Bordeaux Wine

• 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
• Expert tasters predict which ones will be good
• Can analytics be used to come up with a different system for judging wine?
Predicting the Quality of Wine

- March 1990 - Orley Ashenfelter, a Princeton economics professor, claims he can predict wine quality without tasting the wine.
Building a Model

• Ashenfelter used a method called **linear regression**
  • Predicts an outcome variable, or *dependent variable*
  • Predicts using a set of *independent variables*

• Dependent variable: typical price in 1990-1991 wine auctions (approximates quality)

• Independent variables:
  • Age – older wines are more expensive
  • Weather
    • Average Growing Season Temperature
    • Harvest Rain
    • Winter Rain
The Data (1952 – 1978)
The Expert’s Reaction

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

“Ashenfelter is an absolute total sham”

“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”
One-Variable Linear Regression

\[ y = 7.07 \]
\[ y = 0.5(AGST) - 1.25 \]
The Regression Model

- One-variable regression model

\[ y^i = \beta_0 + \beta_1 x^i + \epsilon^i \]

- The best model (choice of coefficients) has the smallest error terms

\( y^i \) = dependent variable (wine price) for the \( i \)th observation
\( x^i \) = independent variable (temperature) for the \( i \)th observation
\( \epsilon^i \) = error term for the \( i \)th observation
\( \beta_0 \) = intercept coefficient
\( \beta_1 \) = regression coefficient for the independent variable
Selecting the Best Model

\[ \text{SSE} = (\varepsilon_1)^2 + (\varepsilon_2)^2 + \ldots + (\varepsilon_N)^2 \]

\[ N = \text{# data points} \]
Selecting the Best Model

\[ \text{Avg Growing Season Temp (Celsius)} \]

\[ \begin{align*}
\text{SSE} &= 10.15 \\
\text{SSE} &= 6.03 \\
\text{SSE} &= 5.73
\end{align*} \]
Other Error Measures

- SSE can be hard to interpret
  - Depends on N
  - Units are hard to understand

- Root-Mean-Square Error (RMSE)

\[ RMSE = \sqrt{\frac{SSE}{N}} \]

- Normalized by N, units of dependent variable
\[ R^2 \]

- Compares the best model to a “baseline” model
- The baseline model does not use any variables
  - Predicts same outcome (price) regardless of the independent variable (temperature)
\[ R^2 \]

\[ \text{SSE} = 5.73 \]
\[ \text{SST} = 10.15 \]

\[ R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \]

\[ = 1 - \frac{5.73}{10.15} \]

\[ = 0.44 \]

---

15.071x – The Statistical Sommelier: An Introduction to Linear Regression
Interpreting $R^2$

$$R^2 = 1 - \frac{SSE}{SST} \quad 0 \leq SSE \leq SST \quad 0 \leq SST$$

- $R^2$ captures value added from using a model
- $R^2 = 0$ means no improvement over baseline
- $R^2 = 1$ means a perfect predictive model
- Unitless and universally interpretable
  - Can still be hard to compare between problems
  - Good models for easy problems will have $R^2 \approx 1$
  - Good models for hard problems can still have $R^2 \approx 0$
Available Independent Variables

• So far, we have only used the Average Growing Season Temperature to predict wine prices

• Many different independent variables could be used
  • Average Growing Season Temperature
  • Harvest Rain
  • Winter Rain
  • Age of Wine (in 1990)
  • Population of France
Multiple Linear Regression

• Using each variable on its own:
  • $R^2 = 0.44$ using Average Growing Season Temperature
  • $R^2 = 0.32$ using Harvest Rain
  • $R^2 = 0.22$ using France Population
  • $R^2 = 0.20$ using Age
  • $R^2 = 0.02$ using Winter Rain

• Multiple linear regression allows us to use all of these variables to improve our predictive ability
The Regression Model

- Multiple linear regression model with \( k \) variables

\[
y^i = \beta_0 + \beta_1 x^i_1 + \beta_2 x^i_2 + \ldots + \beta_k x^i_k + \epsilon^i
\]

- Best model coefficients selected to minimize SSE
# Adding Variables

- Adding more variables can improve the model
- Diminishing returns as more variables are added

<table>
<thead>
<tr>
<th>Variables</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Growing Season Temperature (AGST)</td>
<td>0.44</td>
</tr>
<tr>
<td>AGST, Harvest Rain</td>
<td>0.71</td>
</tr>
<tr>
<td>AGST, Harvest Rain, Age</td>
<td>0.79</td>
</tr>
<tr>
<td>AGST, Harvest Rain, Age, Winter Rain</td>
<td>0.83</td>
</tr>
<tr>
<td>AGST, Harvest Rain, Age, Winter Rain, Population</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Selecting Variables

- Not all available variables should be used
  - Each new variable requires more data
  - Causes overfitting: high $R^2$ on data used to create model, but bad performance on unseen data

- We will see later how to appropriately choose variables to remove
Understanding the Model and Coefficients

| Coefficients: | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|---------|
| (Intercept)   | -4.504e-01 | 1.019e+01  | -0.044  | 0.965202|
| AvgGrowingSeasonTemp | 6.012e-01  | 1.030e-01  | 5.836   | 0.000233*** |
| HarvestRain   | -3.958e-03 | 8.751e-04  | -4.523  | 0.000233*** |
| Age           | 5.847e-04  | 7.900e-02  | 0.007   | 0.994172 |
| WinterRain    | 1.043e-03  | 5.310e-04  | 1.963   | 0.064416 |
| FrancePopulation | -4.953e-05 | 1.667e-04  | -0.297  | 0.769578 |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Correlation

A measure of the linear relationship between variables

+1 = perfect positive linear relationship
0 = no linear relationship
-1 = perfect negative linear relationship
Examples of Correlation

\[ \text{Cor} = 0.14 \]
Examples of Correlation

\[ \text{Cor} = -0.06 \]
Examples of Correlation

$$\text{cor} = -0.99$$

Population of France (thousands)

Age of Wine (Years)
Predictive Ability

• Our wine model had a value of $R^2 = 0.83$

• Tells us our accuracy on the data that we used to build the model

• But how well does the model perform on new data?

• Bordeaux wine buyers profit from being able to predict the quality of a wine years before it matures
### Out-of-Sample $R^2$

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model $R^2$</th>
<th>Test $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGST</td>
<td>0.44</td>
<td>0.79</td>
</tr>
<tr>
<td>AGST, Harvest Rain</td>
<td>0.71</td>
<td>-0.08</td>
</tr>
<tr>
<td>AGST, Harvest Rain, Age</td>
<td>0.79</td>
<td>0.53</td>
</tr>
<tr>
<td>AGST, Harvest Rain, Age, Winter Rain</td>
<td>0.83</td>
<td>0.79</td>
</tr>
<tr>
<td>AGST, Harvest Rain, Age, Winter Rain, Population</td>
<td>0.83</td>
<td>0.76</td>
</tr>
</tbody>
</table>

- Better model $R^2$ does not necessarily mean better test set $R^2$
- Need more data to be conclusive
- Out-of-sample $R^2$ can be negative!
The Results

- **Parker:**
  - 1986 is “very good to sometimes exceptional”

- **Ashenfelter:**
  - 1986 is mediocre
  - 1989 will be “the wine of the century” and 1990 will be even better!

- In wine auctions,
  - 1989 sold for more than twice the price of 1986
  - 1990 sold for even higher prices!

- Later, Ashenfelter predicted 2000 and 2003 would be great
- Parker has stated that “2000 is the greatest vintage Bordeaux has ever produced”
The Analytics Edge

• A linear regression model with only a few variables can predict wine prices well

• In many cases, outperforms wine experts’ opinions

• A quantitative approach to a traditionally qualitative problem