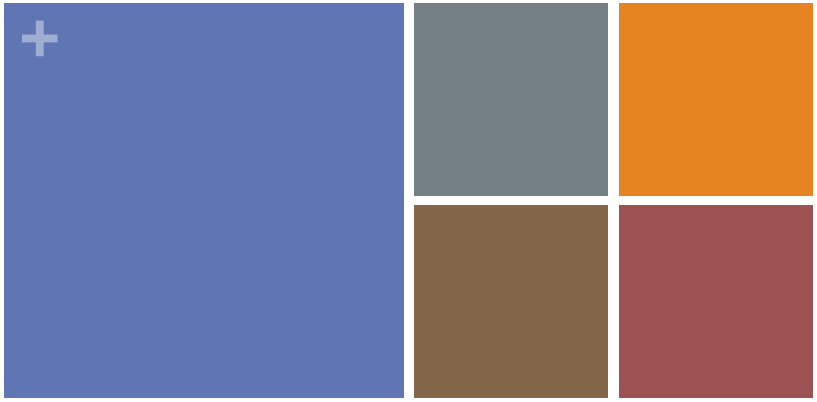


Lecture 2 – Predicting Wine Quality with Linear Regression II

IEOR 142 – Introduction to Machine Learning and Data Analytics
Fall 2018 – Paul Grigas

Announcements

- The first discussion lab will be held this **Friday from 3-4pm**
  - See bCourses announcement for more details
- Homework 1 will be released in a day or two
  - Due date of **Tuesday, September 11**

Today's Agenda

- A correction and more details about least squares estimation
- Model validation and R^2
- Significance, multicollinearity, and other issues
- An improved wine model with categorical variables (most likely next week)

Multiple Linear Regression Review

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

- Parametric method
- Observed data: $(x_i, y_i) \quad i = 1, \ldots, n$
- Each observed x_i is a feature vector: $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$
- Each observed y_i is a continuous response/dependent variable associated with x_i

Correction: Multiple Linear Regression Coefficient Estimates (Last time)

- Let $\mathbf{X}$ be the *n x p* matrix where the i^{th} row is the feature vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$
- Let $\mathbf{y}$ be the *n*-vector of responses y_i
- Then $\text{RSS}(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2$ and (if $\text{rank}(\mathbf{X}) = p$), then one may use calculus/linear algebra to show that the solution of $\min_{\beta \in \mathbb{R}^p} \text{RSS}(\beta)$ is given by:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Multiple Linear Regression Coefficient Estimates (Now the correct way)

- Let $\mathbf{X}$ be the *n x (p + 1)* matrix where the i^{th} row is the appended feature vector $(1, x_{i1}, x_{i2}, \ldots, x_{ip})$
- Let $\mathbf{y}$ be the *n*-vector of responses y_i
- Then the matrix vector product $\mathbf{X}\beta$ is the n-vector of training set predictions associated with the coefficient vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ and the n-vector of residuals is

$$r = \mathbf{y} - \mathbf{X}\beta$$

Multiple Linear Regression Coefficient Estimates, cont.

- Recall the 2-norm of an n-vector z is defined by:

$$\|z\|_2 = \sqrt{z_1^2 + z_2^2 + \ldots + z_n^2}$$

- Then it is easy to see that:

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 \\ &= \|\mathbf{y} - \mathbf{X}\beta\|_2^2\end{aligned}$$

Multiple Linear Regression Coefficient Estimates, cont.

- Using the representation $\text{RSS}(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2$ and assuming that $\text{rank}(\mathbf{X}) = p + 1 < n$, then one may use calculus/linear algebra to show that the solution of $\min_{\beta \in \mathbb{R}^p} \text{RSS}(\beta)$ is given by:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- (Aside on the board in the case of simple linear regression…)

Significance Testing, Multicollinearity, and Other Issues

Multiple Linear Regression Review

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

- Parametric method
- Observed data: $(x_i, y_i) \quad i = 1, \ldots, n$
- Each observed x_i is a feature vector: $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$
- Each observed y_i is a continuous response/dependent variable associated with x_i

Some Important Questions

- Do all of the predictors help to explain the response? Which variables are "significant"?
- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response Y?

Testing the Significance of Regression Coefficients

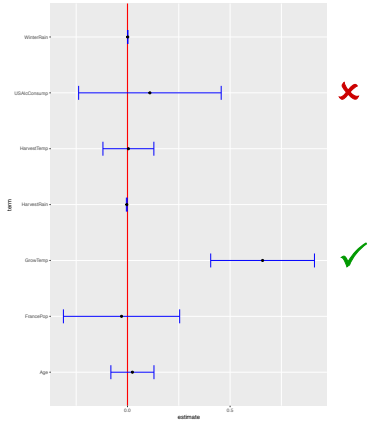
- Is the independent variable X_j useful in predicting the response Y?
  - Does US Alcohol Consumption help to predict log(price index)?
- In other words, is $\beta_j \neq 0$?
- This is an inference question, and can be addressed with a hypothesis test:

$$H_0 : \beta_j = 0 \text{ vs. } H_a : \beta_j \neq 0$$

Testing the Significance of Regression Coefficients

$H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$

- Hypothesis test is equivalent to looking at confidence intervals
- Reject null hypothesis as significance level α if and only if (1-α)% confidence interval does not contain 0



Interlude on "Standard Assumptions" for Linear Regression

A Useful Set of Conceptual Assumptions

- Question: Where do the previous confidence intervals come from?
- Answer: Some of the statistical analysis associated with linear regression is derived from a certain set of assumptions regarding how the data is generated

A Useful Set of Conceptual Assumptions

- 1.) The observed data $(x_i, y_i) \quad i = 1, \ldots, n$ satisfies

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ are the true but unknown regression coefficients and the ϵ_i are noise terms

- 2.) $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent and identically distributed **normal** random variables with mean 0 and variance σ^2
- 3.) If the features $x_1, x_2, \ldots, x_n$ are also regarded as random variables, then they are independent of $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$

Consequences of the assumptions

- Under the previous set of assumptions, it is possible to prove mathematically that:
- 1.) $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is an unbiased estimator of the true vector of coefficients β:

$$\mathbb{E}\left[\hat{\beta} \mid \mathbf{X}\right] = \beta$$

- 2.) The covariance matrix of $\hat{\beta}$ given $\mathbf{X}$ is:

$$\text{cov}(\hat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

- 3.) $\hat{\beta}$ is a normally distributed random vector given $\mathbf{X}$

Constructing a confidence interval

- Given the formula $\text{cov}(\hat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, we can read off the diagonal entries of this matrix to get the standard errors for each coefficient
- Given that $\hat{\beta}$ is normally distributed, we can now easily construct confidence intervals in the usual way, i.e., for some z-score (such as $z^* = 1.96$):

$$\hat{\beta}_j \pm z^* \sqrt{\text{cov}(\hat{\beta} \mid \mathbf{X})_{jj}}$$

- Question: What's the problem?

Constructing a confidence interval

- Question: What's the problem?
- Answer: we usually don't know σ^2 and must estimate that from the data in order to construct the matrix $\text{cov}(\hat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
- Letting $e = \mathbf{y} - \mathbf{X}\hat{\beta}$ denote the vector of training set residuals, then use the estimate:

$$\hat{\sigma}^2 = \frac{\|e\|_2^2}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2$$

Take home message of this interlude

- It is important to understand the assumptions that lead to the results of your analysis (e.g., which variables you retain in your model)
- Ultimately though – regardless of whether you believe or doubt that the assumptions hold for your dataset – it is critical to validate your final model on an out of sample testing set

Back to Significance Testing and Other Issues

Testing the Significance of Regression Coefficients in R

- R shows stars * (literally!) for the significant coefficients
- The more stars, the more significant. To be significant at the 5% level (95% confidence interval), the coefficient must have at least one *
- p-value (Pr(>|t|)) is the boundary point where we switch from significant to not significant (essentially smallest α such that significant at level α)
- Smaller p-values are better

Testing the Significance of Regression Coefficients in R

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.9662699  9.3823951  -0.529  0.60166
WinterRain    0.0011863  0.0005628   2.108  0.04616 *
HarvestRain  -0.0033137  0.0010650  -3.112  0.00491 **
GrowTemp      0.6582753  0.1221937   5.387 1.79e-05 ***
HarvestTemp   0.0044212  0.0599935   0.074  0.94189
Age           0.0240080  0.0507587   0.473  0.64068
FrancePop    -0.0290258  0.1369627  -0.212  0.83403
USAlcConsump  0.1092561  0.1678945   0.651  0.52166
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3307 on 23 degrees of freedom
Multiple R-squared:  0.7894, Adjusted R-squared:  0.7253
F-statistic: 12.31 on 7 and 23 DF,  p-value: 1.859e-06
```

- Are there coefficients that you are not comfortable with?
- Let's return to this question in a moment

Testing the Significant of the Entire Model

- A more basic question: is the model worth anything at all?
- Frame this question as a hypothesis test:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{vs.} \quad H_a : \text{ at least one } \beta_j \neq 0$$

- R reports the F-statistic and corresponding p-value
  - Again, small p-value is good!
  - Why is this not the same as checking the p-value of each coefficient?

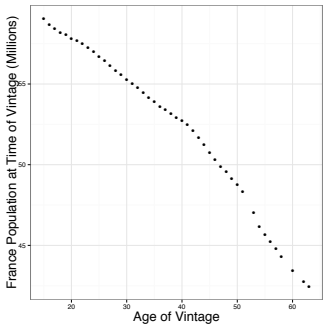
Testing the Significant of the Entire Model


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.9662699  9.3823951  -0.529  0.60166
WinterRain    0.0011863  0.0005628   2.108  0.04616 *
HarvestRain  -0.0033137  0.0010650  -3.112  0.00491 **
GrowTemp      0.6582753  0.1221937   5.387 1.79e-05 ***
HarvestTemp   0.0044212  0.0599935   0.074  0.94189
Age           0.0240080  0.0507587   0.473  0.64068
FrancePop    -0.0290258  0.1369627  -0.212  0.83403
USAlcConsump  0.1092561  0.1678945   0.651  0.52166
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3307 on 23 degrees of freedom
Multiple R-squared:  0.7894,Adjusted R-squared:  0.7253
F-statistic: 12.31 on 7 and 23 DF,  p-value: 1.859e-06
```

- Are there coefficients that you are not comfortable with?
- Why might the last four coefficients not be significant?

Plot of Age versus France Population

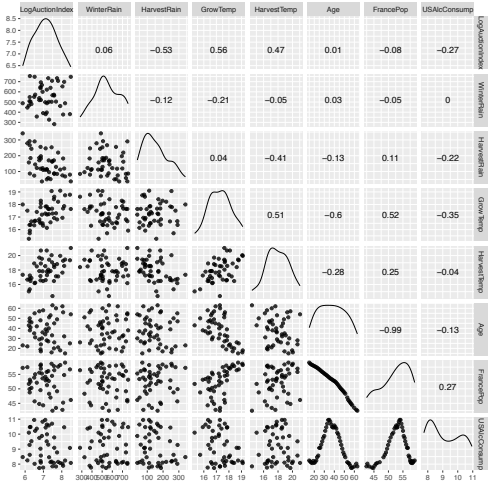


- The data for Age and France population are highly correlated
- This is evidence of **multicollinearity**

Multicollinearity

- Occurs when two or more predictors are highly correlated
- Makes the estimated coefficients $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ very sensitive to noise in the training data
  - Thus can produce very inaccurate estimates which hurts interpretability and possibly predictive performance
- Tell-tale signs:
  - Some of the estimated coefficients have the "wrong" sign
  - Some of the coefficients are not significantly different from zero
- Multicollinearity can usually be fixed by deleting one or more independent variables

Correlation Table



Multicollinearity

- Multicollinearity can exist without evidence of large correlations in the correlation table
- Better to check the **VIFs (variance inflation factors):**

```
WinterRain HarvestRain GrowTemp HarvestTemp Age       FrancePop USAlcConsump
1.295370   1.578682    1.700079 2.198191    66.936256 81.792302 10.441217
```

  - Rule of thumb:
    - VIF > 10: definitely a problem
    - VIF > 5: could be a problem
    - VIF <= 5: probably okay

What is VIF?

- Consider regressing each predictor variable X_j on all of the others:

$$X_j = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_{j-1} X_{j-1} + \alpha_{j+1} X_{j+1} + \ldots + \alpha_p X_p$$

- If the R^2 for the above (call it R_j^2) is equal to 1, then there exists a perfect linear relationship between X_j and all other independent variables (at least according to the training data)
- So, define: $\text{VIF}_j = \dfrac{1}{1 - R_j^2}$

How do we deal with multicollinearity?

- One approach:
  - Remove a variable with high VIF, but if there is a "tie" then keep the variables that you "like"
  - Iterate this procedure
- This issue falls under the realm of **model selection** – the process of finding the best model
- Model selection is still somewhat of an art but we will see some principled approaches later in the course

VIF Values for the Wine Model


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.9662699  9.3823951  -0.529  0.60166
WinterRain    0.0011863  0.0005628   2.108  0.04616 *
HarvestRain  -0.0033137  0.0010650  -3.112  0.00491 **
GrowTemp      0.6582753  0.1221937   5.387 1.79e-05 ***
HarvestTemp   0.0044212  0.0599935   0.074  0.94189
Age           0.0240080  0.0507587   0.473  0.64068
FrancePop    -0.0290258  0.1369627  -0.212  0.83403
USAlcConsump  0.1092561  0.1678945   0.651  0.52166
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 0.3307 on 23 degrees of freedom
Multiple R-squared:  0.7894,Adjusted R-squared:  0.7253
F-statistic: 12.31 on 7 and 23 DF,  p-value: 1.859e-06
```

Coefficient	VIF
WinterRain	1.30
HarvestRain	1.58
GrowTemp	1.70
HarvestTemp	2.20
Age	66.94
FrancePop	81.79
USAlcConsump	10.44

Building our Better Wine Model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.8404548  3.0706463  -2.228  0.03553 *
WinterRain    0.0012145  0.0005359   2.266  0.03274 *
HarvestRain  -0.0033611  0.0010203  -3.294  0.00305 **
GrowTemp      0.6671389  0.1125053   5.930 4.05e-06 ***
HarvestTemp   0.0020543  0.0577600   0.036  0.97192
Age           0.0340519  0.0178084   1.912  0.06787 .
USAlcConsump  0.0933334  0.1471271   0.634  0.53184
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.3241 on 24 degrees of freedom
Multiple R-squared:  0.789, Adjusted R-squared:  0.7362
F-statistic: 14.95 on 6 and 24 DF,  p-value: 4.604e-07
```

Coefficient	VIF
WinterRain	1.22
HarvestRain	1.51
GrowTemp	1.50
HarvestTemp	2.12
Age	8.58
USAlcConsump	8.35

Building our Better Wine Model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.215161   1.672215  -3.119 0.004532 **
WinterRain   0.001119   0.000508   2.202 0.037112 *
HarvestRain -0.003437   0.001001  -3.433 0.002089 **
GrowTemp     0.664336   0.111067   5.981 3.02e-06 ***
HarvestTemp -0.006650   0.055432  -0.120 0.905462
Age          0.023466   0.006143   3.820 0.000785 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.3202 on 25 degrees of freedom
Multiple R-squared:  0.7854, Adjusted R-squared:  0.7425
F-statistic:  18.3 on 5 and 25 DF,  p-value: 1.213e-07
```

Coefficient	VIF
WinterRain	1.13
HarvestRain	1.49
GrowTemp	1.50
HarvestTemp	2.00
Age	1.04

Building our Better Wine Model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.2163945  1.6401825  -3.180 0.003782 **
WinterRain   0.0011116  0.0004949   2.246 0.033424 *
HarvestRain -0.0033766  0.0008504  -3.971 0.000505 ***
GrowTemp     0.6569271  0.0905520   7.255 1.05e-07 ***
Age          0.0235571  0.0059785   3.940 0.000546 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.3141 on 26 degrees of freedom
Multiple R-squared:  0.7853, Adjusted R-squared:  0.7523
F-statistic: 23.78 on 4 and 26 DF,  p-value: 2.307e-08
```

Coefficient	VIF
WinterRain	1.11
HarvestRain	1.12
GrowTemp	1.04
Age	1.03

A Very Good Wine Model


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.2163945  1.6401825  -3.180 0.003782 **
WinterRain   0.0011116  0.0004949   2.246 0.033424 *
HarvestRain -0.0033766  0.0008504  -3.971 0.000505 ***
GrowTemp     0.6569271  0.0905520   7.255 1.05e-07 ***
Age          0.0235571  0.0059785   3.940 0.000546 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3141 on 26 degrees of freedom
Multiple R-squared:  0.7853, Adjusted R-squared:  0.7523
F-statistic: 23.78 on 4 and 26 DF,  p-value: 2.307e-08
```

- $R^2 = 0.79$ (previously 0.79)
- All coefficients are significantly different than zero
- $OSR^2 = 0.75$
- This model is not really different from Orley Ashenfelter's model

Other Potential Fit Problems

- Nonlinear dependence on the features
  - We will discuss this later in the course
- Correlation of residuals
- Non-constant variance of residuals
- Outliers/high-leverage points
- The last three are not a major concern in this course, but it's always healthy to plot your data, including residual plots
- See James Section 3.3.3 for more details

A Better Wine Model Using Categorical Variables

Towards an Even Better Model

- The previous model predicts a price index
- $OSR^2 = 0.75$, pretty good, but not really "great"
- It would be better if we could predict the actual price for a given winery – then we could use the model in direct support of the auction
- This is the "big data" era, yet we have only one price index for each year back to 1952
- Even if we were to look at an individual winery, we would still only have a few dozen data records
- Wouldn't it be great if we could use the separate data from all wineries in all years? Then we could take advantage of more data.
- Let's see how we can do this

Map of Bordeaux Region

All-Wineries Data

	LogAuction	Winery	Age	WinterRain	HarvestRain	GrowTemp	HarvestTemp	FrancePop	USAlcConsump
1 1952	6.653108	Cheval Blanc	63	566.4	165.5	17.28	14.39	42.46	7.85
2 1952	6.861502	Lafite-Rothschild	63	566.4	165.5	17.28	14.39	42.46	7.85
3 1953	6.664192	Cheval Blanc	62	653.3	75.6	16.94	17.64	42.75	8.03
4 1955	6.311426	Cheval Blanc	60	504.3	129.5	17.30	17.13	43.43	7.84
5 1955	6.550209	Lafite-Rothschild	60	504.3	129.5	17.30	17.13	43.43	7.84
6 1959	5.380957	Beychevelle	56	377.0	182.6	17.68	19.28	45.24	7.89
7 1959	7.437242	Cheval Blanc	56	377.0	182.6	17.68	19.28	45.24	7.89
8 1959	7.645302	Lafite-Rothschild	56	377.0	182.6	17.68	19.28	45.24	7.89
9 1960	6.405873	Lafite-Rothschild	55	748.2	290.6	16.67	16.18	45.68	8.02
10 1961	5.813802	Beychevelle	54	747.8	37.7	17.64	21.05	46.16	8.08
11 1961	7.311178	Cheval Blanc	54	747.8	37.7	17.64	21.05	46.16	8.08
12 1961	5.822247	Cos d'Estournel	54	747.8	37.7	17.64	21.05	46.16	8.08
13 1961	6.673045	Lafite-Rothschild	54	747.8	37.7	17.64	21.05	46.16	8.08
14 1962	6.747610	Cheval Blanc	53	639.4	51.8	16.58	17.86	47.00	8.13
15 1962	5.416100	Cos d'Estournel	53	639.4	51.8	16.58	17.86	47.00	8.13
16 1962	6.298839	Lafite-Rothschild	53	639.4	51.8	16.58	17.86	47.00	8.13
17 1964	4.354270	Beychevelle	51	326.5	96.1	17.63	19.43	48.31	8.46
18 1964	6.492785	Cheval Blanc	51	326.5	96.1	17.63	19.43	48.31	8.46
19 1964	6.011610	Lafite-Rothschild	51	326.5	96.1	17.63	19.43	48.31	8.46
20 1966	5.957908	Cheval Blanc	49	734.0	85.2	16.81	18.82	49.16	8.78
...	...	...	...	...	...	...	...	...	...
147 2000	7.060588	Lafite-Rothschild	15	487.8	69.0	18.73	19.45	59.05	8.24

Training and Testing Set (Split by Year)

Full Dataset (N = 147)

	LogAuction	Winery	...	USAlcConsump
1 1952	6.653108	Cheval Blanc	...	7.85
2 1952	6.861502	Lafite-Rothschild	...	7.85
3 1953	6.664192	Cheval Blanc	...	8.03
4 1955	6.311426	Cheval Blanc	...	7.84
5 1955	6.550209	Lafite-Rothschild	...	7.84
6 1959	5.380957	Beychevelle	...	7.89
7 1959	7.437242	Cheval Blanc	...	7.89
8 1959	7.645302	Lafite-Rothschild	...	7.89
9 1960	6.405873	Lafite-Rothschild	...	8.02
10 1961	5.813802	Beychevelle	...	8.08
...	...	...	...	...
81 1985	6.018934	Cheval Blanc	...	9.88
82 1985	4.885072	Cos d'Estournel	...	9.88
83 1985	6.296612	Lafite-Rothschild	...	9.88
84 1986	4.549235	Beychevelle	...	9.75
85 1986	5.721852	Cheval Blanc	...	9.75
86 1986	4.987435	Cos d'Estournel	...	9.75
...	...	...	...	...
146 2000	4.330339	Giscours	...	8.24
147 2000	7.060588	Lafite-Rothschild	...	8.24

Training Set (N = 83)

	LogAuction	Winery	...	USAlcConsump
1 1952	6.653108	Cheval Blanc	...	7.85
2 1952	6.861502	Lafite-Rothschild	...	7.85
3 1953	6.664192	Cheval Blanc	...	8.03
4 1955	6.311426	Cheval Blanc	...	7.84
5 1955	6.550209	Lafite-Rothschild	...	7.84
6 1959	5.380957	Beychevelle	...	7.89
7 1959	7.437242	Cheval Blanc	...	7.89
8 1959	7.645302	Lafite-Rothschild	...	7.89
9 1960	6.405873	Lafite-Rothschild	...	8.02
10 1961	5.813802	Beychevelle	...	8.08
...	...	...	...	...
81 1985	6.018934	Cheval Blanc	...	9.88
82 1985	4.885072	Cos d'Estournel	...	9.88
83 1985	6.296612	Lafite-Rothschild	...	9.88

Testing Set (N = 64)

	LogAuction	Winery	...	USAlcConsump
84 1986	4.549235	Beychevelle	...	9.75
85 1986	5.721852	Cheval Blanc	...	9.75
86 1986	4.987435	Cos d'Estournel	...	9.75
...	...	...	...	...
146 2000	4.330339	Giscours	...	8.24
147 2000	7.060588	Lafite-Rothschild	...	8.24

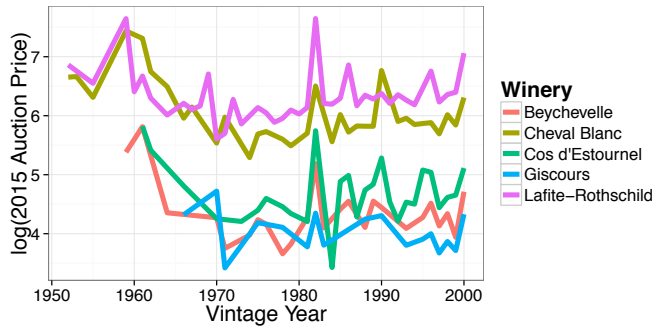
Regression Model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2184029  3.4314946  -0.064   0.9494
Age          0.0528857  0.0118313   4.470 2.62e-05 ***
WinterRain   0.0018288  0.0009861   1.855   0.0674 .
HarvestRain  0.0024980  0.0019812   1.261   0.2111
GrowTemp     0.1209556  0.1926703   0.628   0.5320
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9639 on 78 degrees of freedom
Multiple R-squared:  0.2221,           Adjusted R-squared:  0.1822
F-statistic: 5.567 on 4 and 78 DF,  p-value: 0.0005389
```

- Why is the R^2 low?

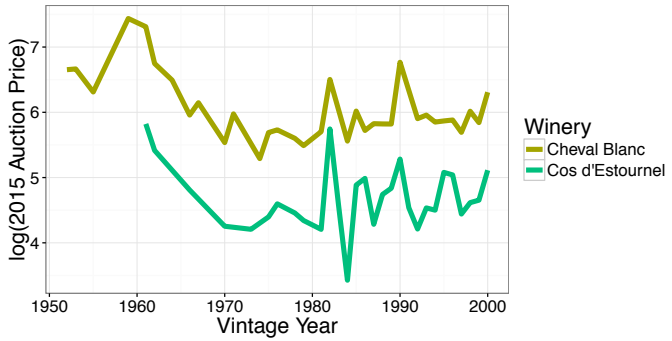
2015 Auction Prices by Winery and Year



Categorical Variables With Two Levels

- For illustration, we will only look at data of two wineries:
  - **Cheval Blanc** – one of the most expensive wineries
  - **Cos d'Estournel** – one of least expensive wineries
- Two categories corresponds to adding one categorical variable
- We will call our variable `WineryCosd'Estournel`
  - Value 1: Wine is from Cos d'Estournel
  - Value 0: Wine is from Cheval Blanc

2015 Auction Prices for Two Wineries



Categorical Variables with Two Levels

		LogAuction	WineryCos d'Estournel	Age	WinterRain	HarvestRain	GrowTemp
1	1952	6.653108	0	63	566.4	165.5	17.28
2	1953	6.664192	0	62	653.3	75.6	16.94
3	1955	6.311426	0	60	504.3	129.5	17.30
4	1959	7.437242	0	56	377.0	182.6	17.68
5	1961	7.311178	0	54	747.8	37.7	17.64
6	1961	5.822247	1	54	747.8	37.7	17.64
7	1962	6.747610	0	53	639.4	51.8	16.58
8	1962	5.416100	1	53	639.4	51.8	16.58
9	1964	6.492785	0	51	326.5	96.1	17.63
10	1966	5.957908	0	49	734.0	85.2	16.81
11	1966	4.809416	1	49	734.0	85.2	16.81
12	1967	6.146929	0	48	646.9	118.1	16.51
13	1970	5.536231	0	45	563.5	88.8	16.92
14	1970	4.254051	1	45	563.5	88.8	16.92
15	1971	5.975843	0	44	488.4	111.9	17.20
16	1973	4.207376	1	42	357.2	122.6	17.41
17	1974	5.290386	0	41	503.6	185.1	16.39
18	1975	5.689684	0	40	501.8	170.5	17.23
19	1975	4.397162	1	40	501.8	170.5	17.23
	...	...	...	...	...	...	...
35	1985	4.885072	1	30	667.1	37.2	17.19

A Two-Category Model

$$\begin{aligned}\text{Auction Price} &= \beta_0 \\ &+ \beta_1 \cdot \text{WineryCos d'Estournel} \\ &+ \beta_2 \cdot \text{Age} \\ &+ \beta_3 \cdot \text{WinterRain} \\ &+ \beta_4 \cdot \text{HarvestRain} \\ &+ \beta_5 \cdot \text{GrowTemp}\end{aligned}$$

- What is the interpretation?

The Two-Category Model

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)            -6.6292230  2.3910006  -2.773  0.00962 **
WineryCos d'Estournel -1.3616758  0.1393778  -9.770 1.12e-10 ***
Age                    0.0357669  0.0072019   4.966 2.79e-05 ***
WinterRain             0.0016274  0.0006888   2.363  0.02506 *
HarvestRain           -0.0015879  0.0015803  -1.005  0.32330
GrowTemp               0.6093432  0.1397853   4.359  0.00015 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3874 on 29 degrees of freedom
Multiple R-squared:  0.8598,       Adjusted R-squared:  0.8356
F-statistic: 35.55 on 5 and 29 DF,  p-value: 1.624e-11
```

- Note that $R^2 = 0.86$
- All variables are significant except HarvestRain

Let's go back to the All-Wineries Data

	LogAuction	Winery	Age	WinterRain	HarvestRain	GrowTemp	HarvestTemp	FrancePop	USAlcConsump
1	1952	6.653108	Cheval Blanc	63	566.4	165.5	17.28	14.39	42.46	7.85
2	1952	6.861502	Lafite-Rothschild	63	566.4	165.5	17.28	14.39	42.46	7.85
3	1953	6.664192	Cheval Blanc	62	653.3	75.6	16.94	17.64	42.75	8.03
4	1955	6.311426	Cheval Blanc	60	504.3	129.5	17.30	17.13	43.43	7.84
5	1955	6.550209	Lafite-Rothschild	60	504.3	129.5	17.30	17.13	43.43	7.84
6	1959	5.380957	Beychevelle	56	377.0	182.6	17.68	19.28	45.24	7.89
7	1959	7.437242	Cheval Blanc	56	377.0	182.6	17.68	19.28	45.24	7.89
8	1959	7.645302	Lafite-Rothschild	56	377.0	182.6	17.68	19.28	45.24	7.89
9	1960	6.405873	Lafite-Rothschild	55	748.2	290.6	16.67	16.18	45.68	8.02
10	1961	5.813802	Beychevelle	54	747.8	37.7	17.64	21.05	46.16	8.08
11	1961	7.311178	Cheval Blanc	54	747.8	37.7	17.64	21.05	46.16	8.08
12	1961	5.822247	Cos d'Estournel	54	747.8	37.7	17.64	21.05	46.16	8.08
13	1961	6.673045	Lafite-Rothschild	54	747.8	37.7	17.64	21.05	46.16	8.08
14	1962	6.747610	Cheval Blanc	53	639.4	51.8	16.58	17.86	47.00	8.13
15	1962	5.416100	Cos d'Estournel	53	639.4	51.8	16.58	17.86	47.00	8.13
16	1962	6.298839	Lafite-Rothschild	53	639.4	51.8	16.58	17.86	47.00	8.13
17	1964	4.354270	Beychevelle	51	326.5	96.1	17.63	19.43	48.31	8.46
18	1964	6.492785	Cheval Blanc	51	326.5	96.1	17.63	19.43	48.31	8.46
19	1964	6.011610	Lafite-Rothschild	51	326.5	96.1	17.63	19.43	48.31	8.46
20	1966	5.957908	Cheval Blanc	49	734.0	85.2	16.81	18.82	49.16	8.78
...	...	...	...	...	...	...	...	...	...	...
83	1985	6.296612	Lafite-Rothschild	30	667.1	37.2	17.19	19.56	55.28	9.88

Categorical Variables with More Than Two Levels

- We need k-1 dummy variables to work with k categories (why?)
- Variable WineryCheval Blanc: 1 if from Cheval Blanc, otherwise 0
- Variable WineryCosd'Estournel: 1 if from Cos d'Estournel, otherwise 0
- Variable WineryGiscours: 1 if from Giscours, otherwise 0
- Variable WineryLafite-Rothschild: 1 if from Lafite-Rothschild, otherwise 0
- All variables 0 if wine from Beychevelle

A Model with More Than Two Categories

$$\begin{aligned}\text{Auction Price} = {} & \beta_0 \\ & + \beta_1 \cdot \text{WineryCheval Blanc} \\ & + \beta_2 \cdot \text{WineryCos d'Estournel} \\ & + \beta_3 \cdot \text{WineryGiscours} \\ & + \beta_4 \cdot \text{WineryLafite} - \text{Rothschild} \\ & + \beta_5 \cdot \text{Age} \\ & + \beta_6 \cdot \text{WinterRain} \\ & + \beta_7 \cdot \text{HarvestRain} \\ & + \beta_8 \cdot \text{GrowTemp}\end{aligned}$$

Model with Categorical Data for Five Wineries

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -4.4945857  1.5757380  -2.852 0.005622 **
WineryCheval Blanc       1.6425245  0.1518157  10.819  < 2e-16 ***
WineryCos d'Estournel    0.2754099  0.1649803   1.669 0.099274 .
WineryGiscours          -0.2992903  0.1934825  -1.547 0.126163
WineryLafite-Rothschild  1.8941459  0.1481200  12.788  < 2e-16 ***
Age                      0.0307904  0.0054819   5.617 3.23e-07 ***
WinterRain               0.0016349  0.0004462   3.665 0.000463 ***
HarvestRain              0.0003949  0.0009050   0.436 0.663899
GrowTemp                 0.3875778  0.0886121   4.374 3.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4336 on 74 degrees of freedom
Multiple R-squared:  0.8506,        Adjusted R-squared:  0.8345
F-statistic: 52.68 on 8 and 74 DF,  p-value: < 2.2e-16
```

- $R^2 = 0.85$ is excellent
- $OSR^2 = 0.81$ is also excellent

Prediction for Cos d’Estournel

- Consider making a prediction for the 2014 vintage
  - Cos d’Estournel winery
  - Aged 1 years in 2015
  - 522.3 mm of winter rain
  - 78.9 mm of harvest rain
  - Average growing season temperature of 18.23 °C

LogAuctionPrice = -4.495 + 1.643*(0) + 0.2754*(1) – 0.299*(0) + 1.894*(0)

+ 0.031*(1) + 0.002*(522.3) + … + 0.388*(18.23)

= 3.762

Prediction for Giscours

- Consider making a prediction for the 2014 vintage
  - Giscours winery
  - Aged 1 years in 2015
  - 522.3 mm of winter rain
  - 78.9 mm of harvest rain
  - Average growing season temperature of 18.23 °C

LogAuctionPrice = -4.495 + 1.643*(0) + 0.2754*(0) – 0.299*(1) + 1.894*(0)

+ 0.031*(1) + 0.002*(522.3) + … + 0.388*(18.23)

= 3.188

Prediction for Beychevelle

- Consider making a prediction for the 2014 vintage
  - Beychevelle winery
  - Aged 1 years in 2015
  - 522.3 mm of winter rain
  - 78.9 mm of harvest rain
  - Average growing season temperature of 18.23 °C

LogAuctionPrice = -4.495 + 1.643*(0) + 0.2754*(0) – 0.299*(0) + 1.894*(0)

+ 0.031*(1) + 0.002*(522.3) + … + 0.388*(18.23)

= 3.487

Showdown in *The New York Times*

- Showdown on the front page of *The New York Times* in 1990
- **Parker**: 1986 vintage will be "very good to sometimes exceptional"
- **Ashenfelter**: 1986 vintage will be mediocre, but 1989 vintage will be "stunningly good"
- Experts like Parker hadn't even had a chance to taste the 1989 vintage

Wine Equation Puts Some Noses Out of Joint: Wine Equation Has Noses Out of Joint
By PETER PASSELL;
New York Times (1923-Current file); Mar 4, 1990; ProQuest Historical Newspapers: The New York Times
pg. 1

Wine Equation Puts Some Noses Out of Joint

By PETER PASSELL

Calculate the winter rain and the harvest rain (in millimeters). Add summer heat in the vineyard (in degrees centigrade). Subtract 12.145. And what do you have? A very, very passionate argument over wine.

Prof. Orley Ashenfelter, a Princeton economist, has devised a mathematical formula for predicting the quality of red wine vintages in France. And the guardians of tradition are fuming.

Robert M. Parker Jr., generally regarded as the most influential wine critic in America, calls Professor Ashenfelter's research "ludicrous and absurd."

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

These reactions are not surprising.

No formula could capture the variations in estate-bottled wines from the same vintage. But if widely accepted, Professor Ashenfelter's approach to judging the quality of the red wines of Bordeaux and Burgundy long before they are drinkable could turn the clubby, exclusive world of wine rating on its ear.

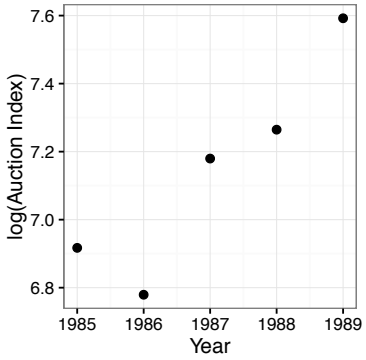
Also for the Human Element

And even if it does not, enthusiasts who care more about drinking wine than talking about it could use what seem to be startlingly accurate forecasts to gain an edge at the expense of wine brokers and dealers.

For the moment, though, Professor Ashenfelter is too distracted to ponder the cosmic implications of removing

Prof. Orley Ashenfelter, whose mathematical formula for predicting the quality of French red wine vintages has left traditionalists fuming.

Years Later, the Winner is Clear



A Convergence of Results

Though most critics never acknowledged the value of Ashenfelter's models, through time the predictions of the models and experts has converged.

Ashenfelter:

> **"Unlike the past, the tasters no longer make any horrendous mistakes. Frankly, I kind of killed myself. I don't have much value added anymore."**

Conclusion

- A linear regression model with only a few variables can predict wine prices well
- In many cases, the model outperforms wine experts' judgments
- A quantitative approach to a traditionally qualitative problem
- Regression capabilities are enhanced with a user who knows how to think with data and learn from the data

- Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

- Thanks to Rob Freund and John Silberholz (MIT) for the wine datasets