THE STATISTICAL SOMMELIER

An Introduction to Linear Regression

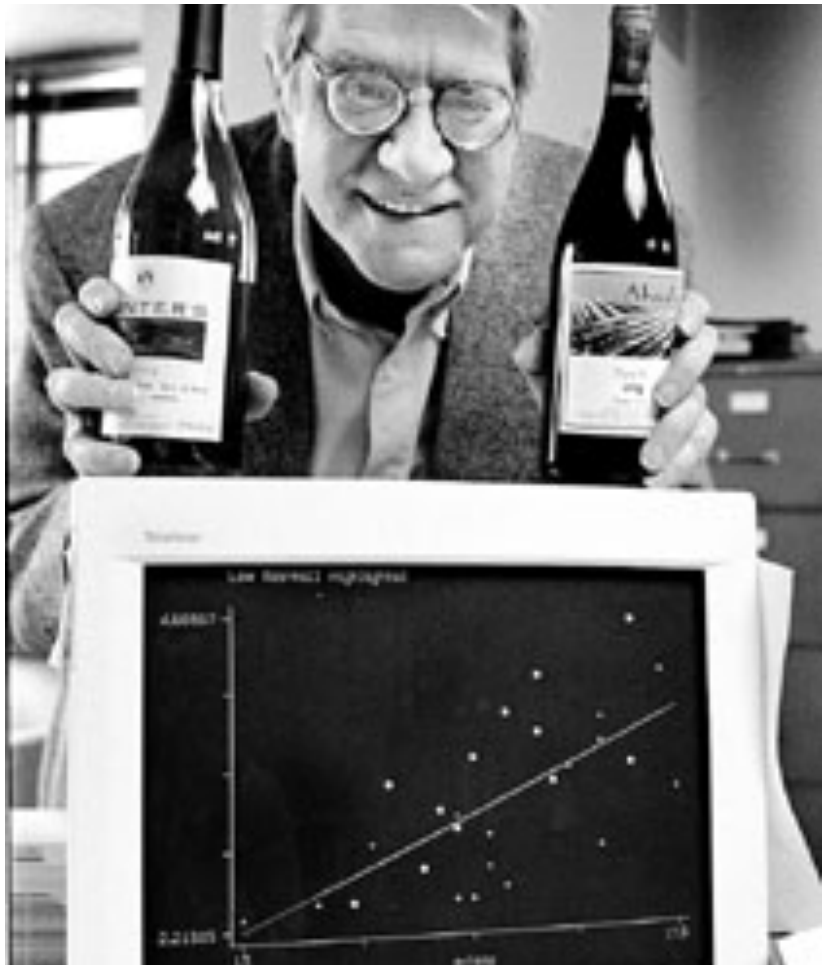
15.071 – The Analytics Edge

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



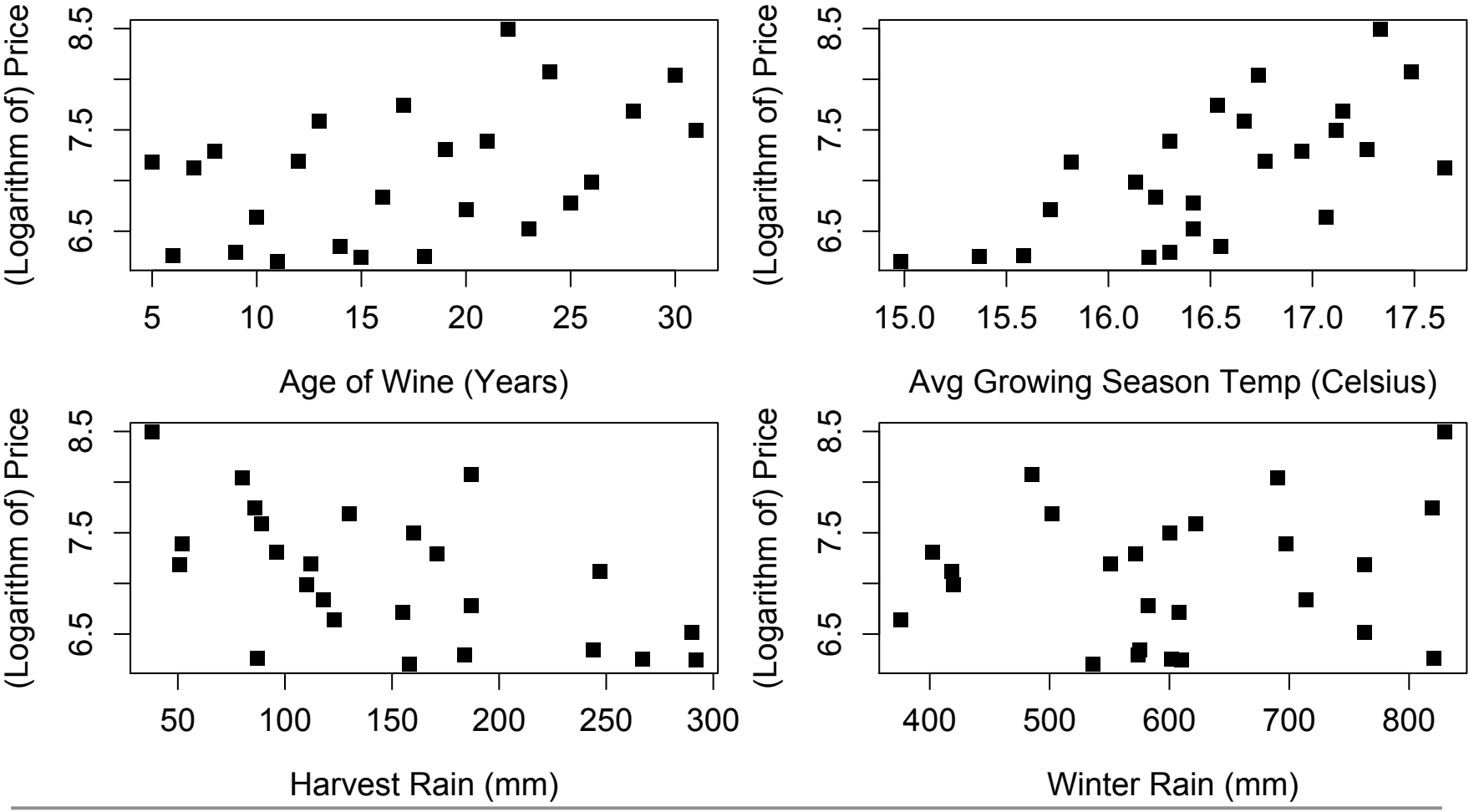
Building a Model

- 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)



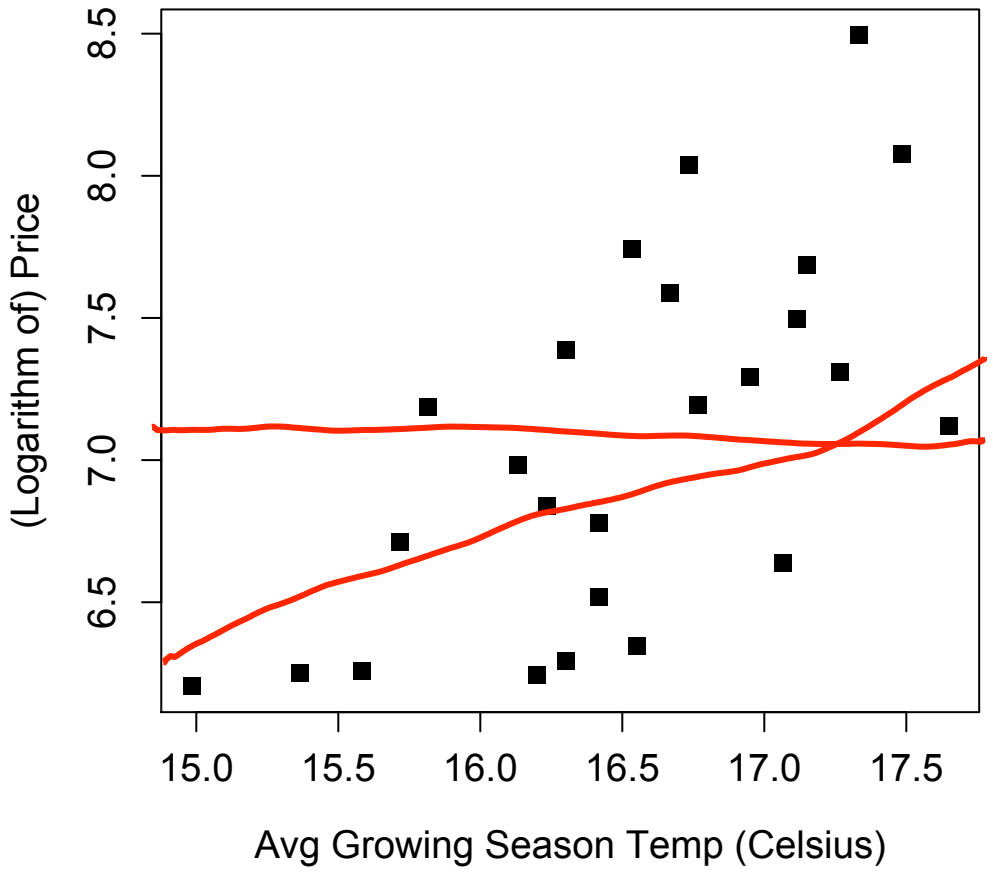
The Expert's Reaction

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

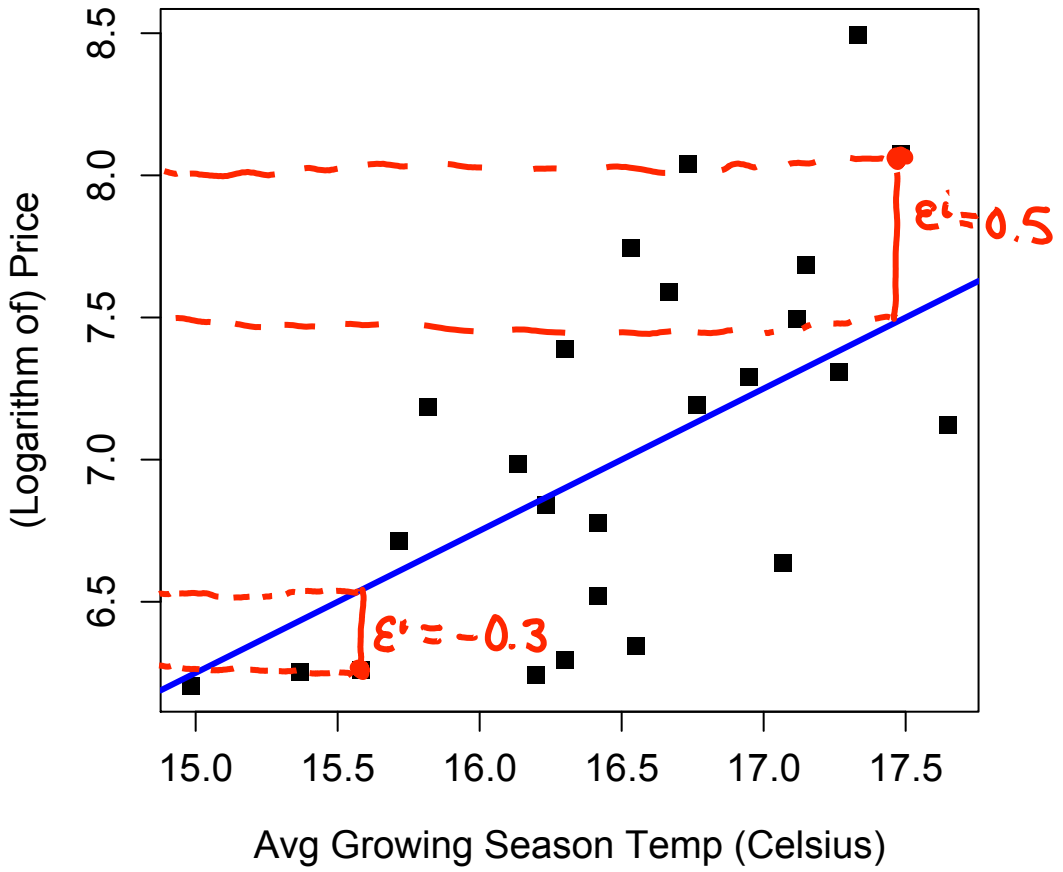
- One-variable regression model

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

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

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

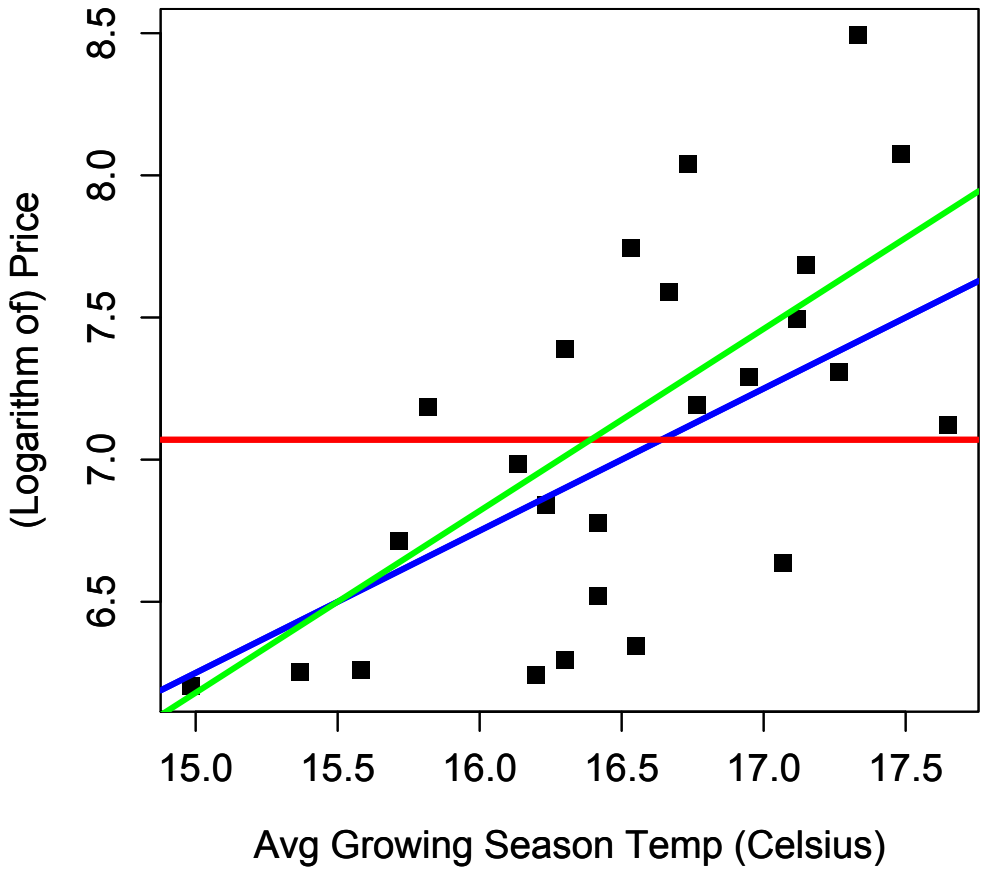
Selecting the Best Model



$$SSE = (\varepsilon^1)^2 + (\varepsilon^2)^2 + \ldots + (\varepsilon^N)^2$$

N = # data points

Selecting the Best Model



SSE = 10.15

SSE = 6.03

SSE = 5.73

Other Error Measures

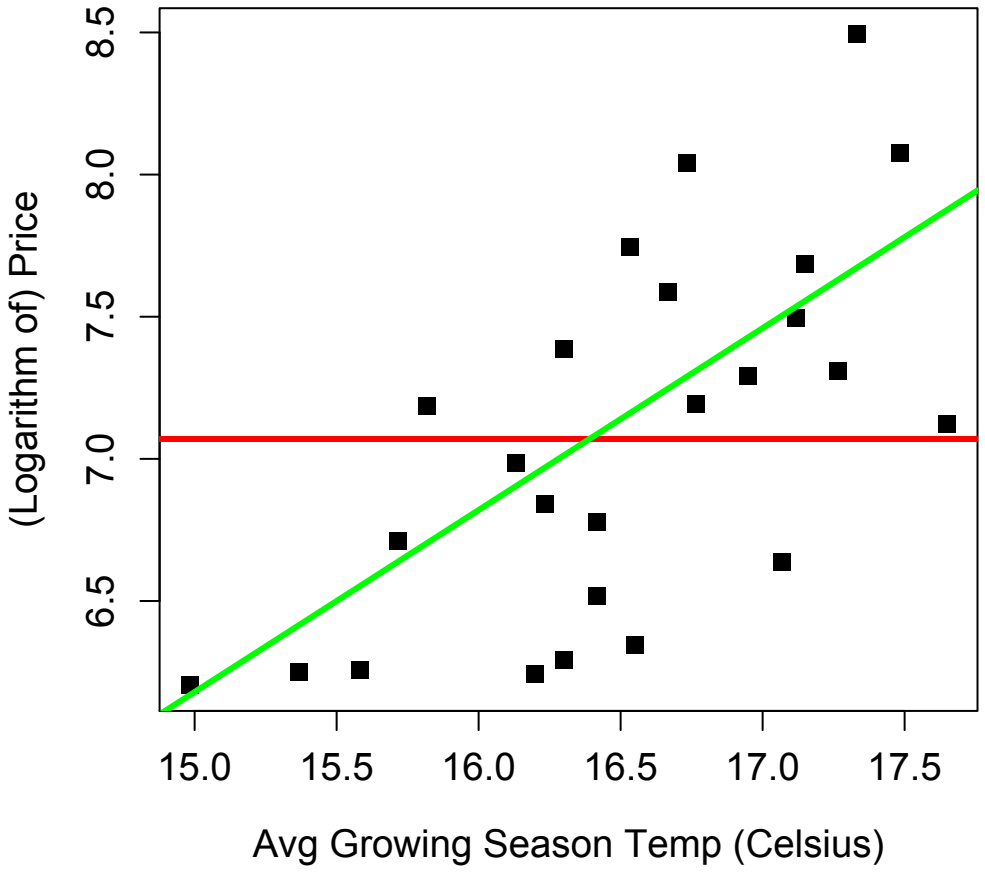
- SSE can be hard to interpret
  - Depends on N
  - Units are hard to understand

- Root-Mean-Square Error (RMSE)

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

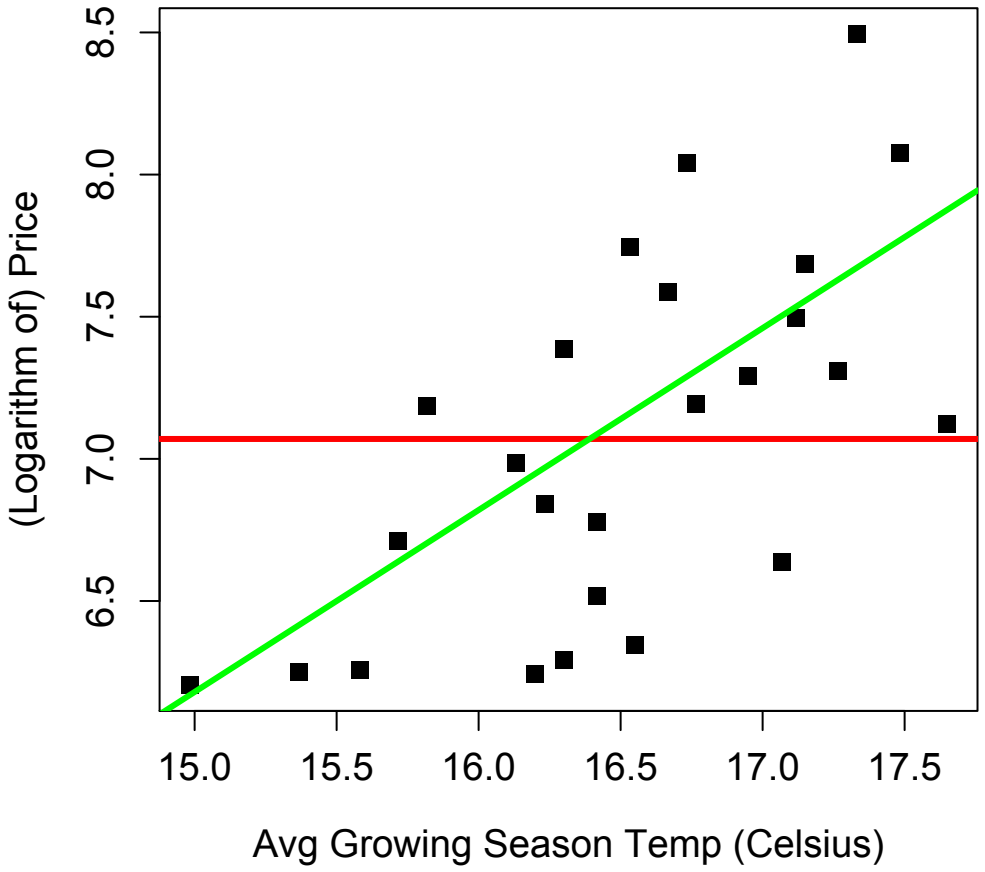
- Normalized by N, units of dependent variable

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

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

$= 1 - \frac{5.73}{10.15}$

$= 0.44$

Interpreting R^2

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

$0 \leq SSE \leq SST$

$0 \leq SST$

- 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

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

y^i = dependent variable (wine price) for the ith observation

x_j^i = jth independent variable for the ith observation

ϵ^i = error term for the ith observation

β_0 = intercept coefficient

β_j = regression coefficient for the jth 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

Understanding the Model and Coefficients

Estimate / Std. Error

```
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
```

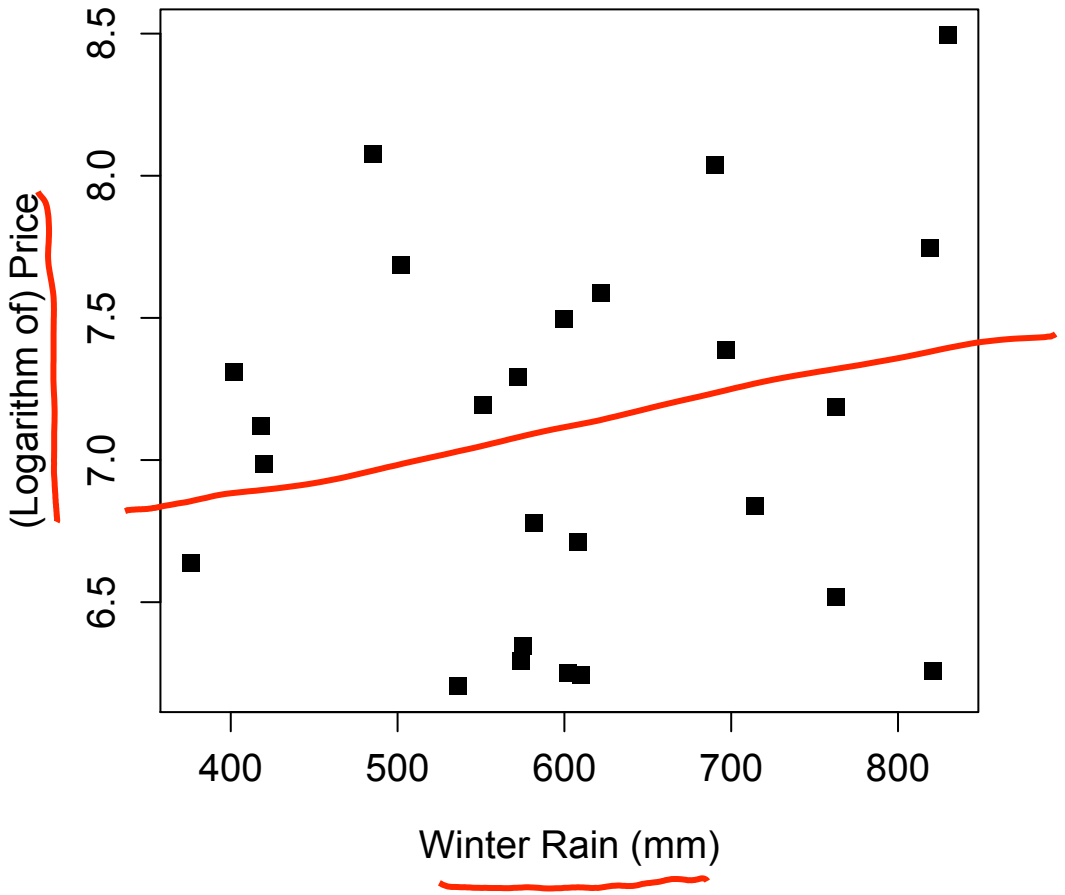
Correlation

A measure of the linear relationship between variables

- +1 = perfect positive linear relationship
- 0 = no linear relationship
- -1 = perfect negative linear relationship

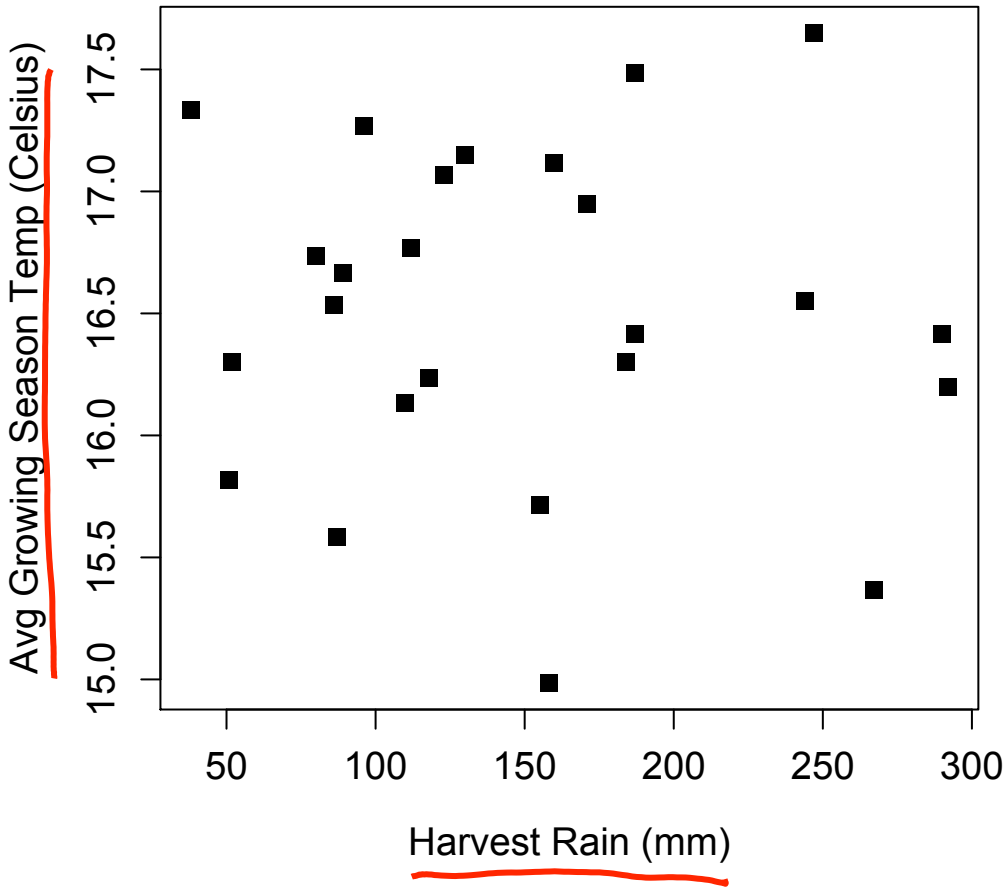
Examples of Correlation

cor
= 0.14



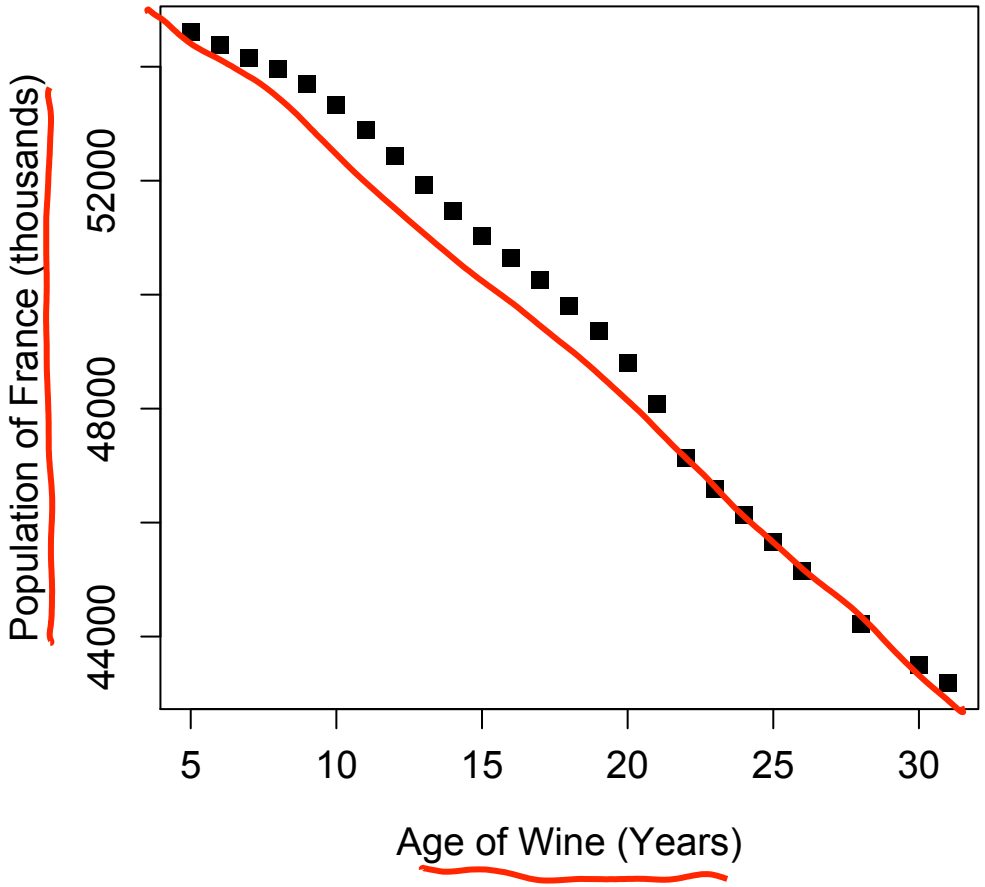
Examples of Correlation

Cor
= -0.06



Examples of Correlation

cor
= −0.99



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

training

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

test

  - Bordeaux wine buyers profit from being able to predict the quality of a wine years before it matures

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!

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

MONEYBALL

The Power of Sports Analytics

15.071 – The Analytics Edge

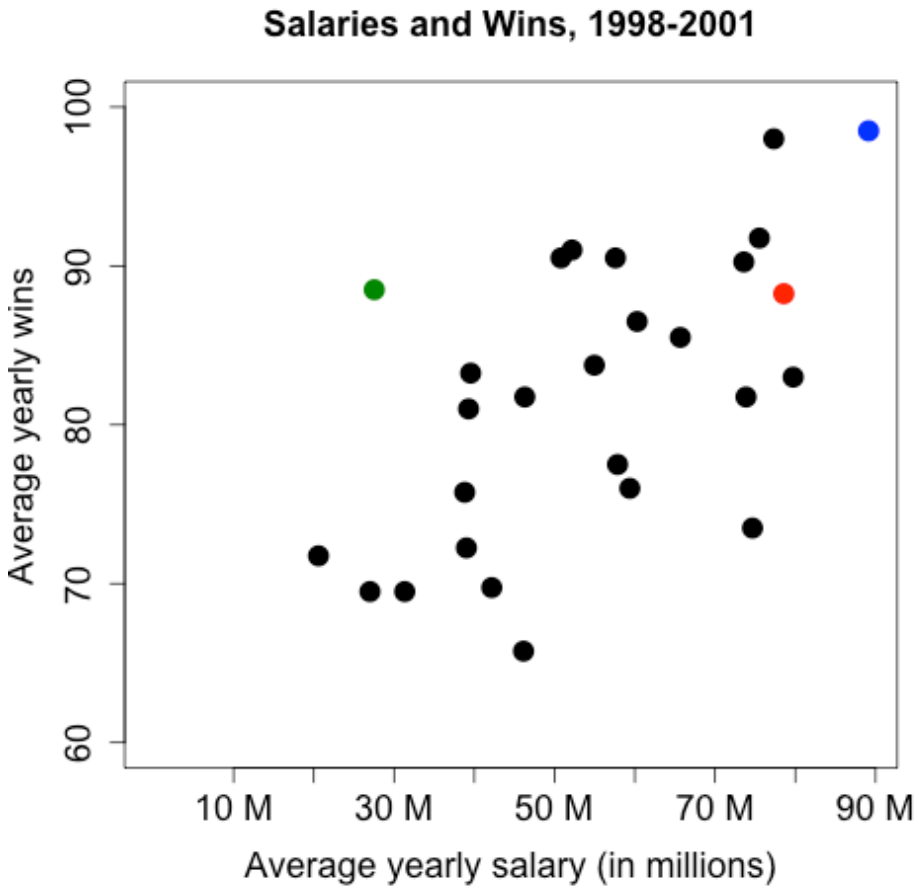
The Story

- *Moneyball* tells the story of the Oakland A's in 2002
  - One of the poorest teams in baseball
    - New ownership and budget cuts in 1995
  - But they were improving

Year	Win %
1997	40%
1998	46%
1999	54%
2000	57%
2001	63%

- How were they doing it?
  - Was it just luck?
- In 2002, the A's lost three key players
- Could they continue winning?

The Problem



- Rich teams can afford the all-star players
- How do the poor teams compete?

Competing as a Poor Team

- Competitive imbalances in the game
  - Rich teams have four times the salary of poor teams
- The Oakland A's can't afford the all-stars, but they are still making it to the playoffs. How?
- They take a quantitative approach and find undervalued players

A Different Approach

- The A's started using a different method to select players
- The traditional way was through scouting
  - Scouts would go watch high school and college players
  - Report back about their skills
  - A lot of talk about speed and athletic build
- The A's selected players based on their statistics, not on their looks

  "The statistics enabled you to find your way past all sorts of sight-based scouting prejudices."

  "We're not selling jeans here"

The Perfect Batter

The A's

A catcher who couldn't throw
Gets on base a lot

The Yankees

A consistent shortshop
Leader in hits and stolen bases

The Perfect Pitcher

The A's

Unconventional delivery
Slow speed

The Yankees

Conventional delivery
Fast speed

Billy Beane

- The general manager since 1997
- Played major league baseball, but never made it big
  - Sees himself as a typical scouting error
- Billy Beane succeeded in using analytics
  - Had a management position
  - Understood the importance of statistics – hired Paul DePodesta (a Harvard graduate) as his assistant
  - Didn't care about being ostracized

Taking a Quantitative View

- Paul DePodesta spent a lot of time looking at the data
- His analysis suggested that some skills were undervalued and some skills were overvalued
- If they could detect the undervalued skills, they could find players at a bargain

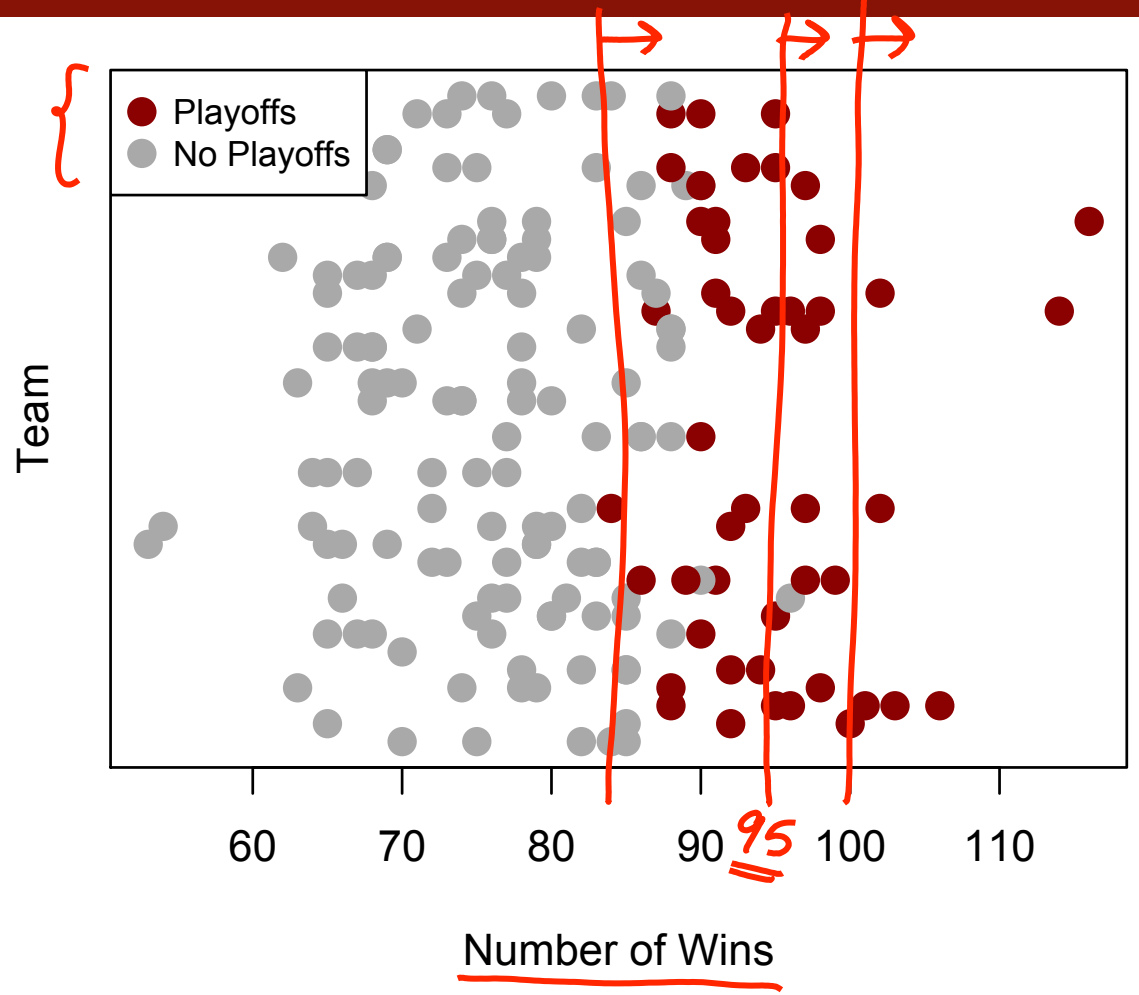
The Goal of a Baseball Team

Making it to the Playoffs

- How many games does a team need to win in the regular season to make it to the playoffs?

- "Paul DePodesta reduced the regular season to a math problem. He judged how many wins it would take to make it to the playoffs: 95."

Making it to the Playoffs



Data from all teams 1996-2001

Winning 95 Games

- How does a team win games?
- They score more runs than their opponent
- But how many more?
- The A's calculated that they needed to score 135 more runs than they allowed during the regular season to expect to win 95 games
- Let's see if we can verify this using linear regression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 80.881375   0.131157  616.67   <2e-16 ***
RD           0.105766   0.001297   81.55   <2e-16 ***
```

$$W = 80.8814 + 0.1058(RD)$$

$$W \geq 95$$

$$80.8814 + 0.1058(RD) \geq 95$$

$$RD \geq \frac{95 - 80.8814}{0.1058} = 133.4 \sim 135$$

The Goal of a Baseball Team

Scoring Runs

- How does a team score more runs?
- The A's discovered that two baseball statistics were significantly more important than anything else
- On-Base Percentage (OBP)
  - Percentage of time a player gets on base (including walks)
- Slugging Percentage (SLG)
  - How far a player gets around the bases on his turn (measures power)

Scoring Runs

- Most teams focused on Batting Average (BA)
  - Getting on base by hitting the ball
- The A's claimed that:
  - On-Base Percentage was the most important
  - Slugging Percentage was important
  - Batting Average was overvalued
- Can we use linear regression to verify which baseball stats are more important to predict runs?

Allowing Runs

- We can use pitching statistics to predict runs allowed
  - Opponents On-Base Percentage (OOBP)
  - Opponents Slugging Percentage (OSLG)

- We get the linear regression model

  Runs Allowed = -837.38 + 2913.60(OOBP) + 1514.29(OSLG)

- $R^2 = 0.91$
- Both variables significant

Predicting Runs and Wins

- Can we predict how many games the 2002 Oakland A's will win using our models?
- The models for runs use team statistics
- Each year, a baseball team is different
- We need to estimate the new team statistics using past player performance
  - Assumes past performance correlates with future performance
  - Assumes few injuries
- We can estimate the team statistics for 2002 by using the 2001 player statistics

Predicting Runs Scored

- At the beginning of the 2002 season, the Oakland A's had 24 batters on their roster
- Using the 2001 regular season statistics for these players
  - Team OBP is 0.339
  - Team SLG is 0.430
- Our regression equation was

$$RS = -804.63 + 2737.77(OBP) + 1584.91(SLG)$$

- Our 2002 prediction for the A's is

$$RS = -804.63 + 2737.77(0.339) + 1584.91(0.430) = 805$$

Predicting Runs Allowed

- At the beginning of the 2002 season, the Oakland A's had 17 pitchers on their roster
- Using the 2001 regular season statistics for these players
  - Team OOBP is 0.307
  - Team OSLG is 0.373
- Our regression equation was

  $$RA = -837.38 + 2913.60(OOBP) + 1514.29(OSLG)$$
- Our 2002 prediction for the A's is

  $$RA = -837.38 + 2913.60(0.307) + 1514.29\ (0.373) = 622$$

Predicting Wins

- Our regression equation to predict wins was

  $$\text{Wins} = 80.8814 + 0.1058(\text{RS} - \text{RA})$$

- We predicted
  - RS = 805
  - RA = 622

- So our prediction for wins is

  $$\text{Wins} = 80.8814 + 0.1058(805 - 622) = 100$$

The Oakland A's

- Paul DePodesta used a similar approach to make predictions
- Predictions closely match actual performance

	Our Prediction	Paul's Prediction	Actual
Runs Scored	805	800 – 820	800

The Oakland A's

- Paul DePodesta used a similar approach to make predictions
- Predictions closely match actual performance

	Our Prediction	Paul's Prediction	Actual
Runs Scored	805	800 – 820	800
Runs Allowed	622	650 – 670	653

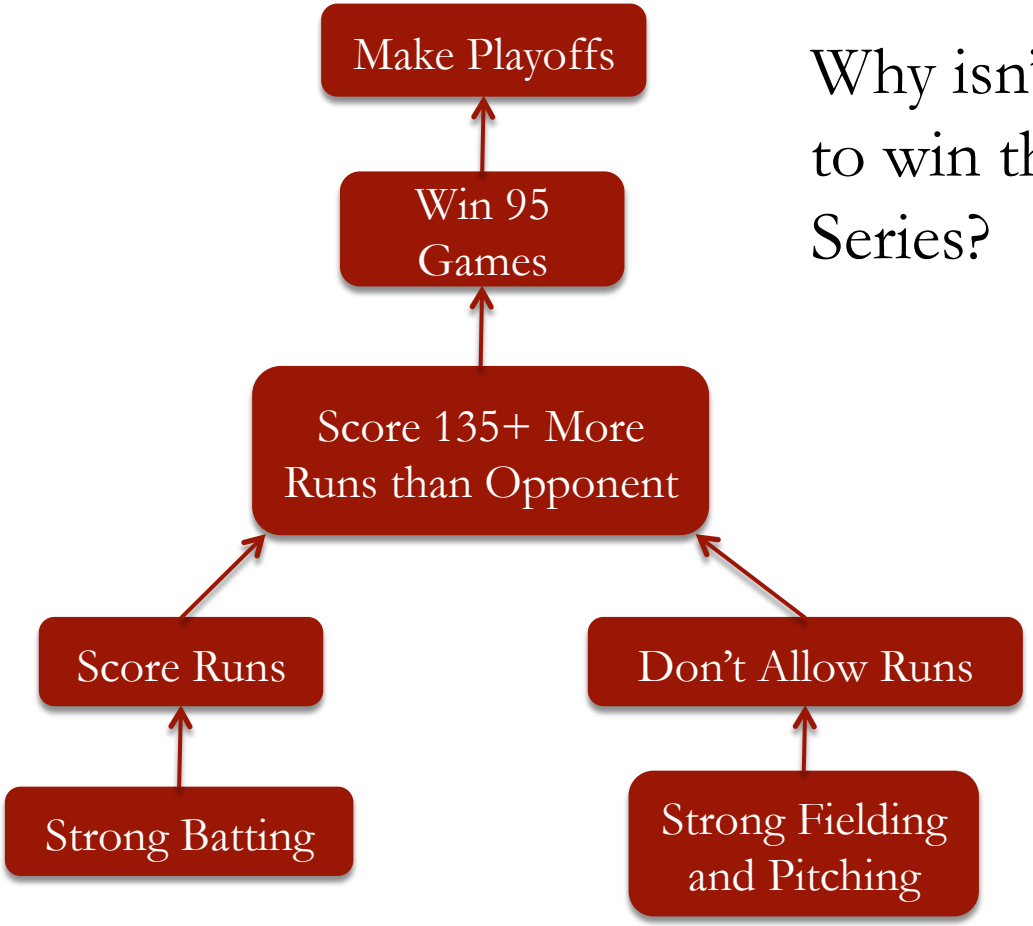
The Oakland A's

- Paul DePodesta used a similar approach to make predictions
- Predictions closely match actual performance

	Our Prediction	Paul's Prediction	Actual
Runs Scored	805	800 – 820	800
Runs Allowed	622	650 – 670	653
Wins	100	93 – 97	103

- The A's set a League record by winning 20 games in a row
- Won one more game than the previous year, and made it to the playoffs

The Goal of a Baseball Team



Why isn't the goal to win the World Series?

Luck in the Playoffs

- Billy and Paul see their job as making sure the team makes it to the playoffs – after that all bets are off
  - The A's made it to the playoffs in 2000, 2001, 2002, 2003
  - But they didn't win the World Series
- Why?
- "Over a long season the luck evens out, and the skill shines through. But in a series of three out of five, or even four out of seven, anything can happen."

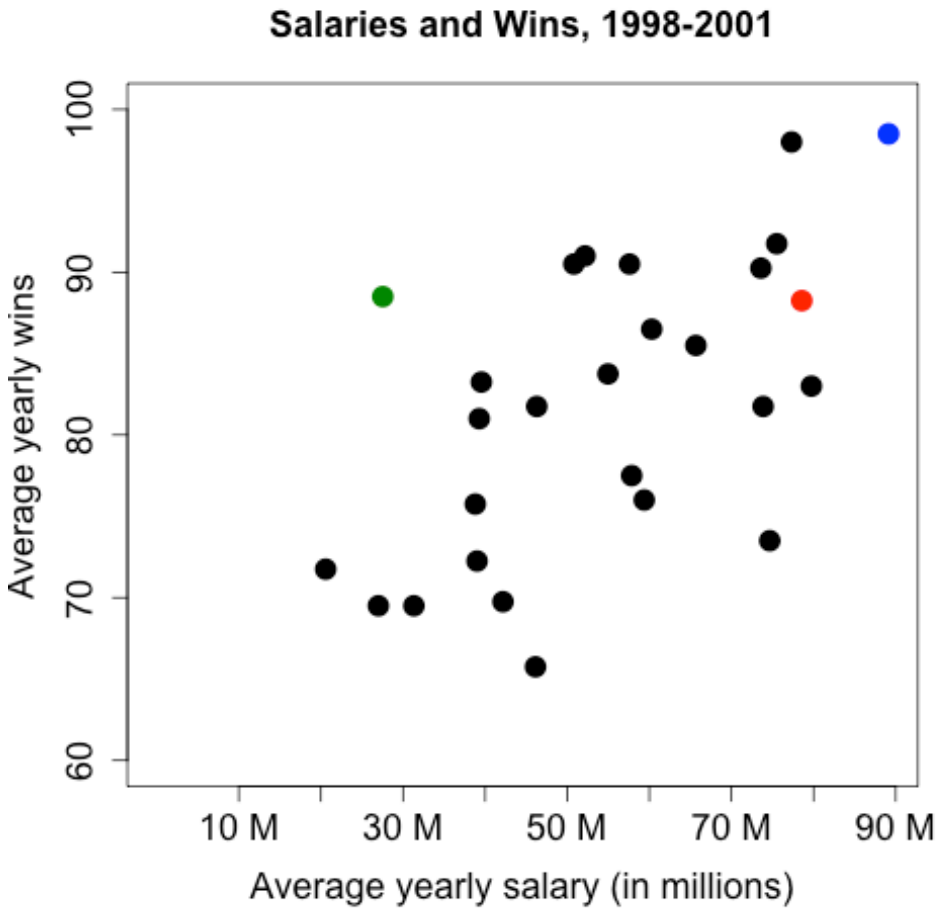
Is Playoff Performance Predictable?

- Using data 1994-2011 (8 teams in the playoffs)
- Correlation between winning the World Series and regular season wins is 0.03
- Winning regular season games gets you to the playoffs
- But in the playoffs, there are too few games for luck to even out
- *Logistic regression* can be used to predict whether or not a team will win the World Series

Other Moneyball Strategies

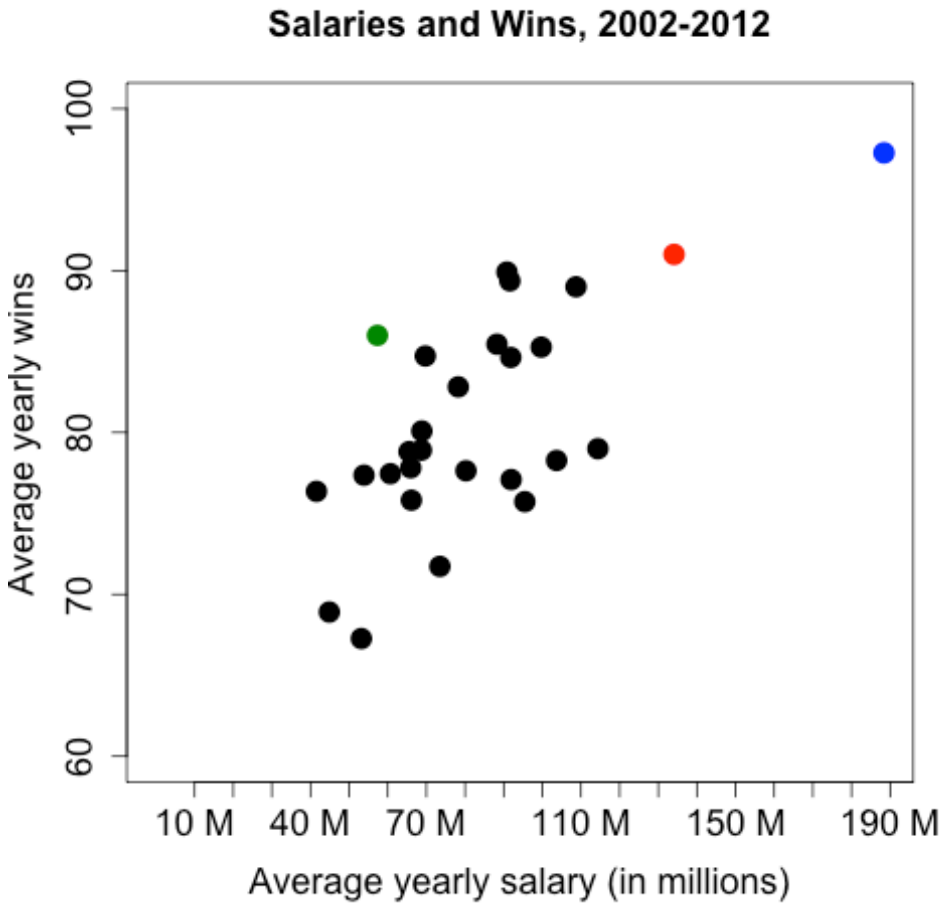
- *Moneyball* also discusses:
  - How it is easier to predict professional success of college players than high school players
  - Stealing bases, sacrifice bunting, and sacrifice flies are overrated
  - Pitching statistics do not accurately measure pitcher ability – pitchers only control strikeouts, home runs, and walks

Where was Baseball in 2002?



- Before Moneyball techniques became more well-known, the A's were an outlier
- 20 more wins than teams with equivalent payrolls
- As many wins as teams with more than double the payroll

Where is Baseball Now?



- Now, the A's are still an efficient team, but they only have 10 more wins than teams with equivalent payrolls
- Fewer inefficiencies

Sabermetrics

- Sabermetrics is a more general term for Moneyball techniques
- There has been a lot of work done in this field
  - Baseball Prospectus (www.baseballprospectus.com)
  - Value Over Replacement Player (VORP)
  - Defense Independent Pitching Statistics (DIPS)
  - *The Extra 2%: How Wall Street Strategies Took a Major League Baseball Team from Worst to First*
    - A story of the Tampa Bay Rays
  - Game-time decisions: batting order, changing pitchers, etc.

Other Baseball Teams and Sports

- Every major league baseball team now has a statistics group
- The Red Sox implemented quantitative ideas and won the World Series for the first time in 86 years
- Analytics are also used in other sports, although it is believed that more teams use statistical analysis than is publically known

The Analytics Edge

- Models allow managers to more accurately value players and minimize risk
  - "In human behavior there was always uncertainty and risk. The goal of the Oakland front office was simply to minimize the risk. Their solution wasn't perfect, it was just better than ... rendering decisions by gut feeling."
- Relatively simple models can be useful