DIMENSIONLESS
TECHNOLOGY

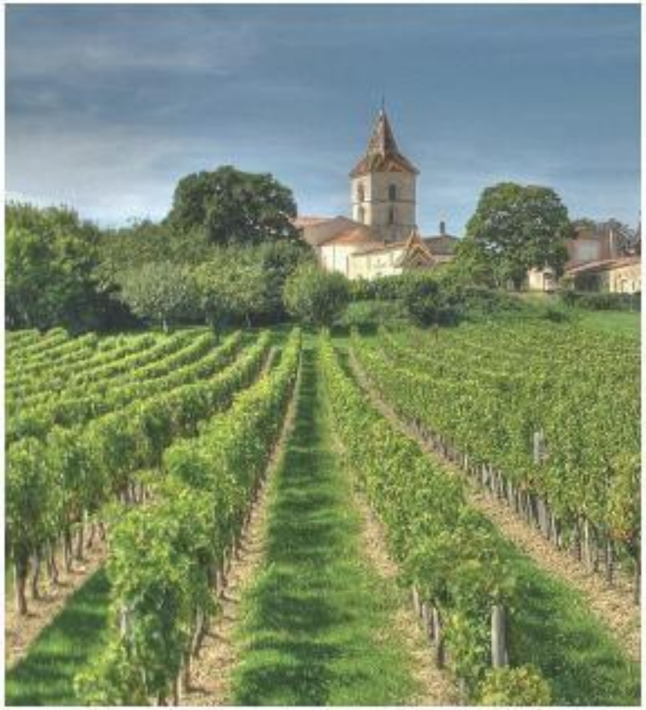
Linear Regression

Linear Regression

Bordeaux, France

Bordeaux Wine

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

https://storage.googleapis.com/dimensionless/Analytics/wine_test.csv

https://storage.googleapis.com/dimensionless/Analytics/wine.csv

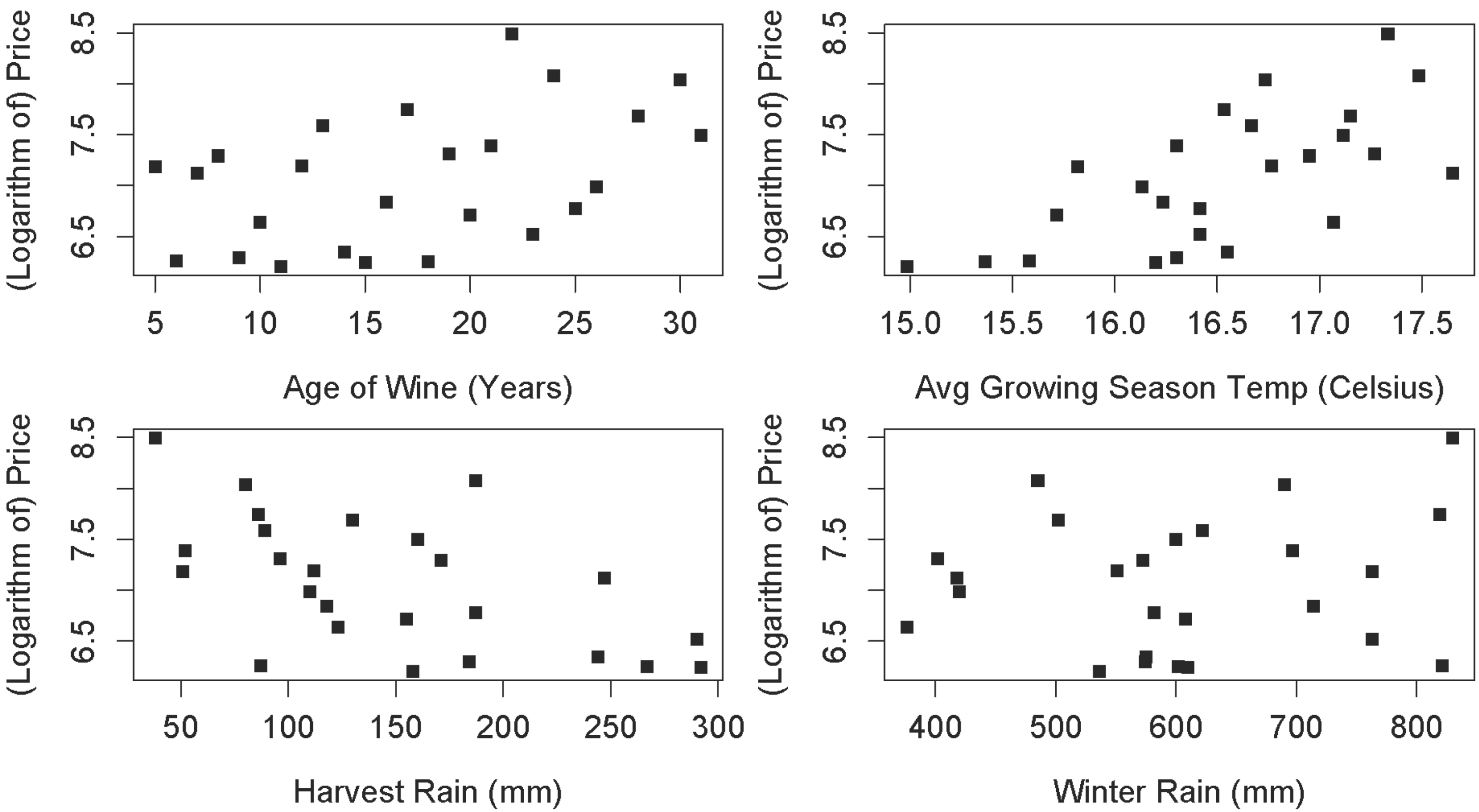
Building the Model

- 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(1952-78)

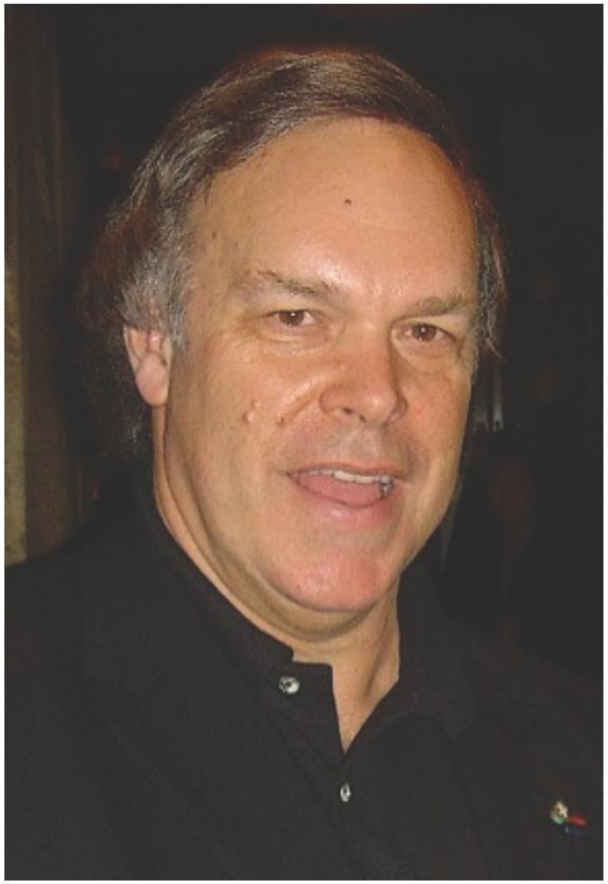


Expert Reaction

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

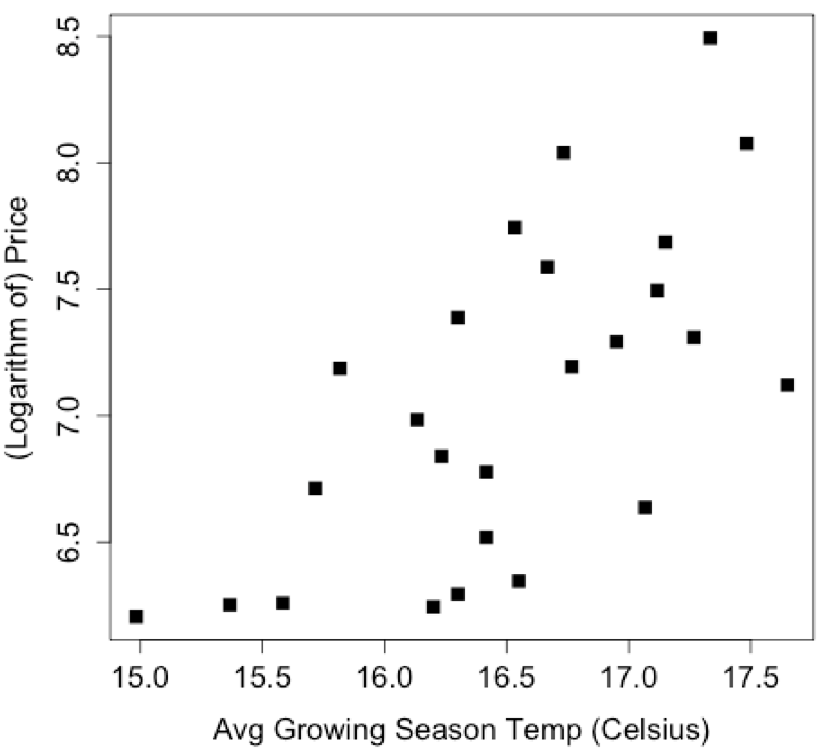
"Ashenfelter is an absolute total sham"

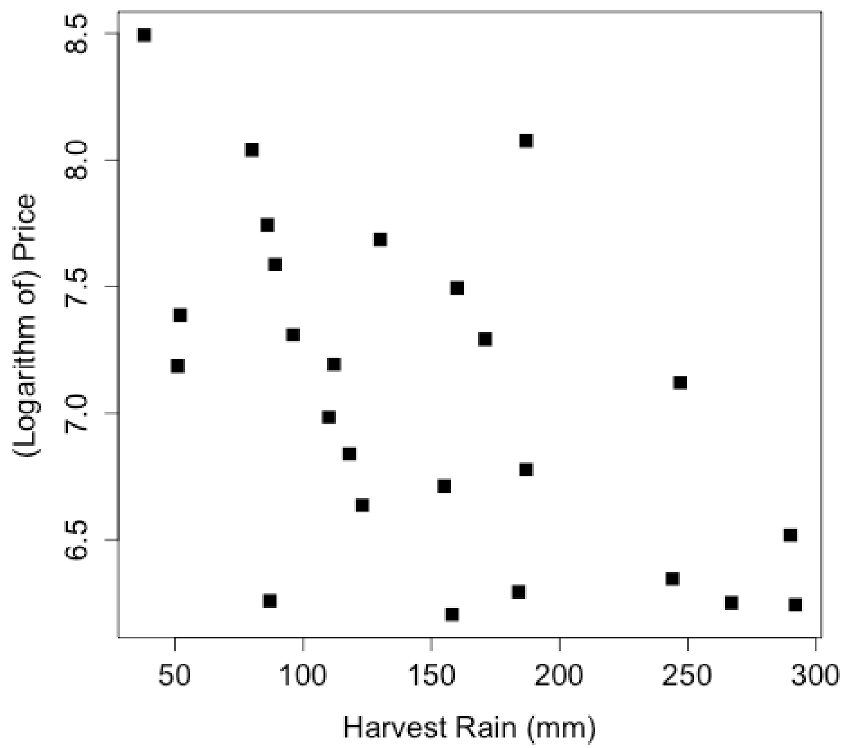
"rather like a movie critic who never goes to see the movie but tells you how good it is based on the actors and the director"



Quick Question

The plots below show the relationship between two of the independent variables considered by Ashenfelter and the price of wine.

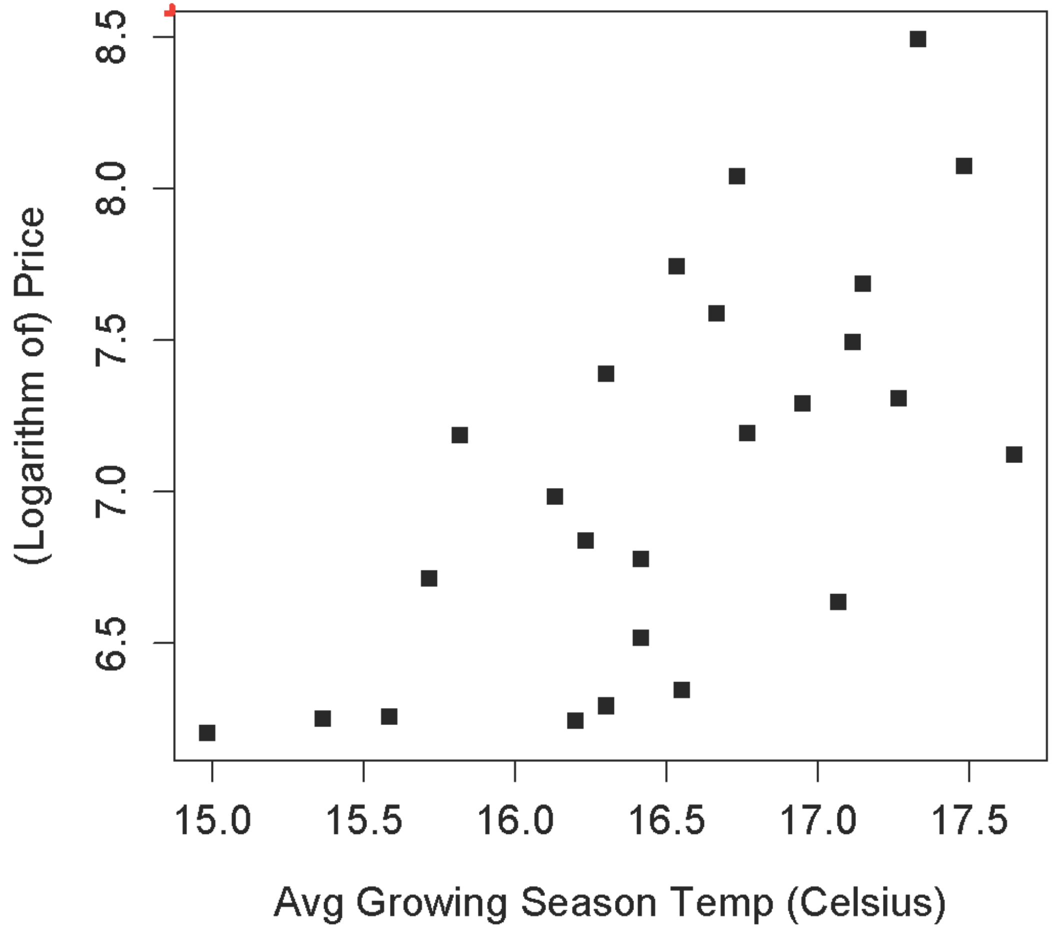




What is the correct relationship between harvest rain, average growing season temperature, and wine prices?

- More harvest rain is associated with a higher price, and higher temperatures is associated with a higher price
- More harvest rain is associated with a higher price, and higher temperatures is associated with a lower price
- More harvest rain is associated with a lower price, and higher temperatures is associated with a higher price
- More harvest rain is associated with a lower price, and higher temperatures is associated with a lower price

One Variable Linear Regression



The Regression Model

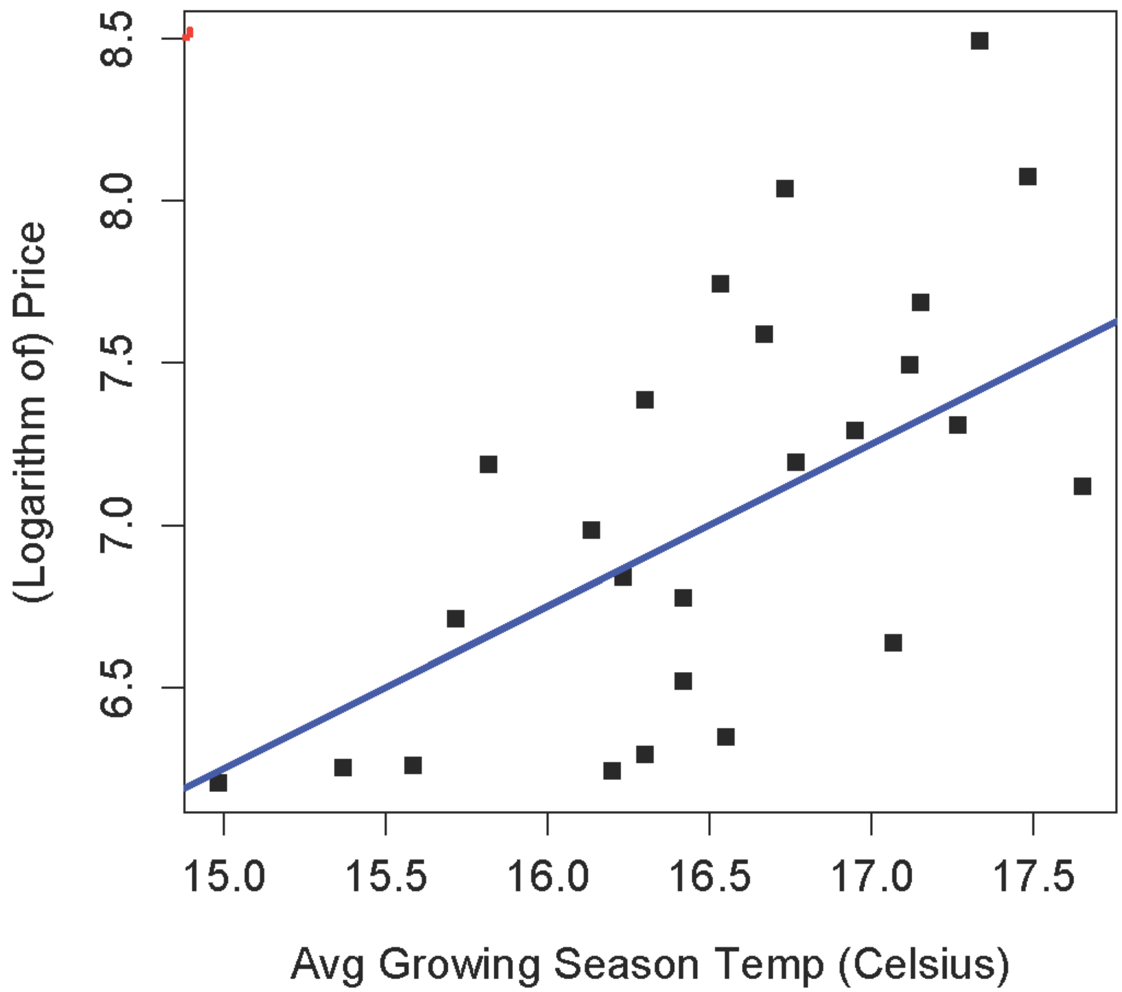
- One-variable regression model

$$y^i = \beta_0 + \beta_1 x^i + \epsilon^i$$

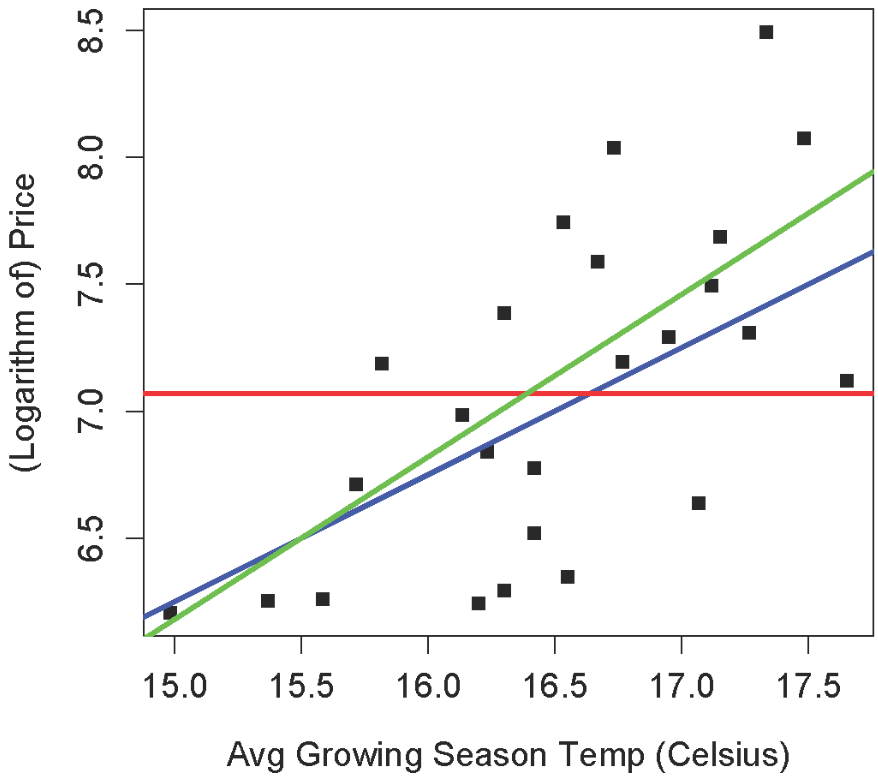
y^i = dependent variable (wine price) for the i^{th} observation
x^i = independent variable (temperature) for the i^{th} observation
ϵ^i = error term for the i^{th} observation
β_0 = intercept coefficient
β_1 = regression coefficient for the independent variable

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

Selecting the Best Model



Selecting the Best Model



SSE = 10.15

SSE = 6.03

SSE = 5.73

Other Error measures

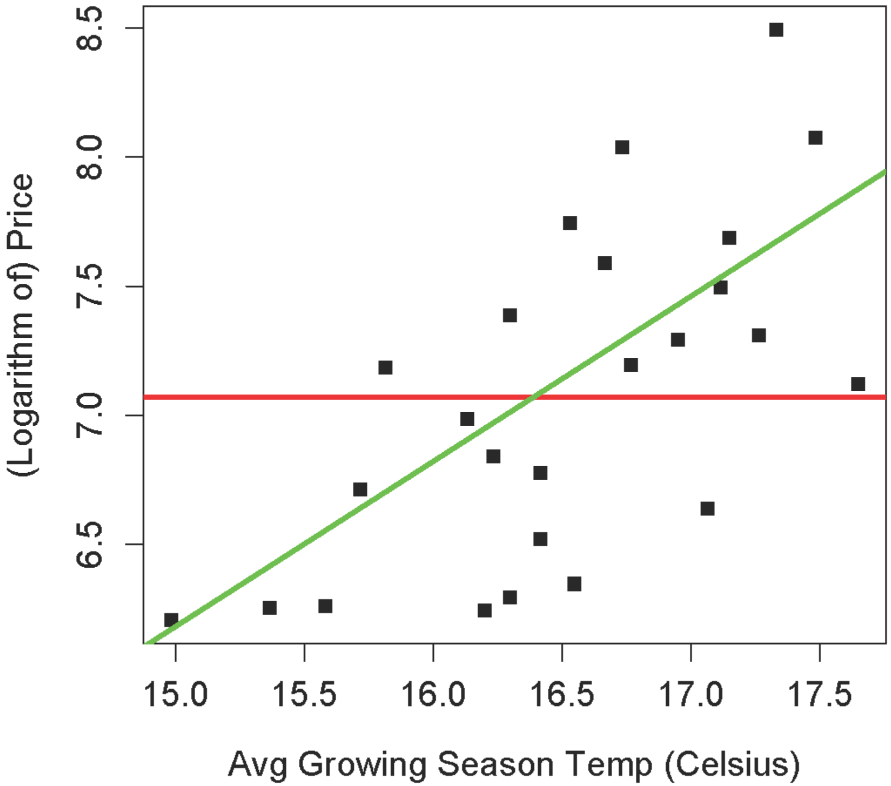
- SSE can be hard to interpret
  - Depends on N
  - Units are hard to understand

- Root-Mean-Square Error (RMSE)

$$RMSE = \sqrt{\frac{SSE}{N}}$$

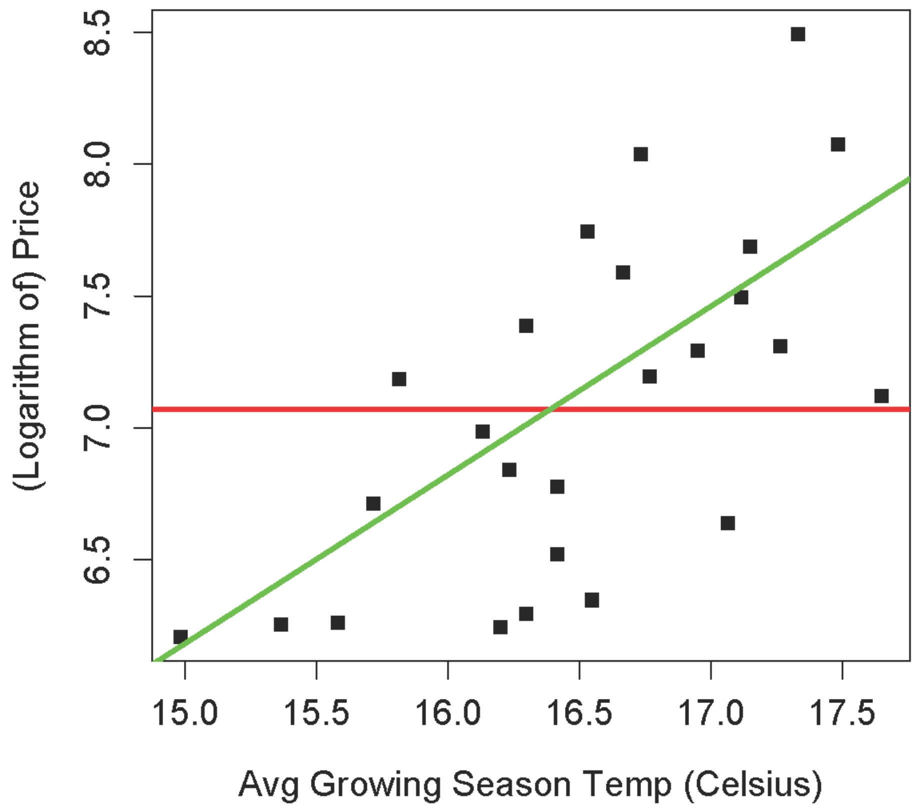
- Normalized by N, units of dependent variable

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



SSE = 5.73

SST = 10.15

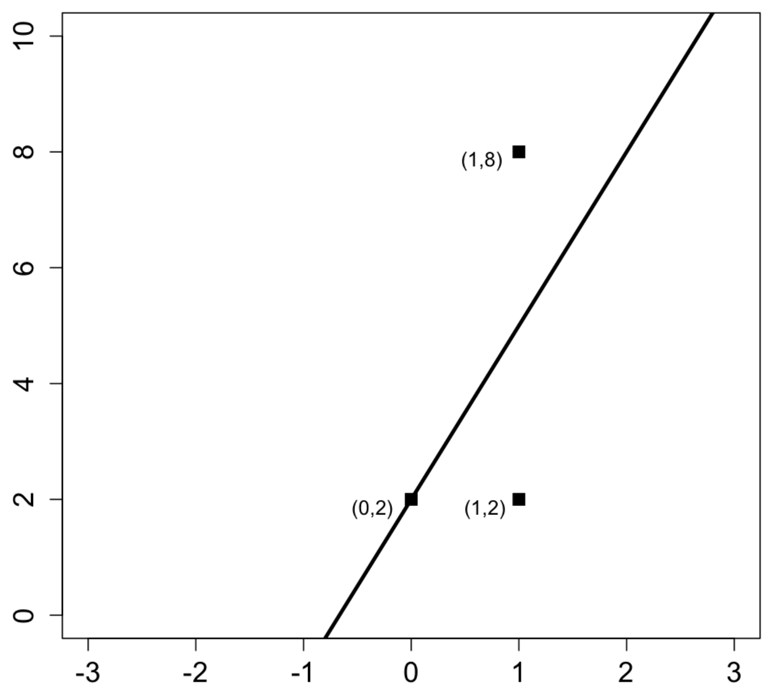
Interpreting R^2

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

- 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 $R^2 \approx 0$

Quick Question

- The following figure shows three data points and the best fit line
- $y = 3x + 2$.
- The x-coordinate, or "x", is our independent variable and the y-coordinate, or "y", is our dependent variable.



Quick Question

- Please answer the following questions using this figure.
  - What is the baseline prediction?
  - What is the Sum of Squared Errors (SSE) ?
  - What is the Total Sum of Squares (SST) ?
  - What is the R^2 of the model?

DIMENSIONLESS
TECHNOLOGY

Multiple Linear Regression

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

$$y^i = \beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \ldots + \beta_k x_k^i + \epsilon^i$$

y^i = dependent variable (wine price) for the i^{th} observation

x_j^i = j^{th} independent variable for the i^{th} observation

ϵ^i = error term for the i^{th} observation

β_0 = intercept coefficient

β_j = regression coefficient for the j^{th} independent variable

- Best model coefficients selected to minimize SSE

Adding Variables

Variables	R^2
Average Growing Season Temperature (AGST)	0.44
AGST, Harvest Rain	0.71
AGST, Harvest Rain, Age	0.79
AGST, Harvest Rain, Age, Winter Rain	0.83
AGST, Harvest Rain, Age, Winter Rain, Population	0.83

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

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

Quick Question

Q:-Suppose we add another variable, Average Winter Temperature, to our model to predict wine price. Is it possible for the model's R^2 value to go down from 0.83 to 0.80?

- ❑ No, the model's R^2 value can only decrease to 0.81 by adding new variables.
- ❑ No the model's R^2 value can not decrease at all , by adding new variables.
- ❑ Yes, the R^2 value could decrease to 0.80.

DIMENSIONLESS
TECHNOLOGY

Linear Regression in R

Linear Regression in R

- Read the file wine.csv. We will call our data frame "wine"
  - wine = read.csv("wine.csv")
- Look at the structure of our data by using the "str" function and explore all the variables
  - str(wine)
- Look at the statistical summary of our data
- using the summary function.
  - summary(wine)

One Variable Regression in R

- Create a one-variable linear regression equation using “AGST” to predict “Price”.
- We’ll call our regression model “model1”

```
model1 = lm(Price ~ AGST, data=wine)
```

- Look at the summary of model1.

```
summary(model1)
```

  - The first thing we see is a description of the function we used to build the model.
  - Then we see a summary of the residuals or error terms.
  - Following that is a description of the coefficients of our model.
  - The first row corresponds to the intercept term.
  - The second row corresponds to our independent variable, AGST.

Contd.

- Compute the sum of squared errors, or SSE, for our model.
- Residuals, or error terms, are stored in the vector “model1$residuals”.
- Compute the Sum of Squared Errors, or SSE,
    - sum(model1$residuals^2)

MLR in R

- Add another variable to our regression model, "HarvestRain".
- We'll call our new model "model2".
- Use the lm function to predict Price, but this time using "AGST" and "HarvestRain"

```
model2 = lm(Price ~ AGST + HarvestRain, data=wine)
```

- Look at the summary of our new model.
- Compute the SSE for this new model.

MLR in R contd.

- Build a IIIrd model with all the independent variables. Name it “model3”.
- Use the lm function to predict Price, but this time using all the independent variables.

```
model3 = lm(Price ~ AGST + HarvestRain + WinterRain + Age +
    FrancePop, data=wine)
```

- Look at the summary of our new model.
- Compute the SSE for this new model.

Quick Question

In R, use the dataset wine.csv to create a linear regression model to predict Price using HarvestRain and WinterRain as independent variables. Using the summary output of this model, answer the following questions:

- What is the "Multiple R-squared" value of your model?
- What is the coefficient for HarvestRain?
- What is the intercept coefficient?

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 1.27e-05 ***
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
```

Removing Variables

- As we just learned, both Age and France Population are insignificant in our model.
- Remove FrancePopulation and create a new model "model4"
  - model4 = lm(Price ~ AGST + HarvestRain + WinterRain + Age, data=wine)
- Look at the summary of model4.
- This model is just as strong, if not stronger than the previous model,
- Before, Age was not significant at all in our model. But now,Age has two stars, meaning that it's very significant in this new model.
- This is due to something called multicollinearity.

Quick Question

Use the dataset wine.csv to create a linear regression model to predict Price using HarvestRain and WinterRain as independent variables, like you did in the previous quick question. Using the summary output of this model, answer the following questions:

- Is the coefficient for HarvestRain significant?
- Is the coefficient for WinterRain significant?

DIMENSIONLESS
TECHNOLOGY

Correlation and Multicollinearity

Correlation

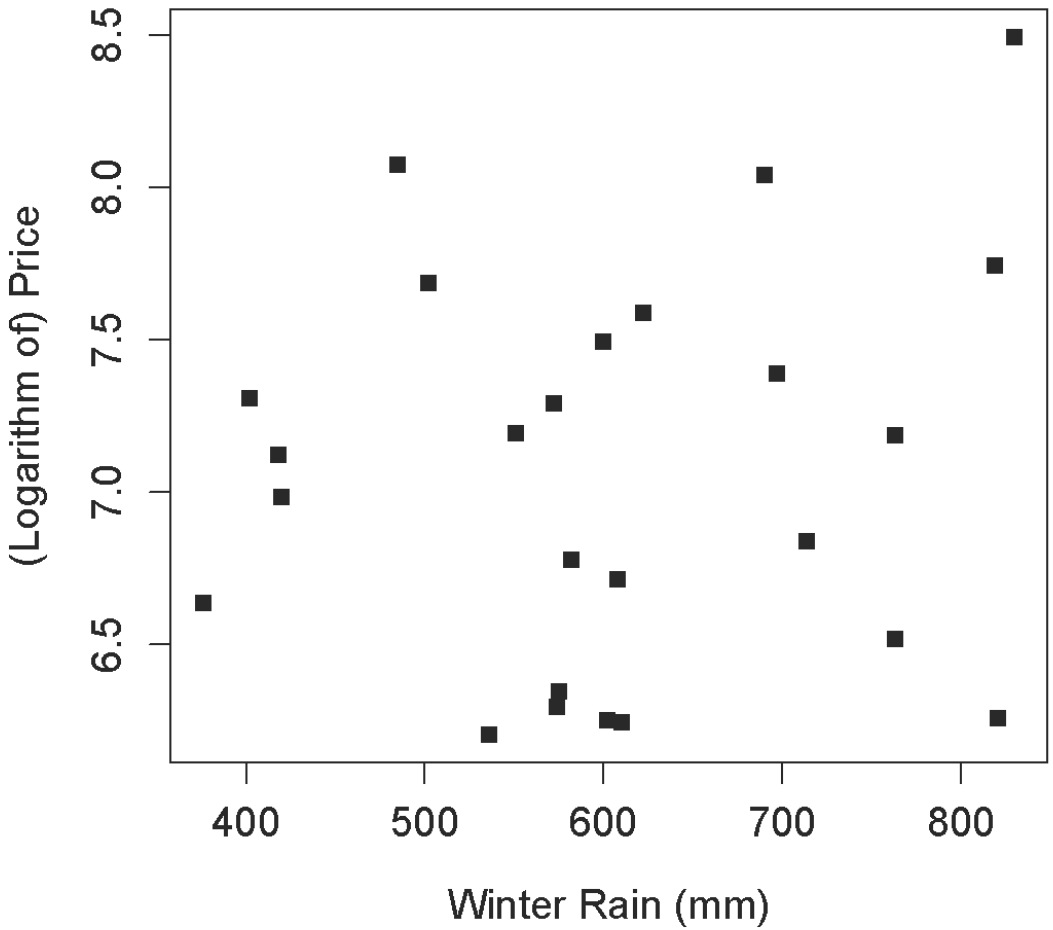
A measure of the linear relationship between variables

+1 = perfect positive linear relationship

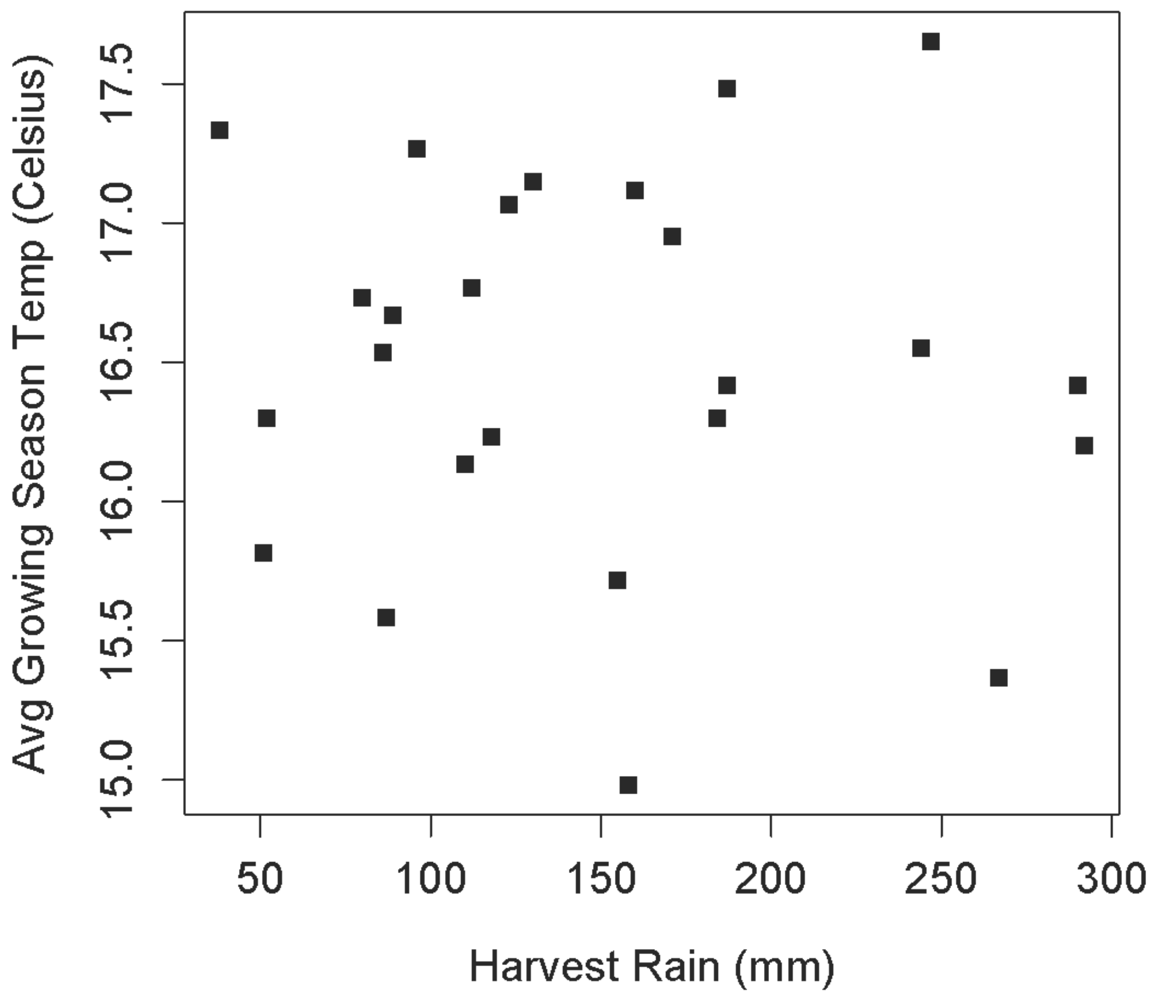
0 = no linear relationship

-1 = perfect negative linear relationship

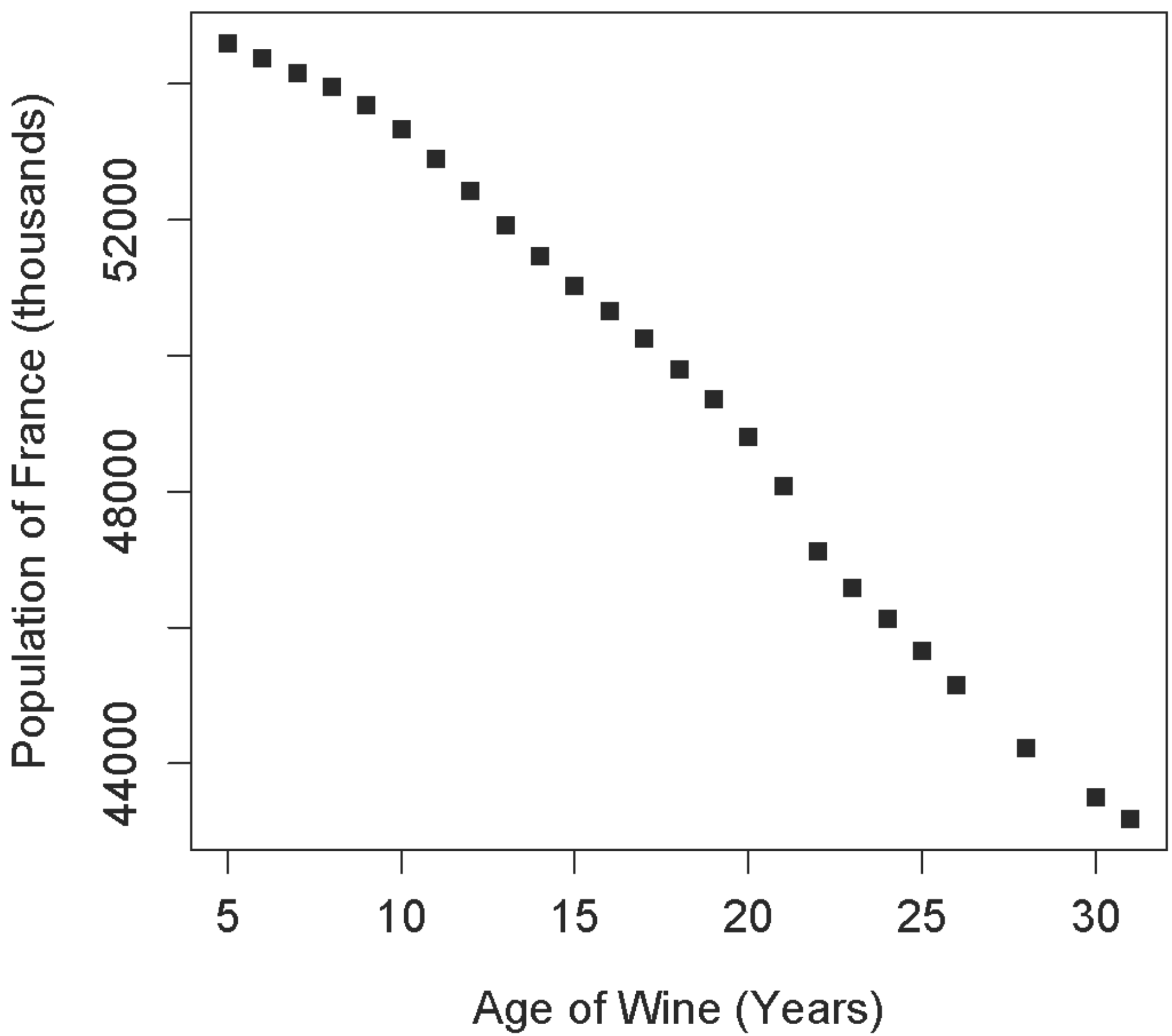
Examples of Correlation



Examples of Correlation



Examples of Correlation



Correlation in R

- We will use “cor()” function
  - cor(wine$WinterRain, wine$Price)
  - cor(wine$Age, wine$FrancePop)
  - cor(wine)

```
> cor(wine)
                   Year      Price   WinterRain        AGST HarvestRain         Age
Year         1.00000000 -0.4477679  0.016970024 -0.24691585  0.02800907 -1.00000000
Price       -0.44776786  1.0000000  0.136650547  0.65956286 -0.56332190  0.44776786
WinterRain   0.01697002  0.1366505  1.000000000 -0.32109061 -0.27544085 -0.01697002
AGST        -0.24691585  0.6595629 -0.321090611  1.00000000 -0.06449593  0.24691585
HarvestRain  0.02800907 -0.5633219 -0.275440854 -0.06449593  1.00000000 -0.02800907
Age         -1.00000000  0.4477679 -0.016970024  0.24691585 -0.02800907  1.00000000
FrancePop    0.99448510 -0.4668616 -0.001621627 -0.25916227  0.04126439 -0.99448510
               FrancePop
Year         0.994485097
Price       -0.466861641
WinterRain  -0.001621627
AGST        -0.259162274
HarvestRain  0.041264394
Age         -0.994485097
FrancePop    1.000000000
```

Removing the Variables

Let's see what would have happened if we had removed both Age and FrancePopulation.

- Let's call it Model5
  - model5 = lm(Price ~ AGST + HarvestRain + WinterRain, data=wine)
- See the summary of "model5"
- AGST and HarvestRain are very significant, and WinterRain is almost significant.
- R^2 reduced to 0.75 from 0.83 (when we used age)
- Removing both would make us miss a significant variable.
- Why can't we keep France population and remove Age?

Multicollinearity

- Multicollinearity reminds us that coefficients are only interpretable in the presence of other variables being used.
- High correlations can even cause coefficients to have an unintuitive sign.
- There is no definitive cut-off value for what makes a correlation too high.
- But typically, a correlation greater than 0.7 or less than -0.7 is cause for concern.
- By seeing the correlation table, we can say Model4 is our best guess.

Quick Question

Using the data set wine.csv, what is the correlation between HarvestRain and WinterRain?

Multicollinearity

- A simple way to detect collinearity is to look at the correlation matrix of the predictors.

- An element of this matrix that is large in absolute value indicates a pair of highly correlated variables, and therefore a collinearity problem in the data.

- Instead of inspecting the correlation matrix, a better way to assess multi- collinearity is to compute the *variance inflation factor* (VIF).

VIF

- The VIF is variance inflation factor the ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own.
- The smallest possible value for VIF is 1, which indicates the complete absence of collinearity.
- Typically in practice there is a small amount of collinearity among the predictors.
- As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

Calculating VIF

- The VIF for each variable can be computed using the formula

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{Xj|X-j}$ is the R^2 from a regression of Xj onto all of the other predictors.

If $R^2_{Xj|X-j}$ is close to one, then collinearity is present, and so the VIF will be large

VIF in R

- Install the package "car"
- Load the package library(car)

➢vif(model3)

```
 AGST    HarvestRain WinterRain    Age     FrancePop
1.274536 1.116584    1.298801  97.219725 98.252693
```

- The high values of vif for Age and FrancePop indicates that there is multi collinearity.

DIMENSIONLESS
TECHNOLOGY

Making Predictions

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

Predictive Ability

- We need to build a model that does well at predicting data it's never seen before.

- The data that we use to build a model is often called the training data and the new data is often called the test data.

- The accuracy of the model on the test data is often referred to as out-of-sample accuracy.

Making Predictions in R

- Load the new data file wine_test.csv into R.
https://storage.googleapis.com/dimensionless/Analytics/wine_test.csv
- We'll call the data frame wineTest.
- Looking at the structure of wineTest,we can see that we have two observations and the same seven variables as before.
- To make predictions for these two test points, we'll use the "predict()" function.

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

Output

Function

Model for Prediction

Data Set

Results

- For the first data point we predict 6.7689, and for the second data point we predict 6.6849.

- We can see that the actual Price for the first data point is 6.95, and the actual Price for the second data point is 6.5.

- Looks like our predictions are pretty good, but we can quantify this by computing the R-square value for our test set.

Computing R^2

- Compute SSE first
  - SSE = sum((wineTest$Price - predictTest)^2)
- Compute SST
  - SST = sum((wineTest$Price - mean(wine$Price))^2)

- R^2 = 0.79, which is pretty good out-of-sample R-squared.

- We should increase the size of our test set to be more confident about the out-of-sample accuracy of our model.

Out Of Sample R^2

Variables	Model R^2	Test R^2
AGST	0.44	0.79
AGST, Harvest Rain	0.71	-0.08
AGST, Harvest Rain, Age	0.79	0.53
AGST, Harvest Rain, Age, Winter Rain	0.83	0.79
AGST, Harvest Rain, Age, Winter Rain, Population	0.83	0.76

- 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!

Quick Question

Which of the following are NOT valid values for an out-of-sample (test set) R^2 ? Select all that apply.

- [ ] -7.0
- [ ] -0.3
- [ ] 0.0
- [ ] 0.6
- [ ] 1.0
- [ ] 2.4

DIMENSIONLESS
TECHNOLOGY

Comparing the Model to the Experts

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"

What we have developed is a linear regression model, a simple but rather powerful model for predicting quality of wines. It only used few variables and we have seen that it predicted wine prices quite well. In fact, in many cases it outperformed wine expert's opinions.