64 36.75 43.06 29.4
5 15.13 17.46
[101] 42.72 45.64 26.65 29.03 47.14 14.98 30.10 40.22 20.17 29.02 35.81 18.26 27.05 19.01 27.85 45.38 25.28 36.5
9 30.10 35.58
[121] 17.21 20.26 33.37 49.99 44.23 30.61 18.64 24.19 34.31 30.10 28.65 38.37 32.78 29.18 34.53 14.91 14.92 13.2
8 15.25 16.52
[141] 15.05 15.45 43.56 25.96 24.31 25.70 37.88 14.04 41.60 29.69 41.54 16.45 21.95 41.74 16.48 15.00 14.16 40.3
7 47.35 29.53
[161] 42.28 15.20 25.15 41.48 27.83 38.05 16.71 14.79 35.35 35.75 18.47 16.89 46.33 41.89 37.33 20.73 23.22 26.0
0 28.65 30.61
[181] 48.54 14.18 14.41 17.54 44.85 19.63 22.05 28.90 37.37 28.84 22.87 40.72 46.73 40.24
```

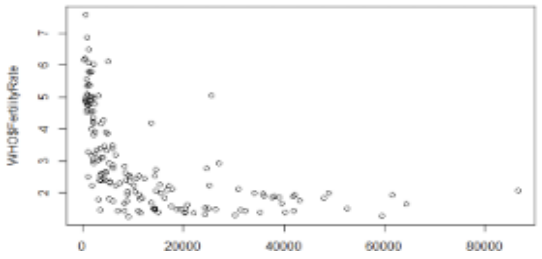
```
WHO$Country[which.min(WHO$Under15)]
```

```
[1] Japan
194 Levels: Afghanistan Albania Algeria Andorra Angola Antigua and Barbuda Argentina Armenia Australia Austria
... Zimbabwe
```

Let's create some plots for exploratory data analysis (EDA). First, Let's create a basic scatterplot of GNI versus FertilityRate.

Hide

```
plot(WHO$GNI, WHO$FertilityRate)
```



Interpreting R Output

```
Call:
lm(formula = Price ~ AGST, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-0.78450 -0.23882 -0.03727  0.38992  0.90318

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.4178     2.4935  -1.371 0.183710
AGST          0.6351     0.1509   4.208 0.000335 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4993 on 23 degrees of freedom
Multiple R-squared:  0.435,  Adjusted R-squared:  0.4105
F-statistic: 17.71 on 1 and 23 DF,  p-value: 0.000335
```

Estimated β_0

Estimated β_1

p-value for each independent variable; Smaller is more significant

AGST is significant at the 0.1% level

R^2

Adding Variables

Independent Variables Used	R^2
AGST	0.4350
AGST, HarvestRain	0.7074
AGST, HarvestRain, WinterRain	0.7537
AGST, HarvestRain, WinterRain, Age	0.8286
AGST, HarvestRain, WinterRain, Age, FrancePopulation	0.8294

- Adding more variables **always** increases R^2
- Diminishing returns as more variables are added
  - Incremental amount in R^2 is getting smaller and smaller
  - This is a usual observation, **NOT** always

Out-of-Sample R^2

We can call the 'predict' function to make predictions for the test points.

```
predictTest = predict(model4, newdata = wineTest)
predictTest
```

```
##        1        2
## 6.768925 6.684910
```

To assess the accuracy of our predictions, we can compute the out-of-sample R-squared.

```
SSE = sum((wineTest$Price - predictTest)^2)
SST = sum((wineTest$Price - mean(wine$Price))^2)
1 - SSE/SST
```

```
## [1] 0.7944278
```

The baseline model is still the average price of the training set!

Out-of-Sample R^2

Independent Variables Used	R^2	Out-of-sample R^2
AGST	0.4350	0.7882
AGST, *HarvestRain*	0.7074	-0.0819
AGST, *HarvestRain*, *WinterRain*	0.7537	0.2626
AGST, HarvestRain, *WinterRain*, *Age*	0.8286	0.7944
AGST, *HarvestRain*, *WinterRain*, *Age*, *FrancePopulation*	0.8294	0.7566

- Out-of-sample R^2 is manually computed using the test set data
- Better R^2 does not necessarily mean better out-of-sample R^2
- Out-of-sample R^2 can be negative!

The Results

- **Parker (in 1991):**
  - 1986 is “very good to sometimes exceptional”
- **Ashenfelter (in 1991):**
  - 1986 is mediocre
  - 1989 will be “the wine of the century” and 1990 will be even better!
- **In 2003,** virtually unanimous agreement that **1989 and 1990 are outstanding**
  - 1989 sold for more than twice the price of 1986
  - 1990 sold for even higher prices!
- In 2003, Ashenfelter predicted 2000 and 2003 would be great
- Parker has stated that “2000 is the greatest vintage Bordeaux has ever produced”

Wine Analytics

informs
http://pubsonline.informs.org/journal/msom/

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 19, No. 2, Spring 2017, pp. 202–215
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Wine Analytics: Fine Wine Pricing and Selection Under Weather and Market Uncertainty

Mert Hakan Hekimoğlu,[a] Burak Kazaz,[b] Scott Webster[c]

[a] Lally School of Management, Rensselaer Polytechnic Institute, Troy, New York 12180; [b] Whitman School of Management, Syracuse University, Syracuse, New York 13244; [c] W. P. Carey School of Business, Arizona State University, Tempe, Arizona 85287
Contact: hekimm@rpi.edu (MHH); bkazaz@syr.edu (BK); scott.webster@asu.edu (SW)

Received: July 22, 2015
Revised: November 12, 2015; March 24, 2016; July 23, 2016
Accepted: September 5, 2016
Published Online in Articles in Advance: November 29, 2016

https://doi.org/10.1287/msom.2016.0602

Abstract. We examine a risk-averse distributor's decision in selecting between bottled wine and wine futures under weather and market uncertainty. At the beginning of every summer, a fine wine distributor has to choose between purchasing bottled wine made from the harvest collected two years ago and wine futures of wine still aging in the barrel from the previous year's harvest. At the end of the summer, after seeing weather and market fluctuations, the distributor can adjust its allocation by trading futures and bottles.

This paper makes three contributions. First, we develop an analytical model to determine the optimal selection of bottled wine and wine futures under weather and market uncertainty. Our model is built on an empirical foundation in which the functional forms describing the evolution of futures and bottle prices are derived from comprehensive data associated with the most influential Bordeaux winemakers. Second, we develop structural properties of optimal decisions. We show that a wine distributor should always invest in wine futures because it increases the expected profit in spite of being a riskier asset than bottled wine. We characterize the influence of variation in various uncertainties in the problem. Third, our study empirically demonstrates for a large distributor the financial benefits of using our model. The hypothetical average profit improvement in our numerical analysis is significant, exceeding 21%, and its value becomes higher under risk aversion. The analysis is beneficial for fine wine distributors, as it provides insights into how to improve their selection in order to make financially healthier allocations.

Wine Analytics

https://www.youtube.com/watch?v=fsdWJtdvQoQ

Lecture 3 wrap-up

- ✓Linear regression method
  - ✓Model
  - ✓Strategy
    - ✓Least squared error
    - ✓Maximum likelihood
  - ✓Algorithm
  - ✓Solve the $\widehat{\boldsymbol{\theta}}$
    - ✓Normal equation
    - ✓Gradient descent method
- ✓Regularization
- ✓Application: quality of wine

Assignment 3

- Get used to R again
  - If you have any questions, send your questions to TA.
  - TA will collect them.

- One bonus question (NOT required):
  - What is the relationship between Ridge regression and MAP?
  - Hint: consider an MAP estimator with Gaussian prior
  - Send your answer to TA (any form, e.g., word, pdf, photo …)

- Due: May 5, 11pm
- TA: Mr. Xiong, xiongym3@mail2.sysu.edu.cn

Next lecture

- Supervised learning
  - Linear regression
  - Logistic regression
  - SVM and kernel
  - Tree models
- Unsupervised learning
  - Clustering
  - PCA
  - EM
- Deep learning
  - Neural networks
  - Convolutional NN
  - Recurrent NN
- Reinforcement learning
  - MDP
  - ADP
  - Deep Q-Network

2020 Data Mining and Machine Learning LN3119
https://wangshan731.github.io/DM-ML/

Questions?

Shan Wang (王杉)

https://wangshan731.github.io/