

Section 4: Predictive Models

Multiple and Logistic Regression, Trees

Vadim Sokolov

Suggested Reading

OpenIntro Statistics, Chapter 8

Previous Section

- ▶ Linear Patterns in Data (Leavitt, House Price)
- ▶ Simple Linear Regression
- ▶ Predictions (Confidence and Prediction Intervals)
- ▶ Least Squares Principle
- ▶ Hypothesis Testing (Google vs SP500)
- ▶ Model Diagnostics (Cancer and Smoking Data)
- ▶ Data transformations (World's Smartest Mammal)

This Section

- ▶ Multiple Regression (Newfood study, Golf Analysis)
- ▶ Interactions (how advertisement change price elasticity?)
- ▶ Predictive analytics cases(Target, Walmart, Airbnb, Stitch Fix)
- ▶ Logistic regression (NBA predictions, Horse predictions, LinkedIn)

Multiple Regression

$$y = \beta_0 + \beta_1 x \quad \text{simple LReg}$$

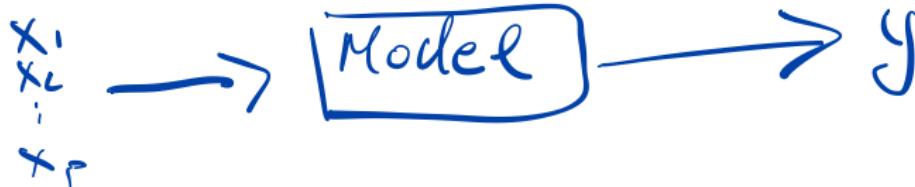
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Many problems involve more than one independent (explanatory) variable or factor which affects the dependent or response variable.

- ▶ Multi-factor asset pricing models (APT). Stock returns, book-to-market ratios, Interest rates

- ▶ Demand for a product given prices of competing brands, advertising, household attributes (to formulate pricing strategies)

- ▶ Internet Analytics What do I like? Suggestions instead of Search! Alexa “book my Xmas vacation,” “buy my best friend a birthday present”



R Regression Commands

Given input-output vectors x and y `cor(...)` computes correlation table

`model = lm(y ~ x)` for linear model (a.k.a regression)

`model = glm(y ~ x)` for logistic regression

`model = lm(y ~ x1+ ... + xp)` for linear multiple regression model

`plot(model)` diagnostics

`plot(cooks.distance(model))` influential points

`rstudent(model)` outliers

$$y_i - \hat{y}_i$$

`summary(model)` provides a summary analysis of our model

`newdata = data.frame(...)` constructs a new input variable

`predict.lm(model, newdata)` provides a prediction at a new input
in Excel

`linest(yrange,xrange)` and `slope(yrange,xrange)`

Regression Model

Y = response or outcome variable

X_1, \dots, X_p = explanatory or input variable

The general relationship is given by

$$Y = f(X_1, \dots, X_p) + \epsilon$$

And a linear relationship is written

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}_{\text{deterministic}} + \epsilon$$

sfock.

MLR Assumptions

The Multiple Linear Regression (MLR) model

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}_{\text{linear}} + \epsilon ; \quad \epsilon \sim N(0, \sigma^2)$$

assumptions follow those of simple linear regression:

1. The conditional mean of Y is **linear** in the X_j variables
2. The errors are normal $N(0, \sigma^2)$.

We write

$$Y | X_1, \dots, X_p \sim N(\underbrace{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}_{\text{centred}}, \sigma^2)$$

Normal
centred
doesn't depend on X

g_i & g_j are indep.

Statistical versus Economic Significance

$\beta_0 + \beta_1 \text{ coeff}$

$\beta_0, \beta_1, \dots, \beta_p$

When looking at the β coefficients there are two issues

1. Statistical Significance: The t-ratios of the β 's
2. Economic Significance: The magnitudes of the β 's

If X_i increases by one unit **holding the other X's constant**

β_i is stat sig.
if $P\text{-val} \leq 0.05$
 $H_0: \beta_i = 0$

Then Y will react by β_i units. **Marginal effects** this means

Units of β_i is $\frac{\text{units of } Y}{\text{units of } X_i}$
They are called **marginal effects**

that input X_i is important.

At the end of the day use your judgment!

s.e. tstat

p-val

Model Diagnostics

1. Predict - *Doubt core about LM assumptions*
2. Interpret - *Need to make sure that LM assumptions are satisfied.*

`plot(model)` provides diagnostics **before** model building!

There are many possible caveats

1. Running simple regressions gives you the wrong answer!
2. **Multiple regression** takes into account the correlation between the factors and the independent variable. It does all the work for you.
3. A variable might be insignificant once we have incorporated a more important predictor variable.

A common sense approach usually works well. If a variable never seems to be significant it typically isn't.

Model Prediction is the great equalizer!!

Example: Newfood Data

Goal of Experiment

- ▶ A six month market test has been performed on the [Newfood product](#).
A breakfast cereal.
- ▶ Build a multiple regression model that gives us good sales forecasts.
- ▶ This dataset is the outcome of a *controlled experiment* in which the values of the independent variables which affect sales are *chosen* by the analyst.

Example: Newfood Data

Analyses the factors which contribute to sales of a new breakfast cereal. Quantify the effects of business decisions such as choice of advertising level, location in store and pricing.

$p=5$

variable	description
y	sales new cereal sales
x_1	price price $\$$
x_2	adv low or high advertising (0 or 1)
x_3	locat bread or breakfast section (0 or 1)
x_4	inc neighborhood income
x_5	svol size of store

Example: Newfood

use all 5 predictors

1. What happens when you regress sales on price, adv, locat?
2. Run the "kitchen-sink" regression. Perform Diagnostic checks.
3. Which variables should we transform?
4. Run the new model. Perform diagnostics and variable selection.
5. What's the largest cooks distance?
6. Provide a summary of coefficients and statistical significance
7. Predict sales when price = 30, adv = 1, income = 8 and svol = 34.

What happens when you predict at the median values of the characteristics? /

might not work when p is large

Example: Newfood

First we examine the **correlation matrix**:

	sales	price	adv	locat	income
Count price	-0.658				
Count adv	0.001	0.000			
Count locat	-0.001	0.000	0.000		
Count income	0.163	-0.131	-0.746	0.000	
Count svol	0.375	-0.179	-0.742	-0.040	0.809

Remember: correlations are not β 's!!

Newfood

Total sales volume is negatively correlated to advertising.

Income is negatively correlated with advertising as well.

How is the negative correlation apt to affect the advertising effects?

	sales	price	adv
price	-0.658		
adv	0.001		
locat	-0.001	0.000 0.000	0.000

good design

There's no correlation in the X's by design!

Newfood model = $\ln(sales)$

SUMMARY (model)

Model = $\ln(sales) \sim$
price + adv + locat

Let's start by only including price, adv, locat

$$sales = 562 - 12.8 \text{ price} + 0.2 \text{ adv} - 0.2 \text{ locat}$$

Coefficients:

	Estimate	Std. Error	t value	P(> t)
β_0 (intercept)	562.31	53.14	10.58	0.000
β_1 price	-12.812	1.780	-7.20	0.000
β_2 adv	0.22	14.54	0.02	0.988
β_3 locat	-0.22	14.54	-0.02	0.988



H₀: $\beta_2 = 0$
cannot reject

- ▶ Why is the marketer likely to be upset by this regression?!
- ▶ Why is the economist happy?

Let's add income and svol to the regression!

$$P(X > t \text{ or } X < -t)$$



Transformation

log is defined for

values > 0

Power model: transform with log-log

$$\text{log(sales)} = 8.41 - 1.74 \text{ logprice} + 0.150 \text{ adv} + 0.0010 \text{ locat} - 0.524 \text{ loginc} + 1.03 \text{ logsvol}$$

Coefficients:

	Estimate	Std. Error	t value	P(> t)
(intercept)	8.407	1.387	6.06	0.000 ✓
logprice	-1.7430	0.2207	-7.90	0.000 ✓
adv	0.1496	0.1005	1.49	0.141 ✓
locat	0.00100	0.06088	0.02	0.997 ✓
loginc	-0.5241	0.4958	-1.06	0.294
logsvol	1.0308	0.2553	4.04	0.000 ✓

It works with
85%

Why no logs for adv and locat variables?

The log(svol) coefficient is close to one!

$R^2 = 60\%$

Transformation

exp is opposite to
log

$$\exp(\log(a)) = a$$

On the transformed scale,

$$\log \text{sales} = 8.41 - \underline{1.74 \log \text{price}} + 0.150 \text{adv} + 0.001 \text{locat} - \underline{0.524 \log \text{inc}} + \underline{1.03 \log \text{svol}}$$

On the un-transformed scale,

$$\text{sales} = e^{8.41} (\text{price})^{-\underline{1.74}} e^{0.15 \text{adv}} e^{\underline{0.001} \text{locat}} (\text{inc})^{-\underline{0.524}} (\text{svol})^{\underline{1.03}}$$

sales/price, income and svol are a **power** sales/adv, locat are
exponential

Interpretation

$$\text{Sales} = \dots \cdot e^{0.15 \cdot \text{adv}} \dots$$

$$\text{adv}: 0, 1$$

$$\text{adv} = 0 \Rightarrow e^{0.15 \cdot 0} = 1$$

$$\text{adv} = 1 \Rightarrow e^{0.15 \cdot 1} = 1.16$$

Interpret your regression model as follows

- ▶ Price elasticity is $\hat{\beta}_{\text{price}} = -1.74$. A 1% increase in price will drop sales 1.74%

- ▶ $\text{adv} = 1$ increases sales by a factor of $e^{0.15} = 1.16$. That's a 16% improvement

Variable Selection: delete locat as its *statistically insignificant*.

Kiue

Prediction

`predict.lm` provides a \hat{Y} -prediction given a new X_f

```
# predict.lm at newdata
> predict.lm(modelnew,newdata,se.fit=T,interval="prediction")
$fit
 $\hat{y}$  fit      lower    upper
lwr      upr
1 5.259691 4.739762 5.77962
$se.fit
[1] 0.05560662
```

95% PI

Exponentiate-back to find sales = $e^{5.2596} = 192.40$. *predicted sales.*

newdata=data.frame(price=30,adv=1,income=8,svol=34)

Interactions among input variables

$$\text{Sales} = \beta_0 + \beta_3 \text{Goog AD} + \boxed{\beta_4 \text{Goog AD} \cdot \text{FBAD}} + \beta_2 \text{FBAD}$$

interaction term

inter. effect

β_4 is signif

► Does gender change the effect of education on wages?

β_2 is not signif

► Do patients recover faster when taking drug A?

main effect

► How does advertisement affect price sensitivity?

Keep main effect

► Interactions are useful. Particularly with dummy variables.

► We build a kitchen-sink model with all possible dummies (day of the week, gender,...)

Marketing:

$$\text{Sales} = \beta_0 + \beta_1 \text{Goog AD} + \beta_2 \text{FBAD}$$

$$\text{Sales} = \beta_0 + (\beta_3 + \beta_4 \text{FBAD}) \text{Goog AD} + \beta_2 \text{FBAD}$$

$$\boxed{\underline{\beta_1} = \beta_3 + \beta_4 \text{FBAD}}$$

Models with Interactions

In many situations, X_1 and X_2 interact when predicting Y

Interaction Model: run the regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

In R: `model = lm(y ~ x1 * x2)` gives $X_1 + X_2 + X_1 X_2$

In R: `model = lm(y ~ x1 : x2)` gives only $X_1 X_2$

The coefficients β_1 and β_2 are marginal effects.

If β_3 is significant there's an interaction effect.

We leave β_1 and β_2 in the model whether they are significant or not.

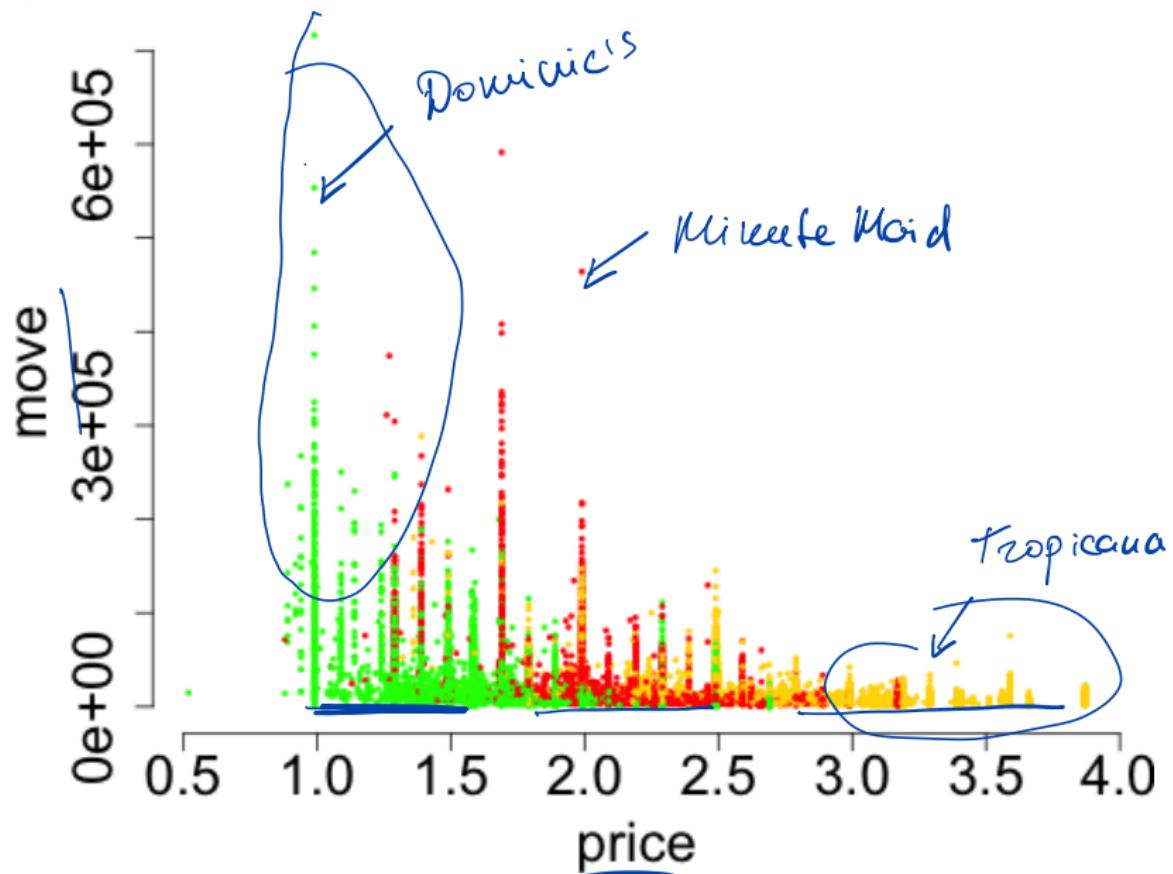
Orange Juice



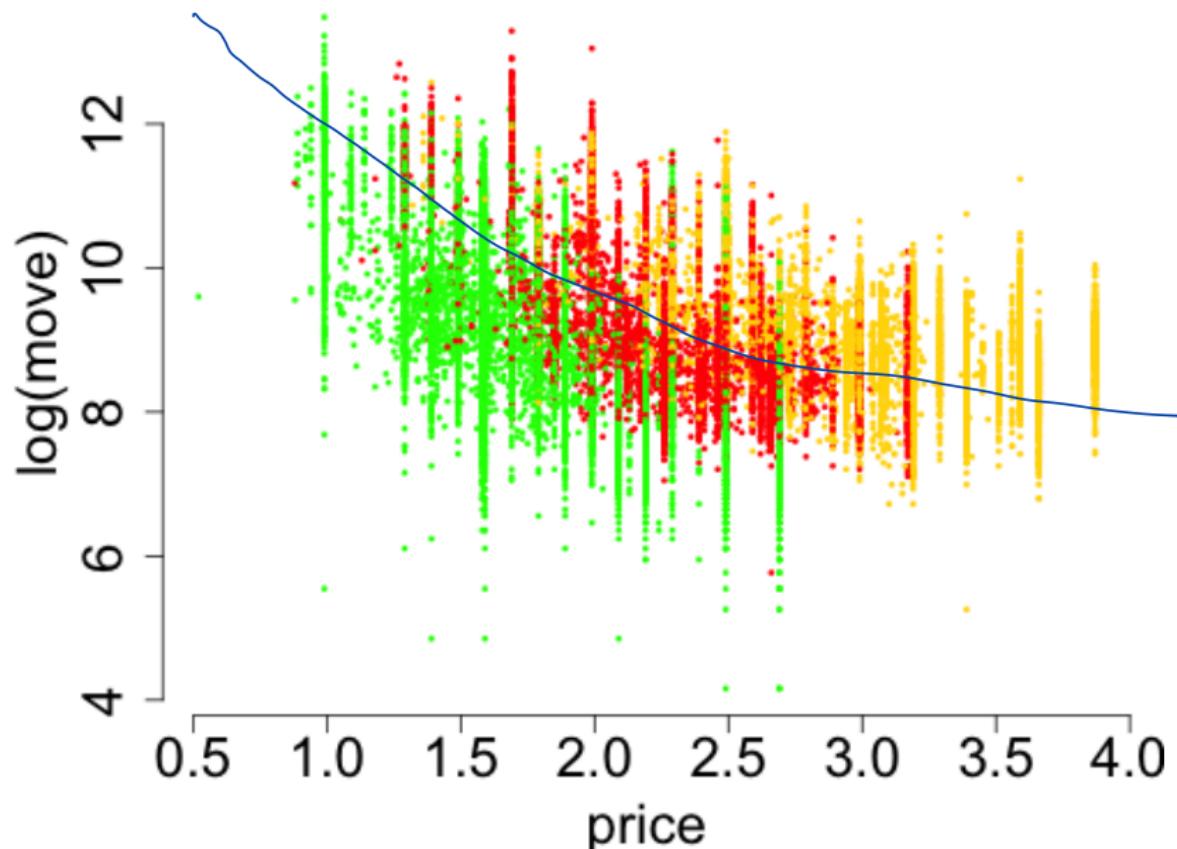
Dominic's

- ▶ 83 Chicagoland Stores (Demographic info for each)
 y x_1
- ▶ Price, sales (log units moved), and whether advertised (feat)
 \underline{y} \underline{x}_1 x_2

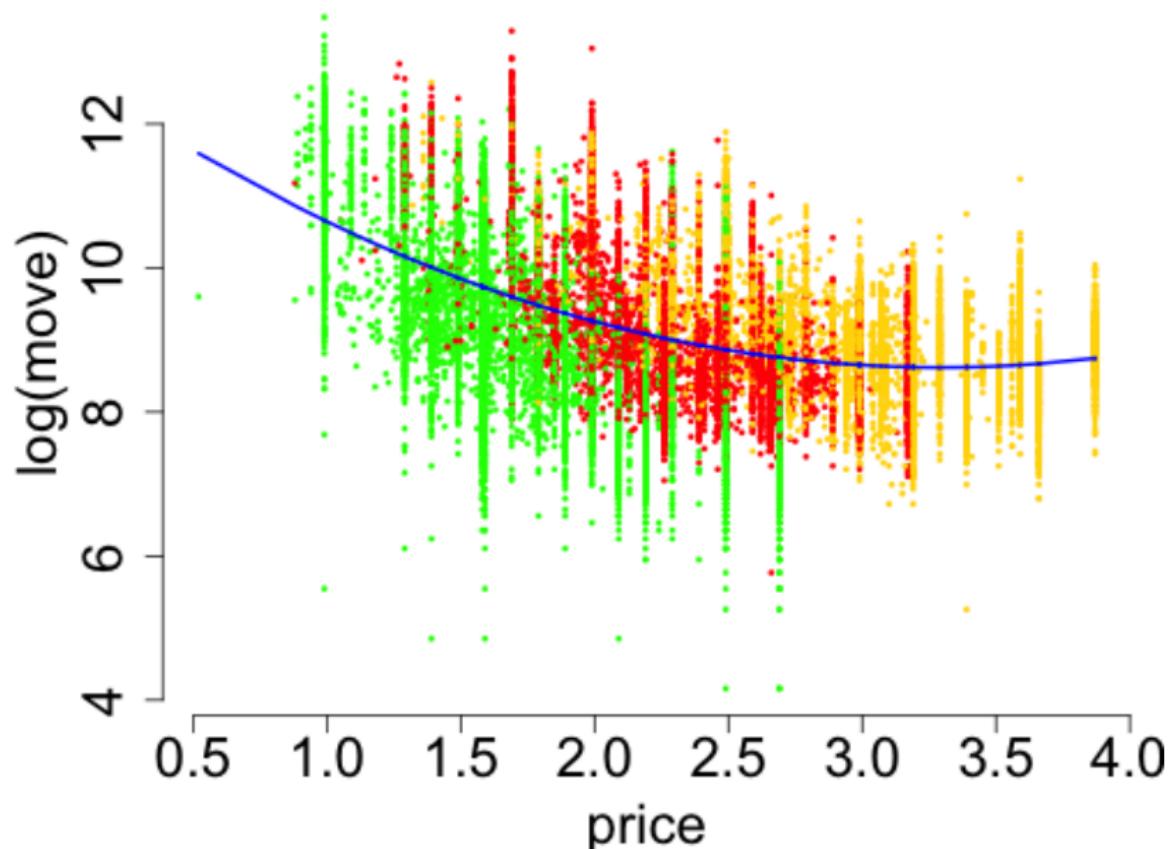
Orange Juice: Price vs Sales



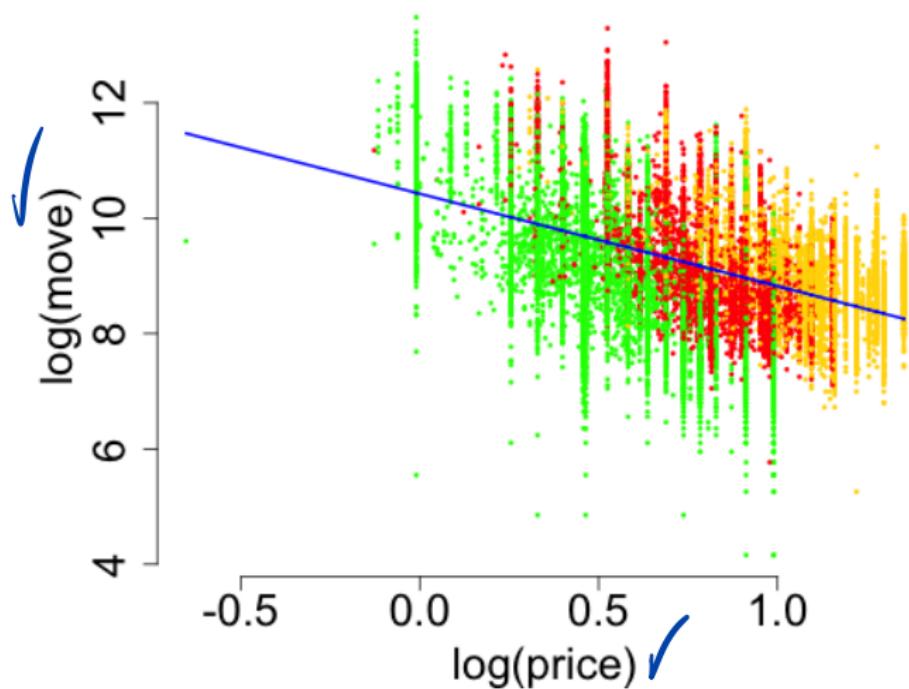
Orange Juice: Price vs log(Sales)



Orange Juice: Price vs log(Sales)



Orange Juice: $\log(\text{Price})$ vs $\log(\text{Sales})$



Why? Multiplicative (rather than additive) change.

How does advertisement affect price sensitivity?

Original model

$$\log(\text{sales}) = \beta_0 + \beta_1 \log(\text{price}) + \beta_2 \text{feat}$$

price elasticity

feat goes from 0 to 1
Price elasticity goes from β_3 to $\beta_3 + \beta_4$

If we feature the brand (in-store display promo or flyer ad), does it affect price sensitivity β_1 ? If we assume it does

$$\beta_1 = \beta_3 + \beta_4 \text{feat}$$

feat k price interact

The new model is

$$\begin{array}{ll} \text{feat} = 0: & \text{interaction} = 0 \\ \text{feat} = 1: & \text{interaction} = \beta_4 \cdot \log(\text{price}) \end{array}$$

$$\log(\text{sales}) = \beta_0 + (\beta_3 + \beta_4 \text{feat}) \log(\text{price}) + \beta_2 \text{feat}$$

$$\log(\text{sales}) = \beta_0 + (\beta_3 + \beta_4) \log(\text{price}) + \beta_2 \text{feat}$$

interaction

After expanding

$$\log(\text{sales}) = \beta_0 + \beta_3 \log(\text{price}) + \boxed{\beta_4 \text{feat} * \log(\text{price})} + \beta_2 \text{feat}$$

How does advertisement affect price sensitivity?

```
> print(lm(logmove) ~ log(price)*feat, data=obj)
```

+

main effects $\log(\text{price})$ & feat
to be included

Call:

```
lm(formula = logmove ~ log(price) * feat, data = obj) along with
```

interaction $\log(\text{price}) \cdot \text{feat}$

Coefficients:

(Intercept)	$\log(\text{price})$	feat	$\log(\text{price}) : \text{feat}$
-------------	----------------------	---------------	------------------------------------

9.6593

-0.9582

1.7144

-0.9773

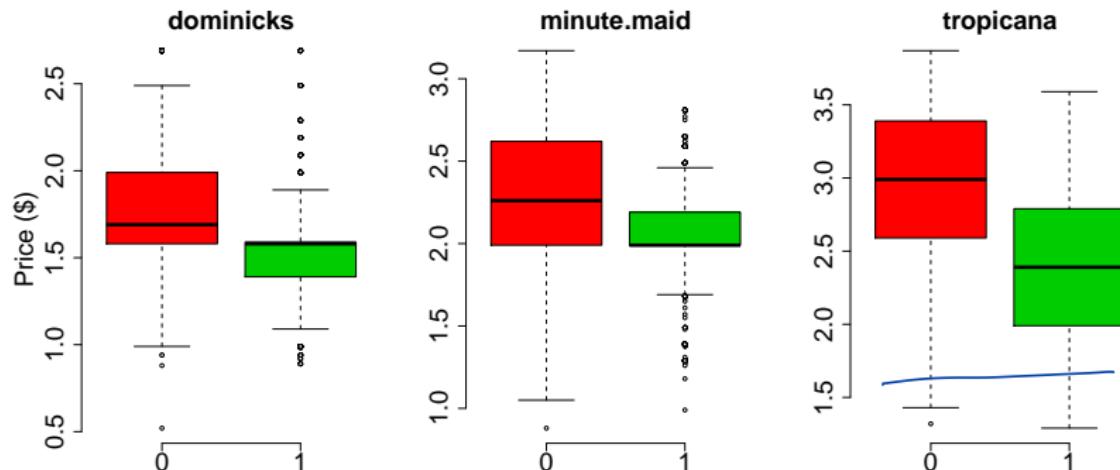
interaction

Advertisement increases price sensitivity from -0.96 to -0.958 - 0.98 = -1.94!

Why?

How does advertisement affect price sensitivity?

One of the reasons is that the price was lowered during the Ad campaign!



0 = not featured, 1 = featured

Dummies

We want to understand effect of the brand on the sales

$$\log(\text{sales}) = \beta_0 + \beta_1 \log(\text{price}) + \beta_2 \text{brand}$$

But brand is not a number!

How can you use it in your regression equation?

We introduce dummy variables

Brand	Intercept	brandminute.maid	brandtropicana
minute.maid	1	1	0
tropicana	1	0	1
dominicks	1	0	0

↙ intercept + effect of dominicks

$$\log(\text{sales}) = \beta_0 + \beta_1 \log(\text{price}) + \beta_{21} \text{brandminute.maid} + \beta_{22} \text{brandtropicana}$$

Effect of Brand on sales brand: Trop, MM,
Dom.

$$\text{Sales} = \beta_0 + \beta_1 \text{Price} + \beta_2 \text{brand}$$

Brand	Trop	MM	Dom
Trop	1	0	0
MM	0	1	0
Dom	0	0	1

Replace brand variable with 3 dummy variables Trop, MM, Dom

$$\text{Sales} = \beta_0 + \beta_1 \text{Price} + \beta_2 \text{Trop} + \beta_3 \text{MM} + \beta_4 \text{Dom}$$

In R: pick nominal brand = Dom β_2 is effect of
 Lump $\beta_0 + \beta_4$ is the new intercept Trop is relevance to
 Dom brand

Dummies

R will automatically do it for you

```
> print(lm(logmove ~ log(price)+brand, data=obj))
```

Call:

```
lm(formula = logmove ~ log(price) + brand, data = obj)
```

Coefficients:

(Intercept)	log(price)	brandminute.maid	brandtropicana
<u>10.8288</u>	<u>-3.1387</u>	<u>0.8702</u>	<u>1.5299</u>

$$\text{log(sales)} = \beta_0 + \beta_1 \text{log(price)} + \beta_3 \text{brandminute.maid} + \beta_4 \text{brandtropicana}$$

β_3 and β_4 are "change relative to reference" (dominicks here).

How does brand affect price sensitivity?

Price sensitivity

Interactions: $\log(\text{move}) \sim \log(\text{price}) * \text{brand}$

When you go

No Interactions: $\log(\text{move}) \sim \log(\text{price}) + \text{brand}$

from Dom to MM

Changes from -3.37 to

Parameter	Interactions	No Interactions
(Intercept)	10.95	10.8288
$\log(\text{price})$	-3.37	-3.1387
brandminute.maid	0.89	0.8702
brandtropicana	0.96239	1.5299
$\log(\text{price}): \text{brandminute.maid}$	0.057	$-3.37 \rightarrow -3.37 + 0.67$
$\log(\text{price}): \text{brandtropicana}$	0.67	

Main [

Trop buyers are least sensitive.

group 1 group 2 control for those
var.

$$y \text{ WLRP} = 40, 1 y$$

effect of a drug age pre-condition
 $\underline{y} = \beta_1 g + \underbrace{\beta_2 x_1 + \dots + \beta_p x_D}_{\text{don't control}}$ gender.

$$H_0: \beta_1 = 0$$

$$\beta_1: p\text{-val} = 0.1$$

Example: Golf Performance Data

Dave Pelz has written two best-selling books for golfers, *Dave Pelz's Short Game Bible*, and *Dave Pelz's Putting Bible*.

- ▶ Dave Pelz was formerly a “rocket scientist” (literally) Data analytics helped him refine his analysis It’s the short-game that matters!
- ▶ The optimal speed for a putt

Best chance to make the putt is one that will leave the ball 17 inches past the hole, if it misses.

Golf Data

2000

Year-end performance data on 195 players from the 2000 PGA Tour.

- 1. nevents, the number of official PGA events included in the statistics
- 2. money, the official dollar winnings of the player
- 3. drivedist, the average number of yards driven on par 4 and par 5 holes
- 4. gir, greens in regulation, measured as the percentage of time that the first (tee) shot on a par 3 hole ends up on the green, or the second shot on a par 4 hole ends up on the green, or the third shot on a par 5 hole ends up on the green
- 5. avgputts, which is the average number of putts per round.

Analyze these data to see which of **nevents**, **rivedist**, **gir**, **avgputts** is most important for winning money.

Golf Data

Regression of Money on all explanatory variables:

```
lm(formula = money ~ nevents + drivelist + gir + avgputts)
```

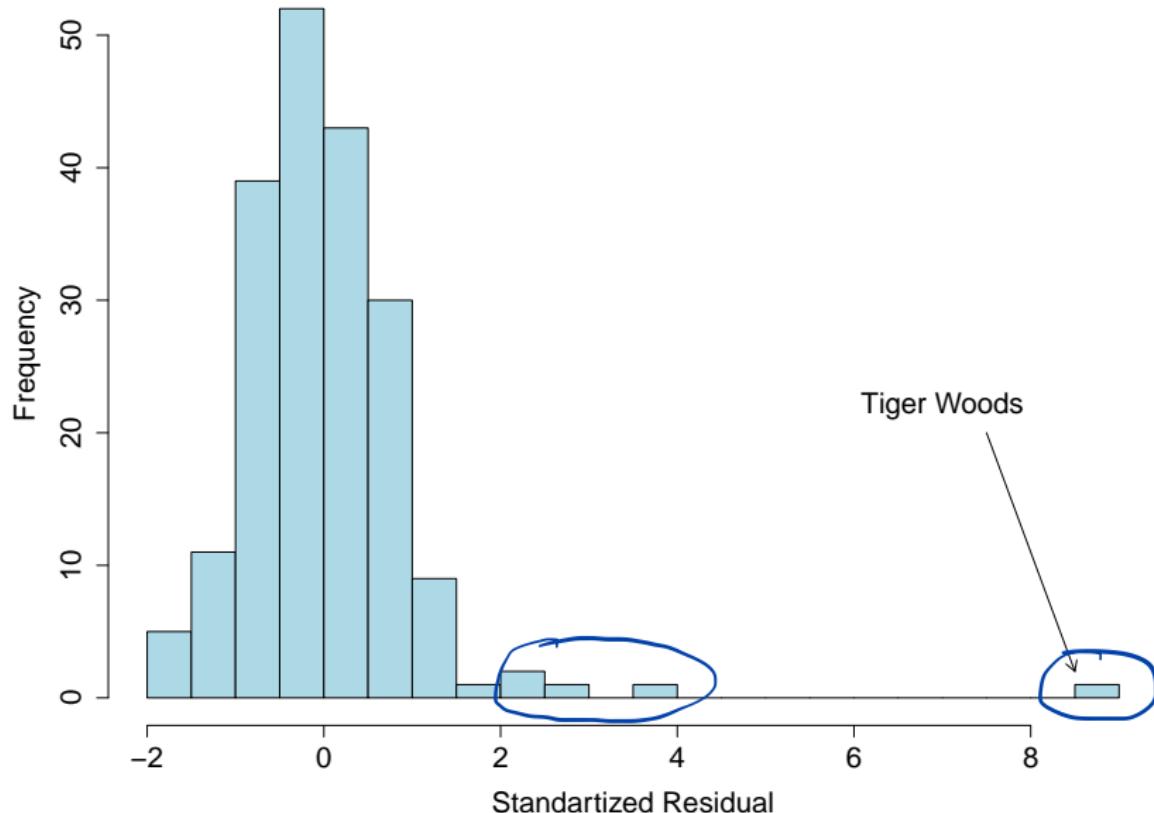
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
--	----------	------------	---------	----------

(Intercept)	<u>14856638</u>	4206466	3.532	0.000518 ***
nevents	<u>-30066</u>	11183	-2.689	0.007815 **
drivelist	<u>21310</u>	6913	3.083	0.002358 **
gir	<u>120855</u>	17429	6.934	6.22e-11 ***
avgputts	<u>-15203045</u>	2000905	-7.598	1.33e-12 ***

$$R^2 = 50\%$$

Residuals



Regression

Transform with $\log(\text{Money})$ as it has much better residual diagnostic plots.

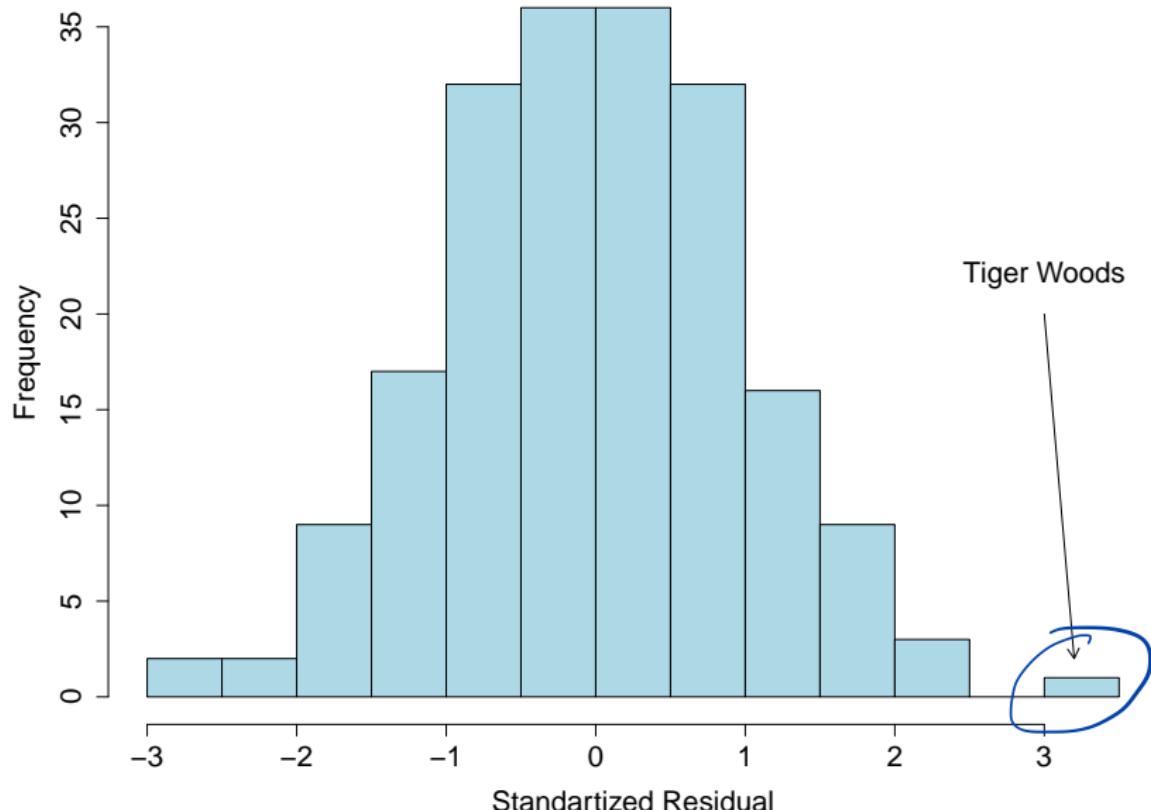
```
lm(formula = log(money) ~ nevents + drivelist + gir + avgputts, data = c)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	36.149228	3.577630	10.104	<2e-16 ***
nevents	-0.008987	0.009511	-0.945	<u>0.3459</u>
drivelist	0.014091	0.005880	2.397	0.0175 *
gir	0.165672	0.014824	11.176	<2e-16 ***
avgputts	-21.128752	1.701784	-12.416	<2e-16 ***

$R^2 = 67\%$. There's still 33% of variation to go

Residuals for log(Money)



Regression

Variable selection: t -stats for nevents is < 1.5 .

```
lm(formula = log(money) ~ drivedist + gir + avgputts, data = d00)
```

Residuals:

Transform + Variable Selection

Min	1Q	Median	3Q	Max
-1.48002	-0.37038	0.00079	0.40227	1.96546

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	36.17370	3.57653	10.114	<2e-16	***
drivedist	0.01463	0.00585	2.501	0.0132	*
gir	0.16577	0.01482	11.186	<2e-16	***
avgputts	<u>-21.36844</u>	1.68230	-12.702	<2e-16	***

The fewer the putts the better golfer you are. Duh!

avgputts per round is hard to decrease by one!

Evaluating the Coefficients

1. Greens in Regulation (GIR) has a $\hat{\beta} = 0.17$. If I can increase my GIR by one, I'll earn $e^{0.17} = 1.18\%$ An extra 18%
2. DriveDis has a $\hat{\beta} = 0.014$. A 10 yard improvement, I'll earn $e^{0.014 \times 10} = e^{0.14} = 1.15\%$ An extra 15%

Caveat: Everyone has gotten better since 2000!

if avg puts go down by 1
 $e^{-0.136}$

Main Findings

Tiger was 9 standard deviations better than the model.

- ▶ Taking logs of Money helps the residuals!
- ▶ An exponential model seems to fit well. The residual diagnostics look good
- ▶ The t -ratios for nevents are under 1.5.

Over-Performers

Outliers: biggest over and under-performers in terms of money winnings, compared with the performance statistics.

Woods, Mickelson, and Els won major championships by playing well when big money prizes were available.

Name	Money	Predicted	Error
Tiger Woods	9,188,321	3,584,241	5,604,080
Phil Mickelson	4,746,457	2,302,171	2,444,286
Ernie Els	3,469,405	1,633,468	1,835,937
Hal Sutton	3,061,444	1,445,904	1,615,540

Under-Performers

Underperformers are given by large negative residuals Glasson and Stankowski should win more money.

Name	Money	Predicted	Error
Kenny Perry	889,381	<u>1,965,740</u>	-1,076,359
Paul Stankowski	669,709	1,808,690	-1,138,981
Bill Glasson	552,795	1,711,530	-1,158,735
Jim McGovern	266,647	1,397,818	-1,131,171

Lets look at 2018 data

Highest earners are

name	nevents	money	drivedist	gir	avgputts
Justin Thomas	23	8,694,821	311.800	68.770	1.714
Dustin Johnson	20	8,457,352	314	70.570	1.699
Justin Rose	18	8,130,678	303.500	69.950	1.732
Bryson DeChambeau	26	8,094,489	305.700	69.650	1.758
Brooks Koepka	17	7,094,047	313.400	68.280	1.747
Bubba Watson	24	5,793,748	313.100	68.210	1.773

Overperformers

name	money	Predicted	Error
Justin Thomas	8,694,821	5,026,220	3,668,601
Dustin Johnson	8,457,352	6,126,775	2,330,577
Justin Rose	8,130,678	4,392,812	3,737,866
Bryson DeChambeau	8,094,489	3,250,898	4,843,591
Brooks Koepka	7,094,047	4,219,781	2,874,266
Bubba Watson	5,793,748	3,018,004	2,775,744
Webb Simpson	5,376,417	2,766,988	2,609,429
Francesco Molinari	5,065,842	2,634,466	2,431,376
Patrick Reed	5,006,267	2,038,455	2,967,812
Satoshi Kodaira	1,471,462	-1,141,085	2,612,547

Underperformers

name	money	Predicted	Error
Trey Mullinax	1,184,245	3,250,089	-2,065,844
J.T. Poston	940,661	3,241,369	-2,300,708
Tom Lovelady	700,783	2,755,854	-2,055,071
Michael Thompson	563,972	2,512,330	-1,948,358
Matt Jones	538,681	2,487,139	-1,948,458
Hunter Mahan	457,337	2,855,898	-2,398,561
Cameron Percy	387,612	3,021,278	-2,633,666
Ricky Barnes	340,591	3,053,262	-2,712,671
Brett Stegmaier	305,607	2,432,494	-2,126,887

Let's Look at 2020

name	drivedist	gir	avgputts	residual
Bryson DeChambeau	344.4	71.53	1.748	<u>1,658,171</u>
Jason Kokrak	309.7	68.65	1.676	1469110
Matthew Wolff	314.4	64.24	1.659	1407259
Patrick Cantlay	303.1	68.06	1.75	1406472
Xander Schauffele	304.9	68.52	1.669	1207284
Martin Laird	299.8	78.57	1.768	1006939
Justin Thomas	301.3	63.43	1.679	858123

Standard deviation of the residual is 334,595. DeChambeau is 5 sigmas away
from the average PGA player!

Findings

Here's three interesting effects:

- ▶ Tiger Woods is 8 standard deviations better!
- ▶ Increasing driving distance by 10 yards makes you 15% more money
- ▶ Increasing GIR by one makes you 18% more money.
- ▶ Detect Under- and Over-Performers

Go Play!!

Regression

1. Input and Plot Data In R: `plot` and `summary` commands
2. "Kitchen-Sink" Regression `lm` command with all variables
3. Residual Diagnostics and `plot(model)` Fitted values and Standardised residuals. Outliers and Influence
4. Transformation?
Correct the 4-in-1 plots and assumptions.
5. VIF. selection

Regression Strategy

1. Variable Selection t -state and p -values from `summary(model)`
2. Final Regression Re-run the model. Interpret the coefficients
`summary(model)`. Economic and Statistical Significance
3. Prediction `predict.lm`. Out-of-sample forecasting A model is only as good as its predictions!!

Machine Learning Tools

There's the list of methods we'll go through

1. Linear Regression ✓
2. Multiple Regression ✓
3. K-Nearest Neighbor Now
4. Simple Tree
5. Random Forests/Bagging
6. Boosting
7. Classification
8. Deep Learning
Nonlinearity. Keras.

Focus is prediction

Next week

Can do it up

Logistic Regression

Support Vector Machine (SVM)

} final lecture

Metrics of how well model predicts:

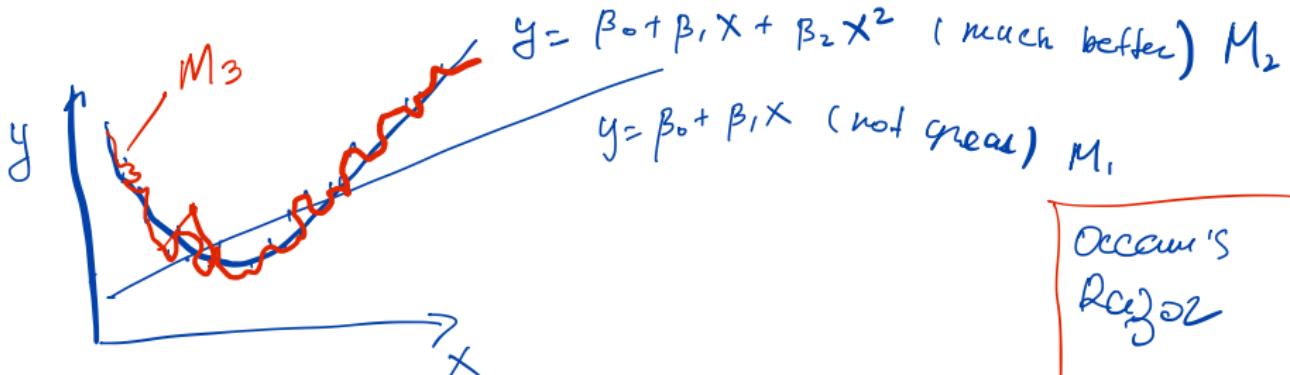
Dataset \Rightarrow Fit model \Rightarrow compare predicted values at $y(\hat{y})$ & observed ones

$$\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 = \text{MSE} \quad (\text{mean squared error})$$

$$\sqrt{\text{MSE}} = \boxed{\text{RMSE}} \quad \text{most popular}$$

Root mean squared error

$$\frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| = \text{MAPE} \quad (\text{mean absolute percent error})$$



M_2 is more complex than M_1

M_3 is more complex than M_1 & M_2 if it fits data better

RMSF

High

M_1

middle

too complex
 M_3 Overfitting

lowest

- is sensitive to noise
- Models noise & not the important pattern

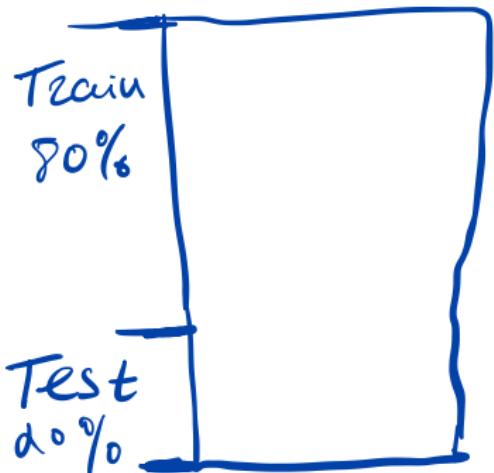
How to choose between M_1 , M_2 & M_3

Task is to predict.

Calculate RMSE on a new dataset

split data

Test data set



Model with lowest
RMSE on test data set
is the best

Trick: Re-order (shuffle)
rows before splitting

Boston Housing Prices

Boston Housing Data (MASS package in R).

14 features (columns) and 506 observations (rows).

- ▶ CRIM - per capita crime rate by town
- ▶ ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- ▶ INDUS - proportion of non-retail business acres per town.
- ▶ CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- ▶ NOX - nitric oxides concentration (parts per 10 million)
- ▶ RM - average number of rooms per dwelling
- ▶ AGE - proportion of owner-occupied units built prior to 1940
- ▶ DIS - weighted distances to five Boston employment centres
- ▶ RAD - index of accessibility to radial highways
- ▶ TAX - full-value property-tax rate per \$10,000
- ▶ PTRATIO - pupil-teacher ratio by town
- ▶ B - $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
- ▶ LSTAT - % lower status of the population
- ▶ MEDV - Median value of owner-occupied homes in \$1000's

g

Predicting Boston Housing Prices

Here we fit different model to the Boston Housing Data, which is available in the MASS package of R and it has 14 features (columns) and 506 observations (rows).

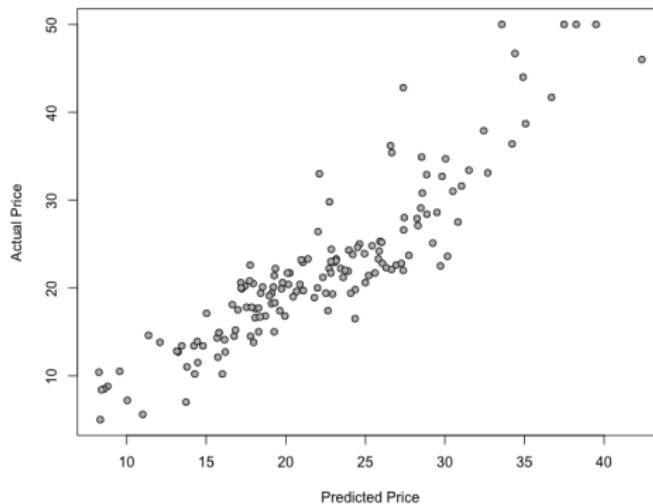
- ▶ Variable to predict: MEDV (median value of owner-occupied homes in 1000s).
- ▶ Features include CRIM (per capita crime rate), DIS (distance to Boston employment centers), RM (average number of rooms per dwelling), LSTAT (percent of population with lower socio-economic status), among others

Multiple Regression

- ▶ Kitchen sink LM: $y = \text{MEDV}$, rest are independent variables. RMSE = 4.38
- ▶ Next we try

$$\log(\text{MEDV}) \sim \text{CRIM} + \text{CHAS} + \text{NOX} + \text{RM} + \text{DIS} + \text{PTRATIO} + \text{RAD} + \text{B} + \text{LSTAT}$$

The RMSE here is 4.18



k-Nearest Neighbors

Med V k LStat
\$ %

Points that are "closer" to the place I am trying to predict should be more relevant...

How about averaging the closest 20 neighbors?

What do I mean by closest? We will choose the 20 points that are closest to the X value we are trying to predict.

This is what is called the k-nearest neighbors algorithm

506 obs.

Med V Lstat Lstat ^{↓ sort}
0.1
0.12
⋮

Lstat = 9%

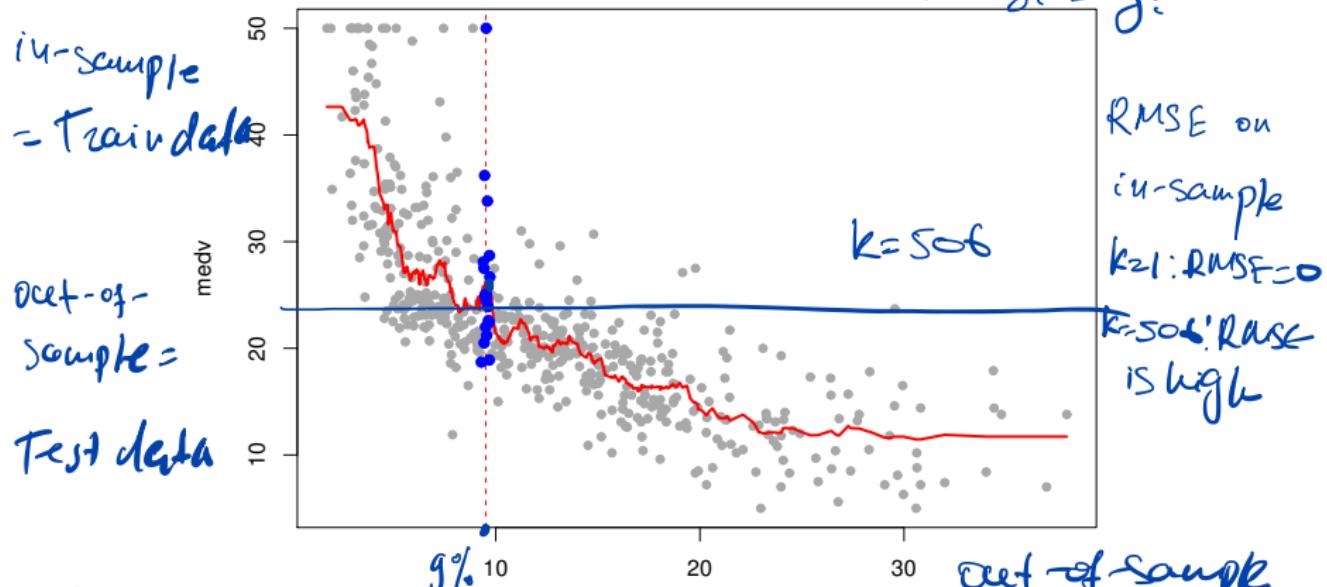
first 20 rows k avg. Med V values

k-Nearest Neighbors

How many obs. to avg.

How to choose k ?

$$k=20 \quad k=506; \bar{g}_i = \bar{g}$$
$$k=1; \bar{g}_i = g_i$$

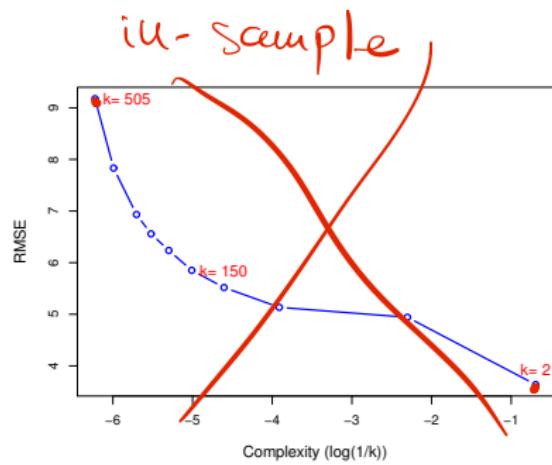


Try all possible values of k_{stat}
& calculate out-of-sample RMSE

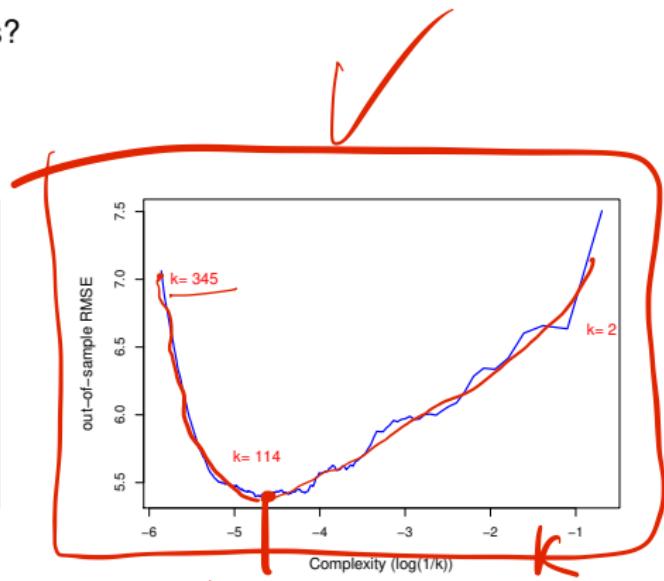
out-of-sample
RMSE: $k=1$
 $k=506$

k-nearest neighbors

What is the accuracy of different models?



In-Sample RMSE

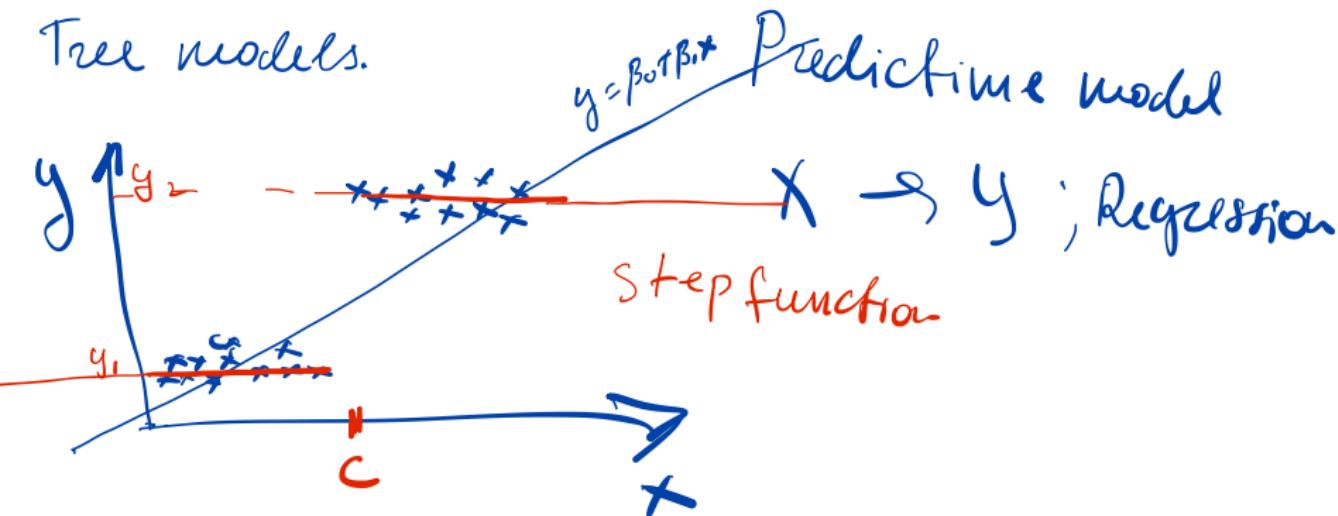


$k=46$

Out-of-sample RMSE

Now, the model where $k = 46$ looks like the most accurate choice!!

Tree models.



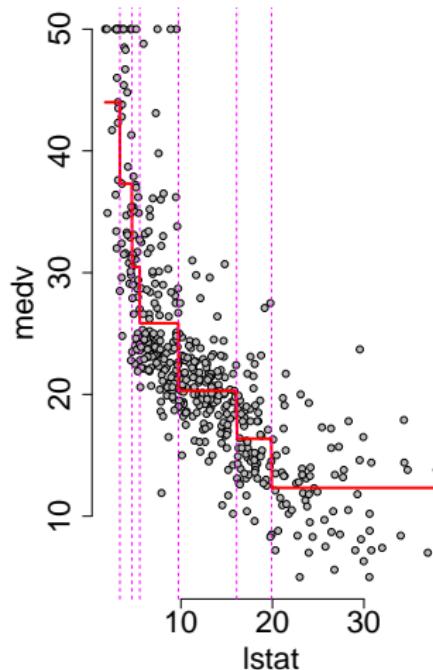
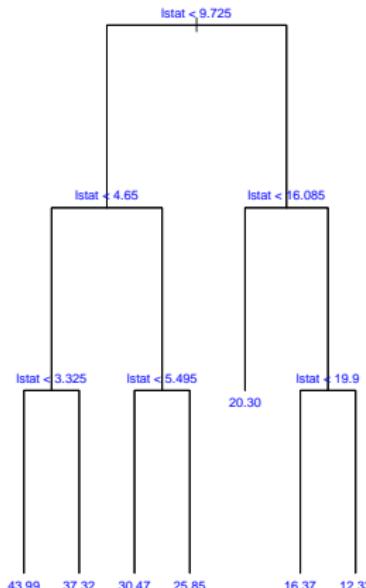
if $x < c$ then $\hat{y} = y_1$

if $x > c$ then $\hat{y} = y_2$

Tree Models: Random Forests and XGBoost

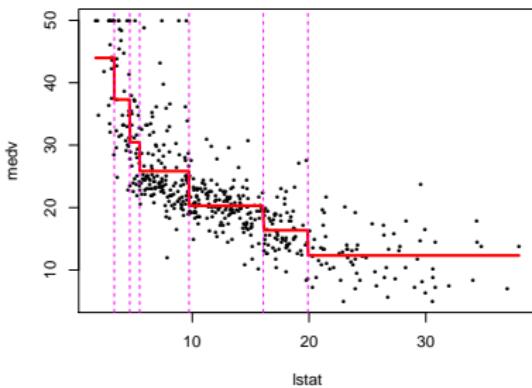
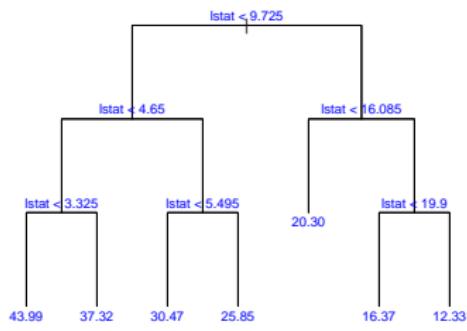
Tree = piecewise regression (a.k.a step function).

Categorical and numeric y and x very nicely and is fast
The leaves of the tree have our best prediction ...



Regression Trees

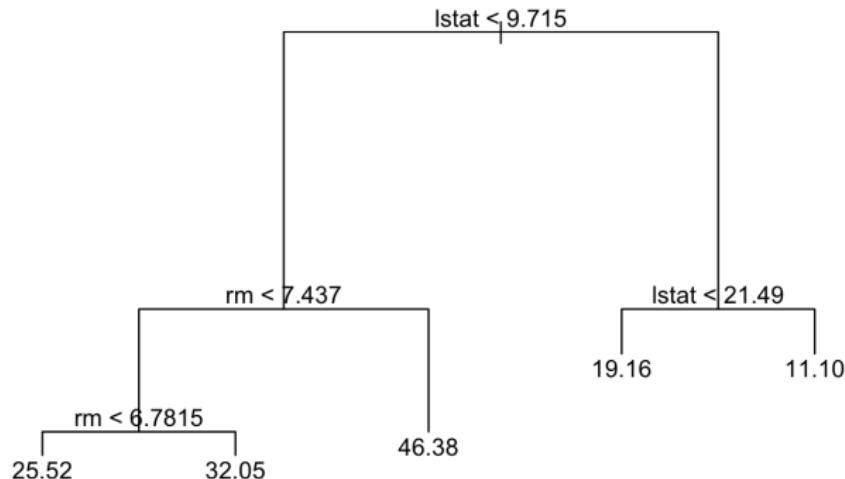
At left is the tree fit to the data. At each interior node there is a decision rule of the form $\{x < c\}$. If $x < c$ you go left, otherwise you go right. Each observation is sent down the tree until it hits a bottom node or leaf of the tree.



The set of bottom nodes gives us a partition of the predictor (x) space into disjoint regions. At right, the vertical lines display the partition. With just one x , this is just a set of intervals.

Regression Trees

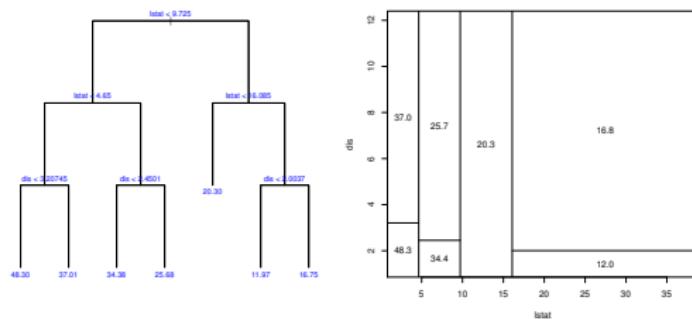
Use 10-fold cross-validation. Set number of leaves to be 5 in pruning. Build a tree in the shape of:



This tree model achieves an out-of-sample MSE of 5.01.

Regression Trees

Here is a tree with $x = (x_1, x_2) = (\text{lstat}, \text{dis})$ and $y = \text{medv}$. Now the decision rules can use either of the two x ?s.

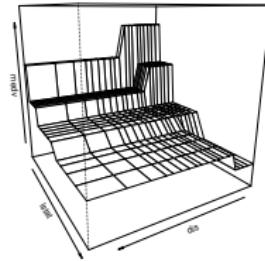


At right is the partition of the x space corresponding to the set of bottom nodes (leaves). The average y for training observations assigned to a region is printed in each region and at the bottom nodes.

Regression Trees

This is the regression function given by the tree.

It is a step function which can seem dumb, but it delivers non-linearity and interactions in a simple way and works with a lot of variables.



Notice the interaction.

The effect of dis depends on lstat!!

Bagging What's wrong with bagging?

- Generate B samples using bootstrap
- Train B trees that are similar.

Treat the sample as if it were the population and then take iid draws.

That is, you sample with replacement so that you can get the same original sample value more than once in a bootstrap sample.

To Bootstrap Aggregate (Bag) we:

- ▶ Take B bootstrap samples from the training data, each of the same size as the training data.
- ▶ Fit a large tree to each bootstrap sample (we know how to do this fast!). This will give us B trees.
- ▶ Combine the results from each of the B trees to get an overall prediction.

Random Forest



remove
randomly
all but k
columns

$K = \sqrt{P}$
 $P = \# \text{cols in original data}$

Each tree is less
similar to others.

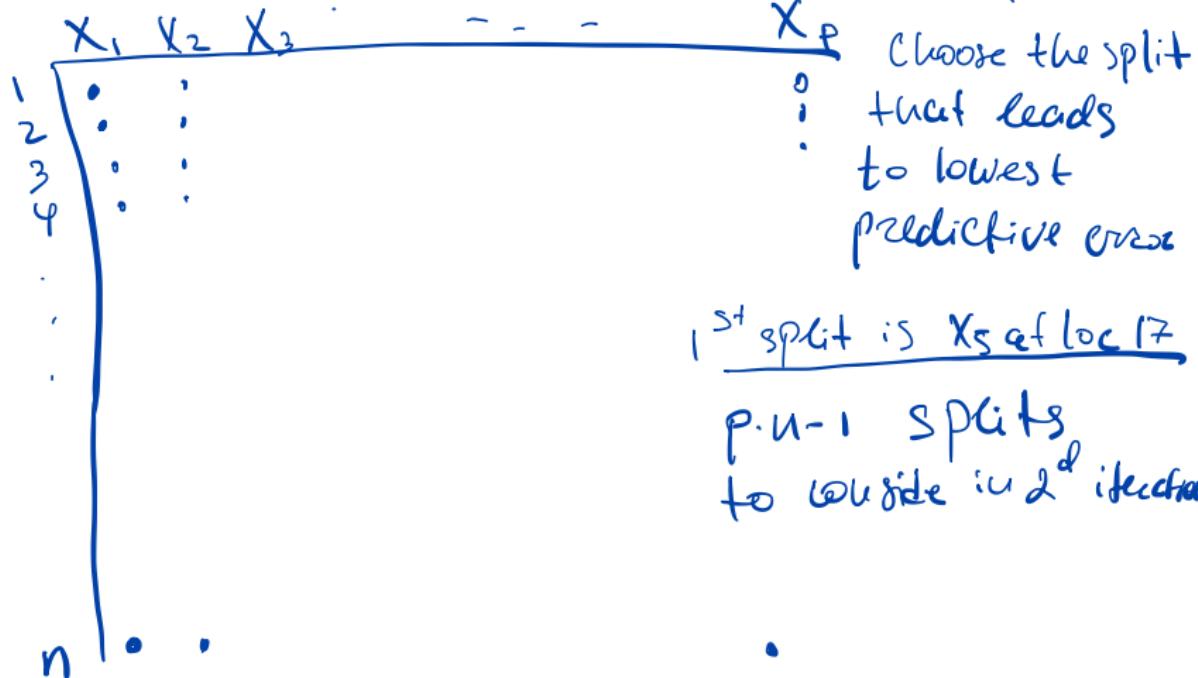
Very Practical!
Used often

How to fit a tree?

p inputs

CART algorithm

n observations



Bagging and Random Forest

- ▶ For numeric y we can combine the results easily by making our overall prediction the average of the predictions from each of the B trees.
- ▶ For categorical y , it is not quite so obvious how you want to combine the results from the different trees.
- ▶ Often people let the trees vote: given x get a prediction from each tree and the category that gets the most votes (out of B ballots) is the prediction.
- ▶ Alternatively, you could average the \hat{p} from each tree. Most software seems to follow the vote plan.

Random Forest and Bagging

Include all 13 predictors for each split of the tree (a.k.a bagging)

Achieves an out-of-sample MSE of 3.66.

After we limit the number of predictors to be 6, we can achieve an even lower MSE of 3.35.

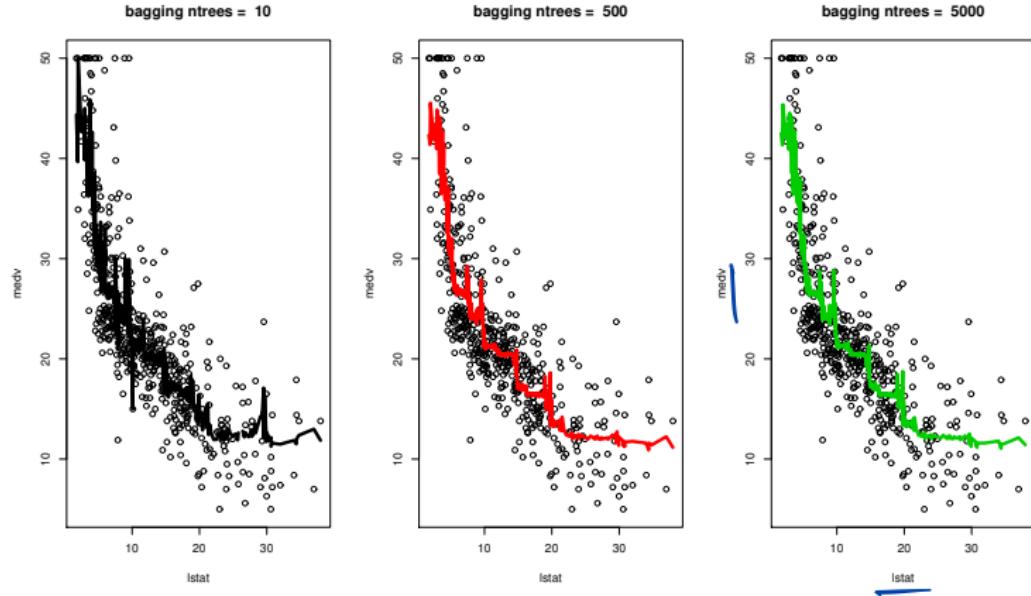
Random Forest beats Bagging.

Bagging and Random Forest

With 10 trees our fit is too jumbly.

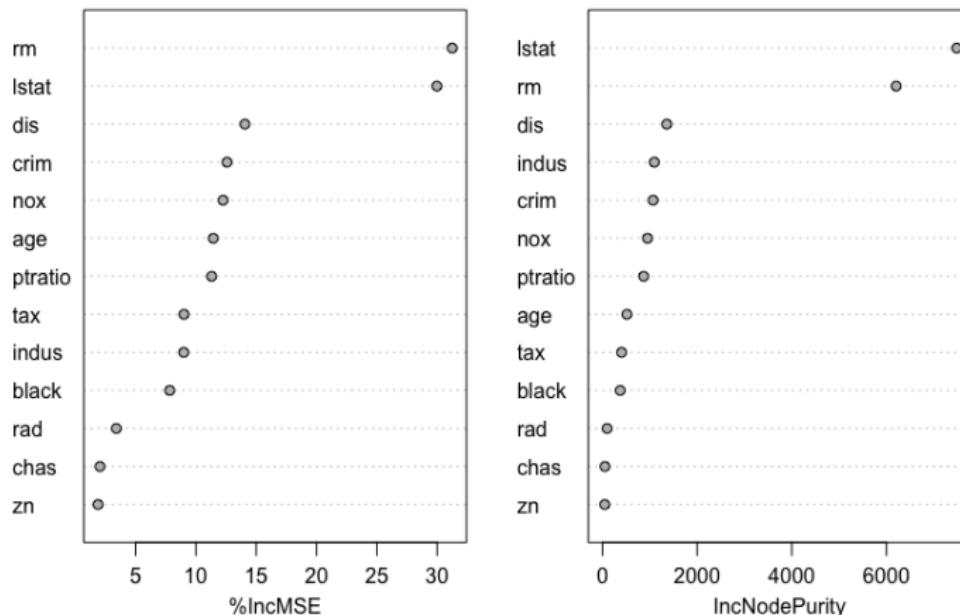
With 1,000 and 5,000 trees the fit is not bad and very similar.

Note that although our method is based multiple trees (average over) so we no longer have a simple step function!!



Random Forest and Bagging

Use an importance function to check the effect of each variable:



Across all trees in random forest, Istat (the wealth level) and rm (house size) are by far the two most important variables.

Boosting: the most popular predictive model in the last 15 years.

1. Fit Shallow Tree m_1 into data. say with 5 terminating nodes (Regions)
2. $z_i = m_1(x_i) - y_i$; $y_i = \underline{m_1(x_i)} - \underline{z_i}$
if I know z_i then I can get exactly y_i .

3. Build predictive model for z_i :
 $m_2(x_i) : x_i \rightarrow z_i$; m_2 is a shallow tree with 5 terminal nodes

$$\hat{y}_i = m_1(x_i) - m_2(x_i), \quad y_i - (m_1(x_i) - m_2(x_i)) = z_i^{(2)}$$

4. $m_3 : x_i \rightarrow z_i^{(2)}$ $\hat{y}_i = m_1(x_i) - m_2(x_i) - m_3(x_i)$
Keep doing until residuals have no pattern
that can be exploited.

Boosting

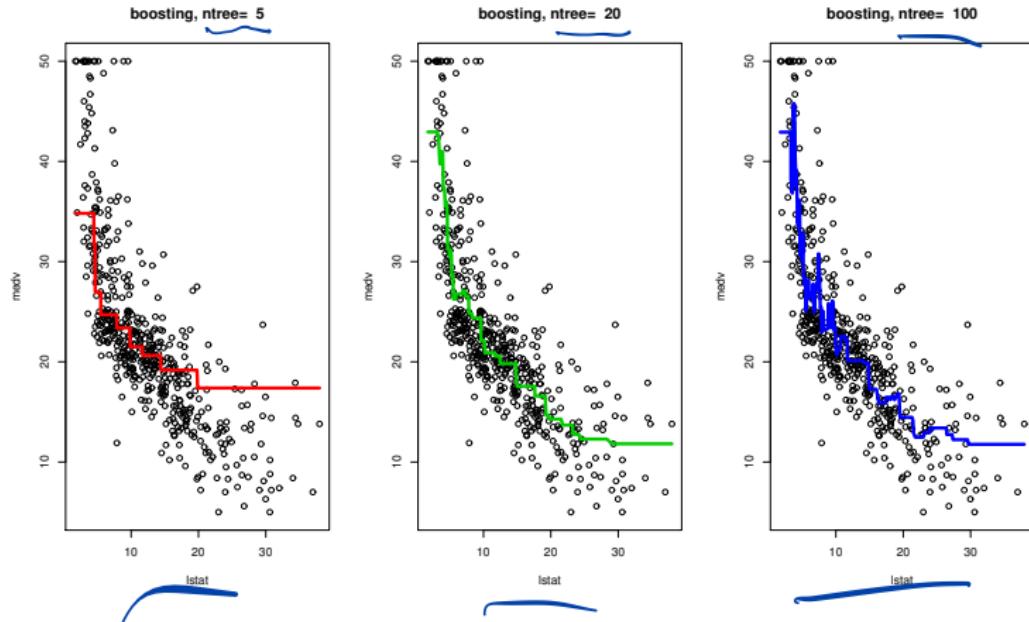
Like Random Forests, boosting is an ensemble method in that the overall fit it produced from many trees. The idea however, is totally different!!

In Boosting we:

- ▶ Fit the data with a single tree.
- ▶ Crush the fit so that it does not work very well.
- ▶ Look at the part of y not captured by the crushed tree and fit a new tree to what is “left over”.
- ▶ Crush the new tree. Your new fit is the sum of the two trees.
- ▶ Repeat the above steps iteratively. At each iteration you fit “what is left over” with a tree, crush the tree, and then add the new crushed tree into the fit.
- ▶ Your final fit is the sum of many trees.

Boosting

Here are some boosting fits where we vary the number of trees, but fix the depth at 2 (suitable with 1 x) and shrinkage = λ at .2.



Boosting

LM: β - marginal effect

Tree models are not interpr.

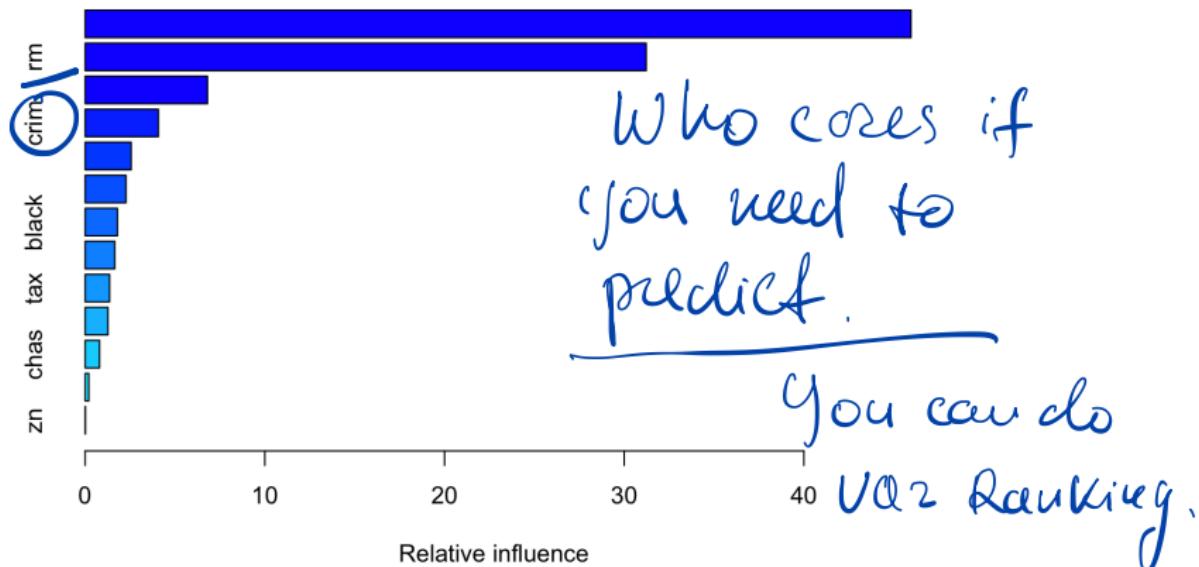
Train the Boosting model with 5000 trees and depth of 4,

Black Box

Out-of-sample MSE of 3.44, which is only slightly worse than Random Forest.

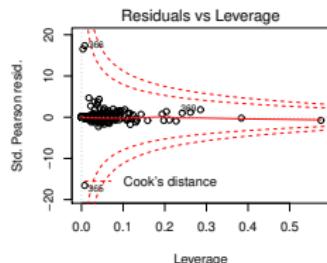
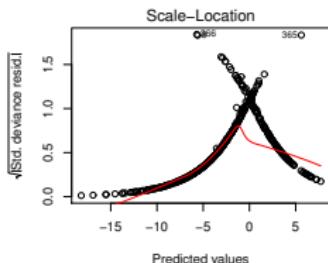
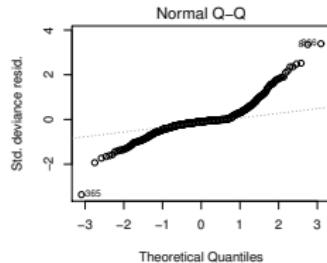
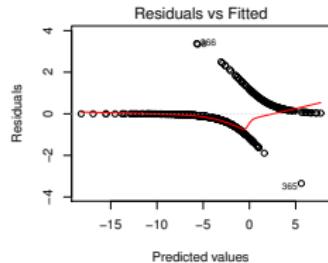
And we can still observe the significance of lstat and rm.

Models.



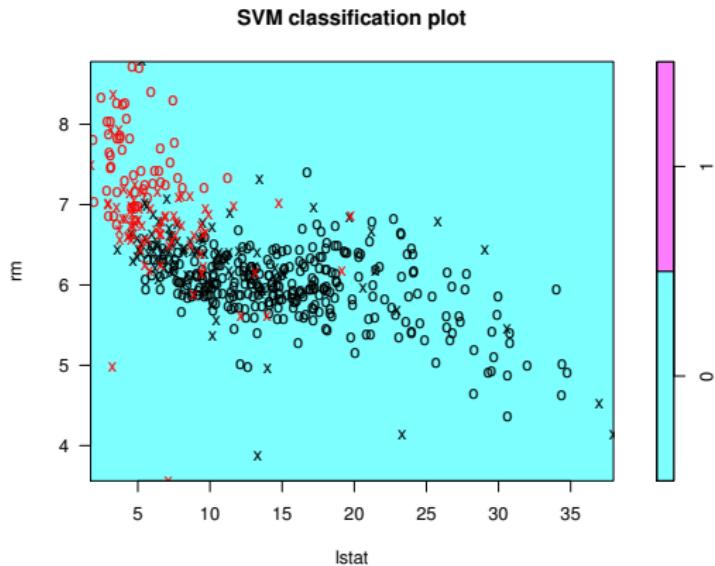
Classification: Logistic Regression

We classified the response as 1 and 0, based on $\text{medv} > 25$ and $\text{medv} \leq 25$. We then tried the logistic regression and give the diagnostic plot:



Classification: SVM

We include all the variables in the model, but the plot here only highlight the support vectors related to rm and lstat.



Deep Learning: Keras

First layer is dense with 200 neurons. Includes `input_shape` which gives the dimensionality of the input data. Then add a dense layer with just a single neuron to serve as the output layer.

Out-of-sample MSE of 11.47.

A dense DL model doesn't do particularly well probably due to over-fitting on such a small set.

Target

[Target](#) and other retailers use predictive analytics to study consumer purchasing behaviour to see what type of coupons or promotions you might like

Here's a famous story about a father and his daughter. Target predicted that his daughter was pregnant from her purchasing behaviour long before they were buying diapers

Here's the original link ...

[Target and Pregnancy](#)

Getting a customer to a routine is the key

- ▶ M.I.T experiment: t-shaped maze with chocolate at the end and behind the barrier that opens after a loud click
- ▶ While each animal wandered through the maze, its brain was working furiously
- ▶ As the scientists repeated the experiment, again and again, the rats eventually stopped sniffing corners and making wrong turns and began to zip through the maze with more and more speed
- ▶ As each rat learned how to complete the maze more quickly, its mental activity decreased

Learning routines from data is the basis for modern marketing

- ▶ Habits is a three-step loop: cue, a trigger (go into automatic mode), then the routine
- ▶ Febreze: original ads were targeting a wrong routine (kill the smell), no sails. They the ad said: use Febreze after cleaning each room. Now it is one of the most successful products.
- ▶ Target used the fact that customers who going through a major life event change their habits (routines). They can identify due dates from registry.

Walmart

Walmart began using predictive analytics in 2004. Mining trillions of bytes' worth of sales data from recent hurricanes

Determine what customers most want to purchase leading up to a storm.

Strawberry Pop-Tarts are one of the most purchased food items, especially after storms, as they require no heating and can be eaten at any meal

Walmart and Hurricanes

Germany's Otto

Otto sells other brands, does not stock those goods itself, hard to avoid one of the two evils: shipping delays until all the orders are ready for fulfilment, or lots of boxes arriving at different times.

- ▶ Analyze around 3 billion past transactions and 200 variables—past sales, searches on Otto's site and weather information. They predict what customers will buy a week before they order. This system has proved so reliable, predicting with 90% accuracy what will be sold within 30 days, that Otto allows it automatically to purchase around 200,000 items a month from third-party brands with no human intervention.

Economist

Germany's Otto

Stitch Fix CEO Says AI Is 'So Real' and Incredibly Valuable

Stitch Fix asks customers for insights and feedback alongside their size and color preference for items, even the ones customers didn't like or buy, in exchange for a clear value proposition.

The breadth and depth of their data are valuable.

Their model relies on a combination of data science – machine learning, AI and natural language processing – and human stylists; on top of complex customer profiles built by data, stylists can layer the nuances of buying and wearing clothes.

Uber Pool

Bayes predicts where you're going to be dropped off.

Uber constructs prior probabilities for riders, Uber cars, and popular places.

Combine to construct a joint probability table

Then calculate the posterior probability of destination for each person and pool travellers together

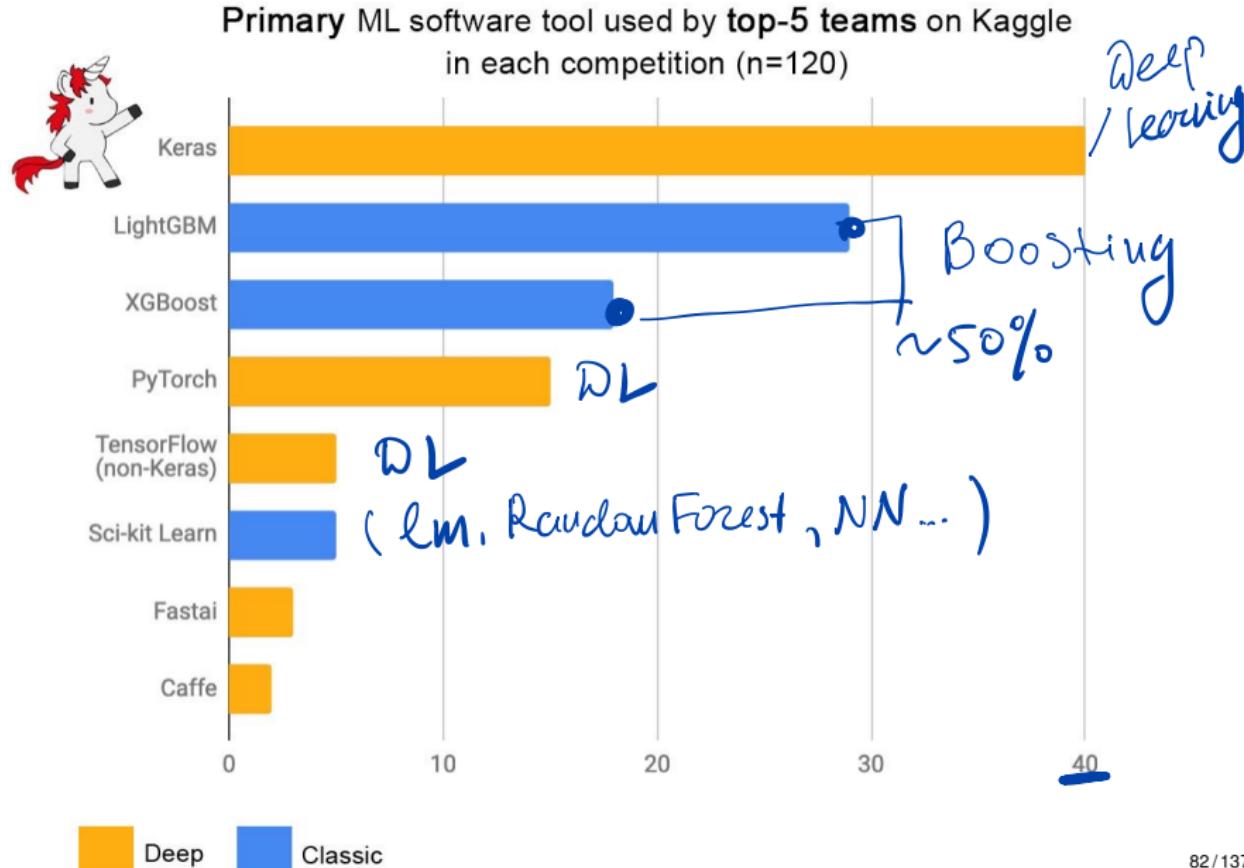
Uber Pool

Kaggle: Predictive Culture

353 Competitions			
	Passenger Screening Algorithm Challenge Improve the accuracy of the Department of Homeland Security's threat recognition algorithms <small>Featured · 2 years ago · 📰 terrorism, image data, object detection</small>	\$1,500,000	518 teams
	Zillow Prize: Zillow's Home Value Prediction (Zestimate) Can you improve the algorithm that changed the world of real estate? <small>Featured · 2 years ago · 📰 real estate, housing</small>	\$1,200,000	3,775 teams
	Data Science Bowl 2017 Can you improve lung cancer detection? <small>Featured · 2 years ago · 📰 healthcare, binary classification, image data</small>	\$1,000,000	1,972 teams
	Heritage Health Prize Identify patients who will be admitted to a hospital within the next year using historical claims data. (Ent..) <small>Featured · 6 years ago</small>	\$500,000	1,351 teams

Most frequently used predictive models

Kaggle
Netflix



Airbnb

Airbnb New User Bookings Prediction Competition New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries.

Accurately predict where a new user will book their first travel experience

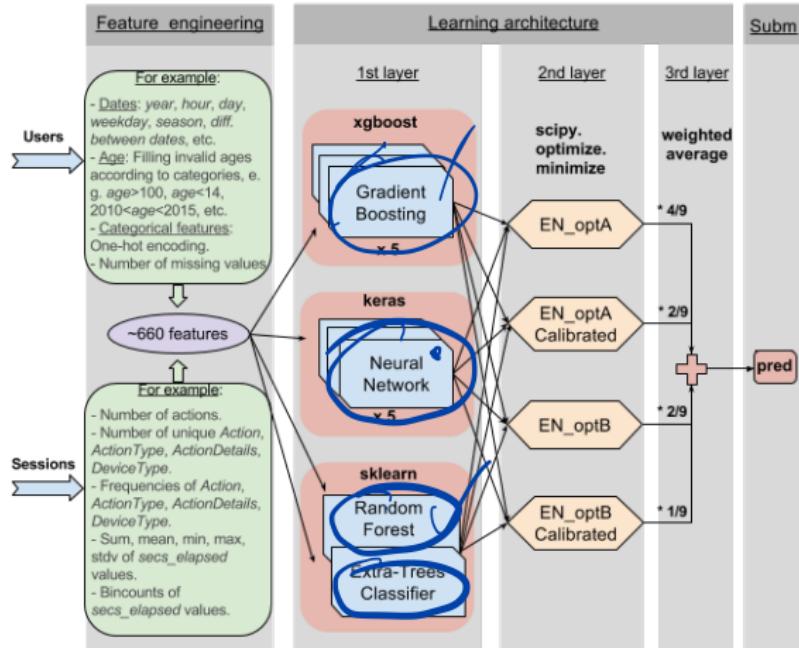
Airbnb can then personalize content, decrease the average time to first booking, and better forecast demand.

12 classes—major destinations, and a did not book category

$$w_1 M_1(x) + w_2 M_2(x) + \dots + w_B M_B(x)$$

$$\sum_{i=1}^B w_i = 1$$

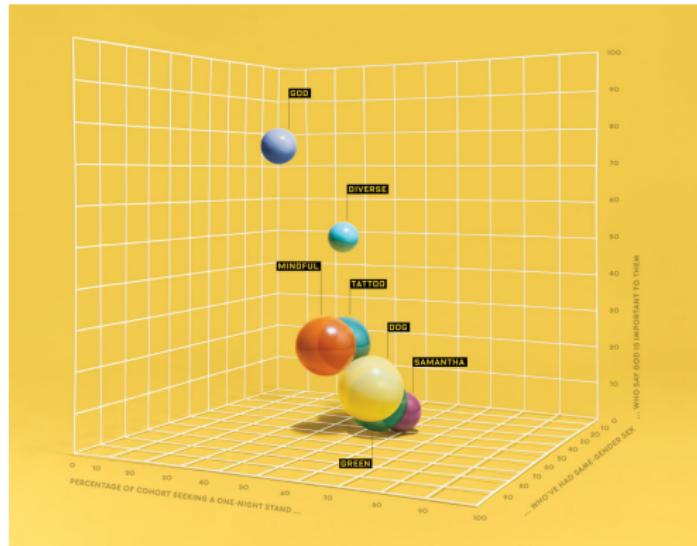
List of users, demographics, web session records, and content data



Use optimization to find optimal weights.

Winner has the best out-of-sample prediction!!

Hacking OkCupid



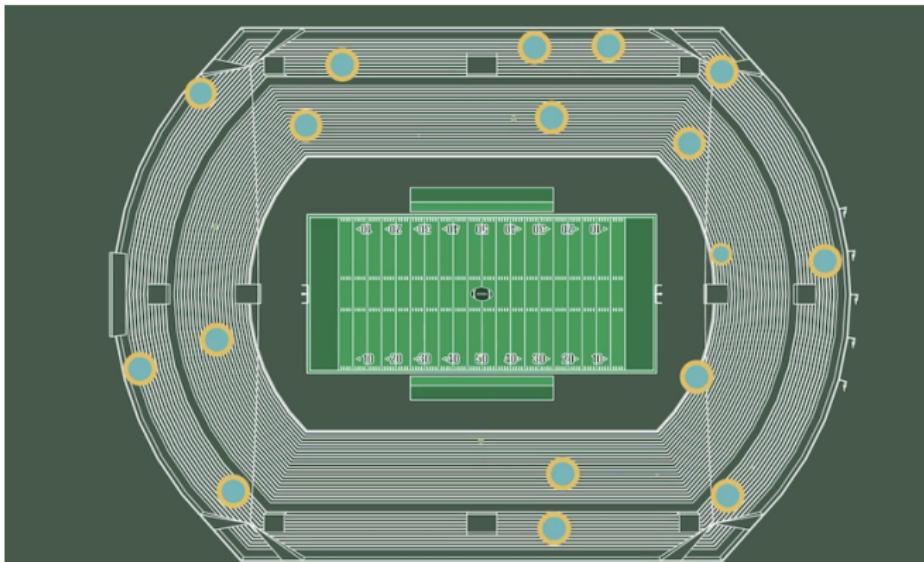
Sorted daters into seven clusters, like "Diverse" and "Mindful," each with distinct characteristics.

[Wired article](#)

[NOVA Video](#)

NFL Dynamic Pricing

Predict price demand for any given Lions game for any given seat in the stadium



<https://grahamschool.uchicago.edu/academic-programs/masters-degrees/analytics/nfl-capstone>

NFL Dynamic Pricing

We submitted our report on June 2016 suggesting that some areas of the stadium were priced efficiently and some were underpriced or overpriced.

On Feb 2017, Detroit Jock City wrote

"Detroit Lions tickets will cost a little more on average for 2017, but some areas of the stadium will decrease or hold steady."

Detroit Lions: Ticket Prices Get Modest Increase for 2017 Season



by Matt Snyder 2 years ago

[Follow @snyder_matthew](#)

TWEET

SHARE

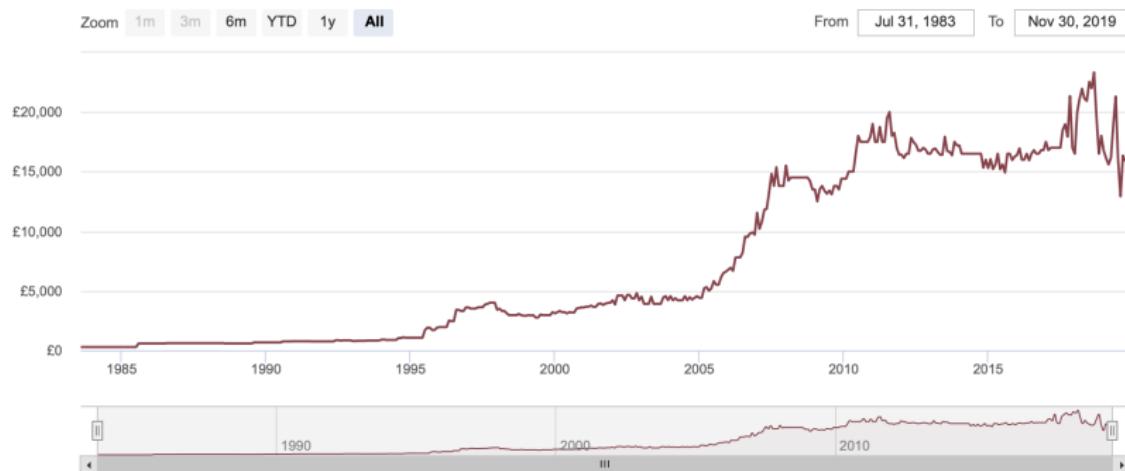


COMMENT

Detroit Lions tickets will cost a little more on average for 2017, but some areas of the stadium will decrease or hold steady.

<https://detroitjockcity.com/2017/02/10/detroit-lions-2017-ticket-prices/>

Wine: Latour 1982 Price History



wininvestment

Château Latour: grand vin

Bottle of Bordeaux wine sells for £135,000 at Christie's

⌚ 28 May 2011



Share

A single bottle of wine has sold for £135,000 in auction.

The six-litre bottle of 1961 Chateau Latour was sold in Hong Kong by London-based Christie's auction house.

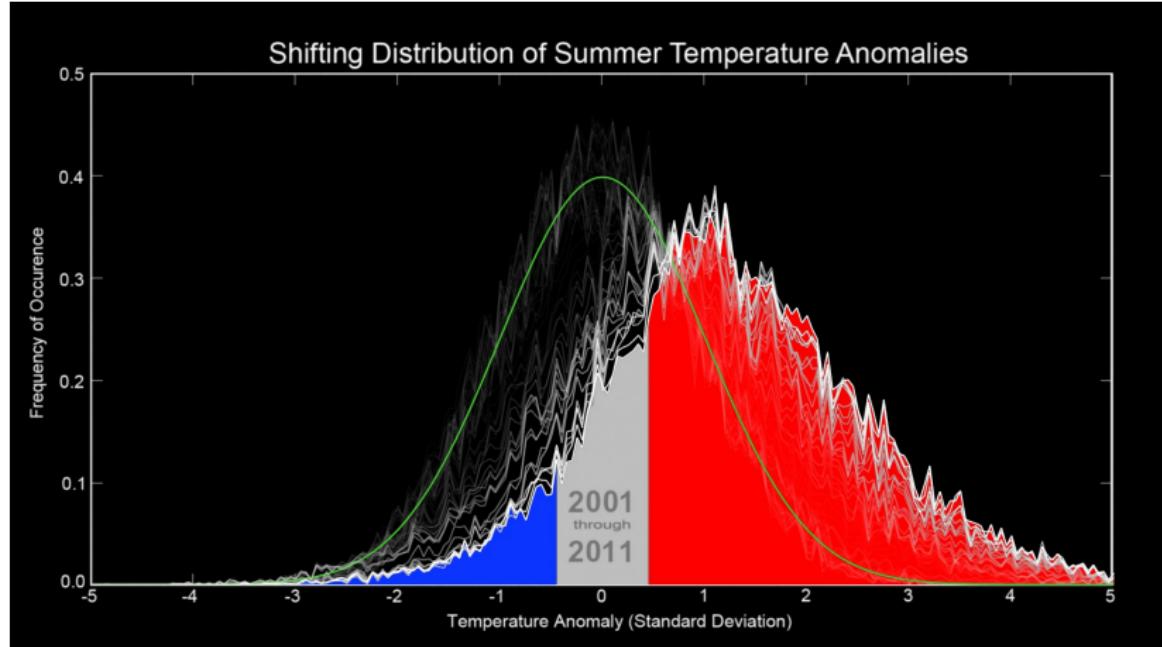
The sum was more than three times the expected price. Wine experts said the bottle was of "perfect provenance".

It would take someone earning the average UK wage more than five years to save up for the bottle. After tax, Prime Minister David Cameron could not afford it with his annual salary.



An expensive tipple - many houses cost less than the £135,000 bottle

Global Warming



Shifting Distribution of Northern Hemisphere Summer Temperature Anomalies,
1951-2011

NASA article with animation

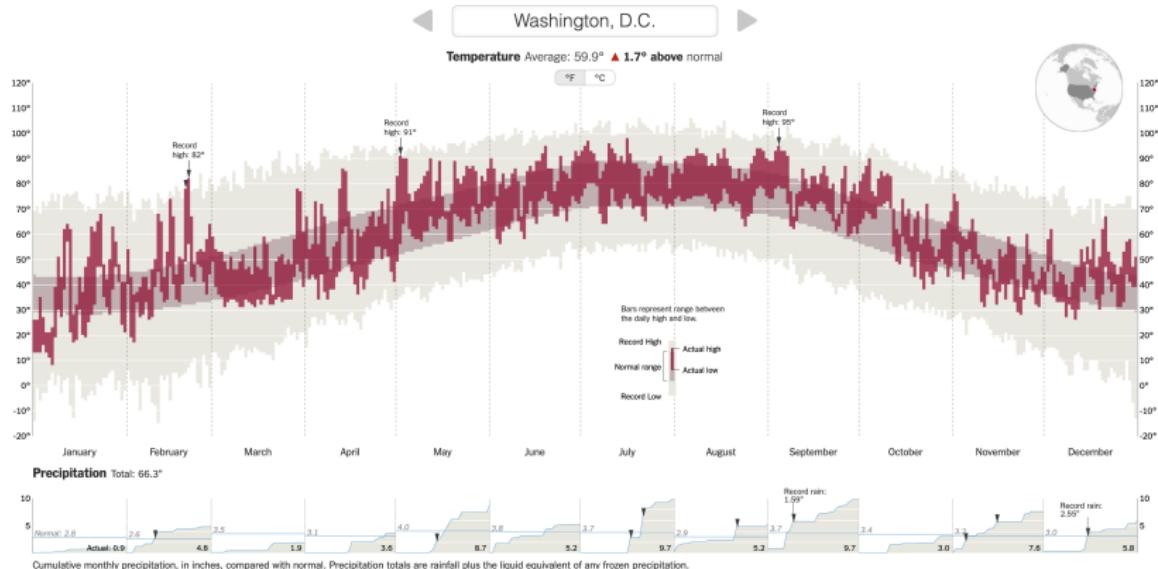
Climate statistics and public policy

Change in global mean temperature is not one of the most sensitive indicator

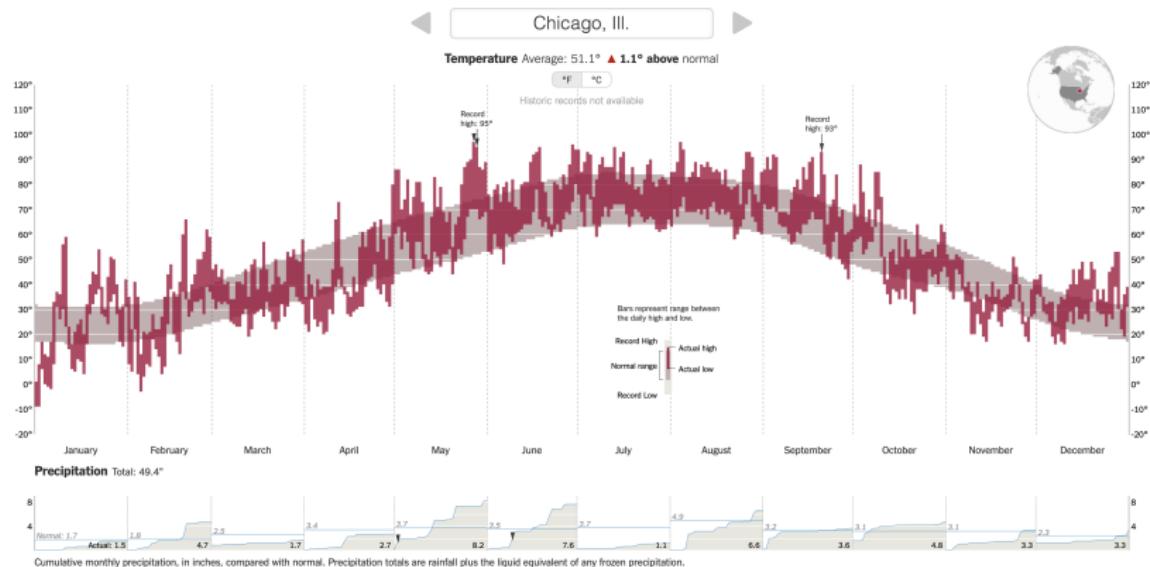
- ▶ Sea surface temperature and Land surface temperature
- ▶ Sea level rise (thermal expansion and ice melt): Greenland and West Antarctic are melting + glacial melt
- ▶ Ocean acidification: CO₂ gets absorbed by water, it produces carbolic acid
- ▶ Seasonal changes; winter - summer temperature has been decreasing since 1954. Shift changes (earlier seasons) lead to ecological effects
- ▶ Hurricanes: increase in maximum wind velocity = sea surface temperature + the difference between sea surface temperature and the average air temperature in the outflow of the hurricane

Guttorp paper

2018 was the fourth-warmest year on record.

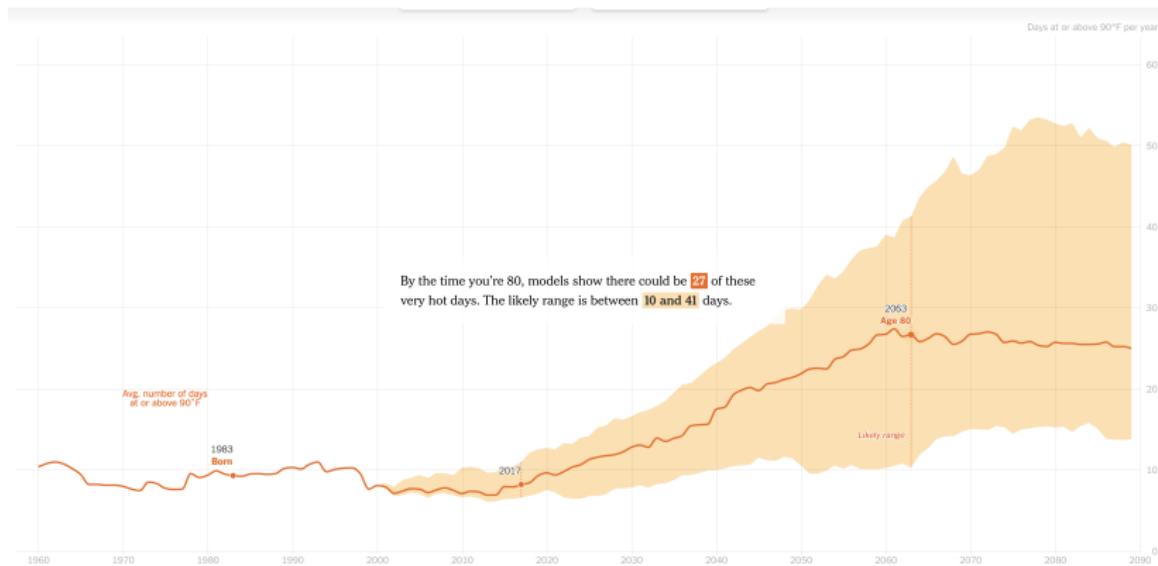


2018 was the fourth-warmest year on record.



NYT article

How Much Hotter Is Your Hometown Than When You Were Born?



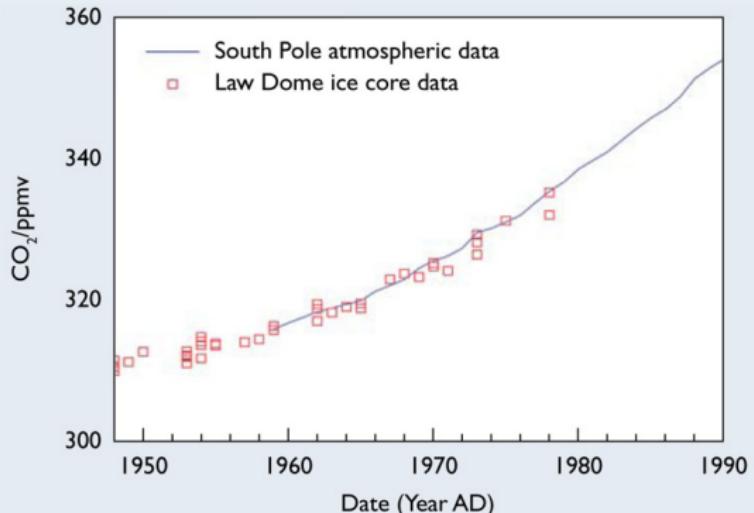
NYT article

Ice cores is an important source of data

Ice core. Cylinder of ice drilled out of an ice sheet or glacier. Most ice core records come from Antarctica and Greenland.

The oldest continuous ice core records to date extend 123,000 years in Greenland and 800,000 years in Antarctica.

Fig 1: Measurements of CO_2 from the Law Dome ice core⁽¹⁾ fall onto the line of annual average atmospheric measurements from South Pole⁽²⁾



Ice Core Datasets

Ice Core Basics

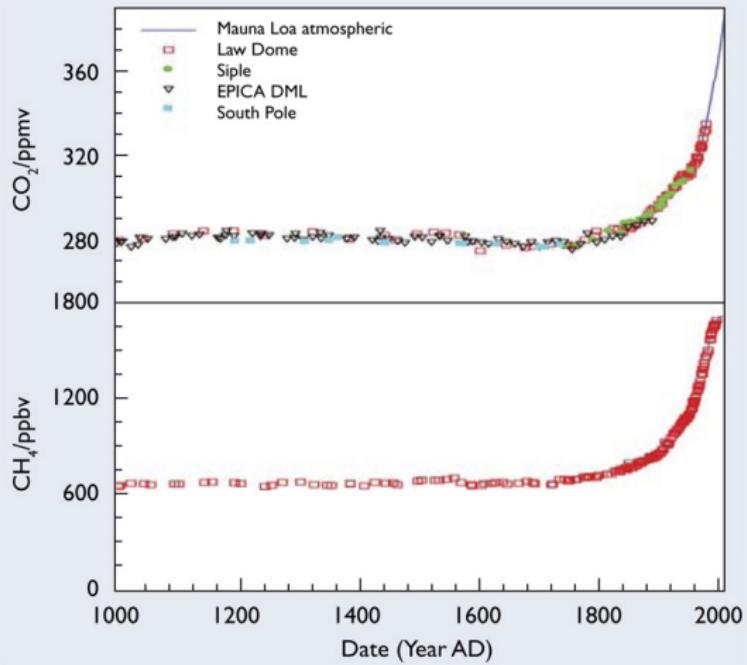
- ▶ has been around since the 1950s
- ▶ Mostly from Greenland and Antarctica
- ▶ bubbles in the ice core preserve actual samples of the world's ancient atmosphere

The World Data Center (WDC) for Paleoclimatology maintains archives of ice core data from polar and low-latitude mountain glaciers and ice caps throughout the world. Proxy climate indicators include oxygen isotopes, methane concentrations, dust content, as well as many other parameters.

<https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/ice-core>

CO₂ was stable over the last millennium

Fig. 2: CO₂ and CH₄ over the last 1,000 years⁽¹⁻⁴⁾



In the early 19th century CO₂ concentration started to rise, and its concentration is now nearly 40% higher than it was before the industrial revolution

Things we learned from ice core

Ice cores contain information about past temperature, and about many other aspects of the environment.

- ▶ Atmospheric carbon dioxide levels are now 40% higher than before the industrial revolution. This increase is due to fossil fuel usage and deforestation.
- ▶ The magnitude and rate of the recent increase are almost certainly unprecedented over the last 800,000 years.
- ▶ Methane also shows a huge and unprecedented increase in concentration over the last two centuries.

BAS article, [The Verge Article](#)

Gates thinks we can use more renewables and nuclear

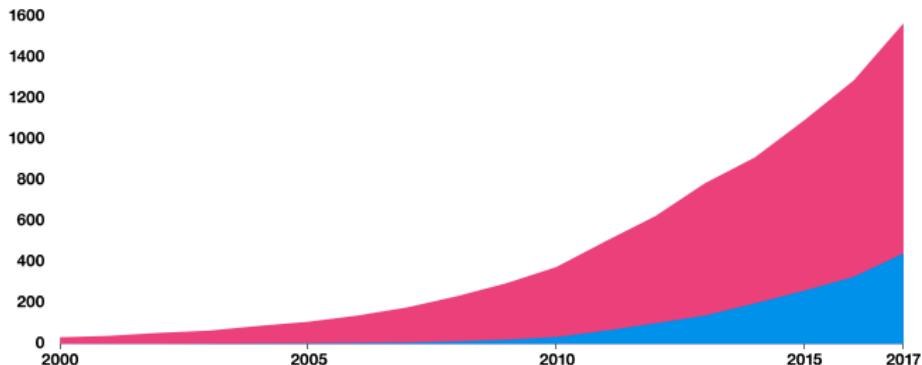


Rising renewables

Global wind and solar power generation is expanding as the world shifts from fossil fuels to carbon-free energy sources.

Wind Solar

Terawatt-hours per year



Source: BP Statistical Review of World Energy

Need better storage + generation (wind/sun) technology

Source: <https://www.gatesnotes.com/Energy/A-critical-step-to-reduce-climate-change>

Business Statistics: 41000

**Predictive Analytics
Logistic Regression**

Vadim Sokolov

The University of Chicago Booth School of Business

<http://vsokolov.org/courses/41000/>

Predictive Analytics

General Introduction

Predictive Analytics is the most widely used tool for high dimensional input-output analysis

$$Y = F(X) \text{ where } X = (X_1, \dots, X_p)$$

- ▶ Consumer Demand (Amazon, Airbnb, ...)
- ▶ Maps (Bing, Uber)
- ▶ Pricing
- ▶ Healthcare

The applications are endless

Logistic Regression: Classification : Y takes values 0 or 1

When the Y we are trying to predict is *categorical* (or *qualitative*) we say that we have a *classification* problem.

For a numeric (or *quantitative*) Y we predict its value

For a binary output we predict the probability its going to happen

$$p(Y = 1 | X = x)$$

where X is our usual list of predictors, X_1, \dots, X_p

Logistic Regression

Suppose that we have a binary response, Y taking the value 0 or 1

- ▶ Win or lose
- ▶ Sick or healthy
- ▶ Buy or not buy
- ▶ Pay or default

The goal is to predict the probability that Y equals 1

You can then do **classification** and categorize a new data-point

Example: Default Data

Here's a typical problem

Assessing credit risk and default data ...

- ▶ Y : whether or not a customer defaults on their credit card (No or Yes)
- ▶ X : The average balance that customer has remaining on their credit card after making their monthly payment.

... plus as many other features you think might predict Y ...

Logistic Regression

Y is an indicator: $Y = 0$ or 1 .

X is our usual set of predictors/covariates

We need to model the probability that $Y = 1$ as

$$p(Y = 1 | X_1, \dots, X_p) = f(\beta_1 X_1 + \dots + \beta_p X_p)$$

where f is increasing and $0 < f(X) < 1$. The **logit-transform** is given by

$$f(x) = e^x / (1 + e^x)$$

Logistic Regression

$$P = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)} = P(Y=1|X)$$

The logistic regression model is linear in **log-odds**

$$\log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

When x_i goes up by 1 unit log-odds go up by β_i

These models are easy to fit in R:

`glm(Y ~ X1 + X2, family = binomial)`

log-odds is
linear function
of X

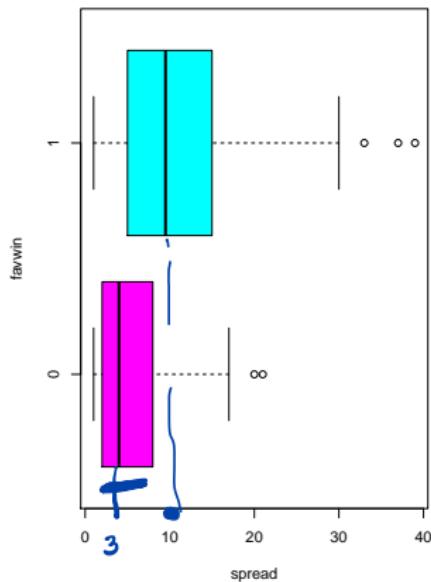
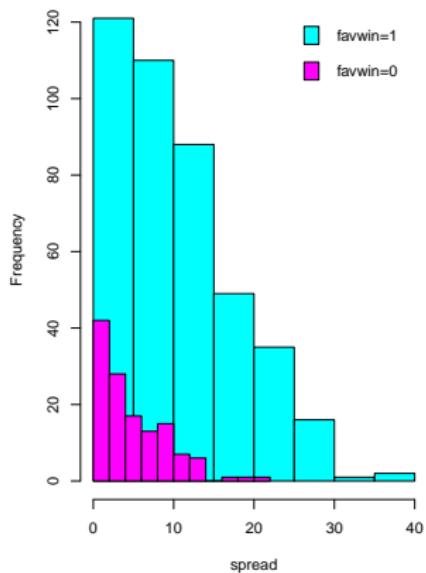
- “g” is for **generalized**; **binomial** indicates $Y = 0$ or 1

- “**glm**” has a bunch of other options.

if β_i go up by 1
log-odds go up by 1

Example: NBA point spread

Does the Vegas point spread predict whether the favorite wins or not?



Turquoise = Favorites does win, Purple = Favorite does not win

R: Logistic Regression

In R: the output gives us ...

```
nbareg = glm(favwin~spread-1, family=binomial)
```

```
summary(nbareg)
```

Call:

```
glm(formula = favwin ~ spread - 1, family = binomial)
```

Coefficients:

	Estimate	Std. Error	z value	P(> z)
spread	0.15600	0.01377	11.33	<2e-16 ***

```
# prediction
```

```
newweek=c(8,4)
```

do not use β_0

y is 0 or 1

The β measures how our log-odds change! $\beta = 0.156$

NBA Point Spread Prediction

“Plug-in” the values for the new game into our logistic regression

$$P(\text{favwin} | \text{spread}) = \frac{e^{\beta x}}{1 + e^{\beta x}}$$

Check that when $\beta = 0$ we have $p = \frac{1}{2}$. $\frac{1}{1+1} = \frac{1}{2}$

- Given our new values spread= 8 or spread= 4,

The win probabilities are 77% and 65%, respectively. Clearly, the bigger spread means a higher chance of winning.

$$\frac{e^{0.15 \cdot 8}}{1 + e^{0.15 \cdot 8}} = 0.77$$

Credit Card Default

10,000 observations

```
> head(Default)
```

	default	student	balance	income
1	No	No	729.5265	44361.625
2	No	Yes	817.1804	12106.135
3	No	No	1073.5492	31767.139
4	No	No	529.2506	35704.494
5	No	No	785.6559	38463.496
6	No	Yes	919.5885	7491.559

Let's build a logistic regression model

Call:

```
glm(formula = default ~ balance, family = binomial, data = Default)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.065e+01	3.612e-01	-29.49	<2e-16 ***	
balance	5.499e-03	2.204e-04	24.95	<2e-16 ***	✓

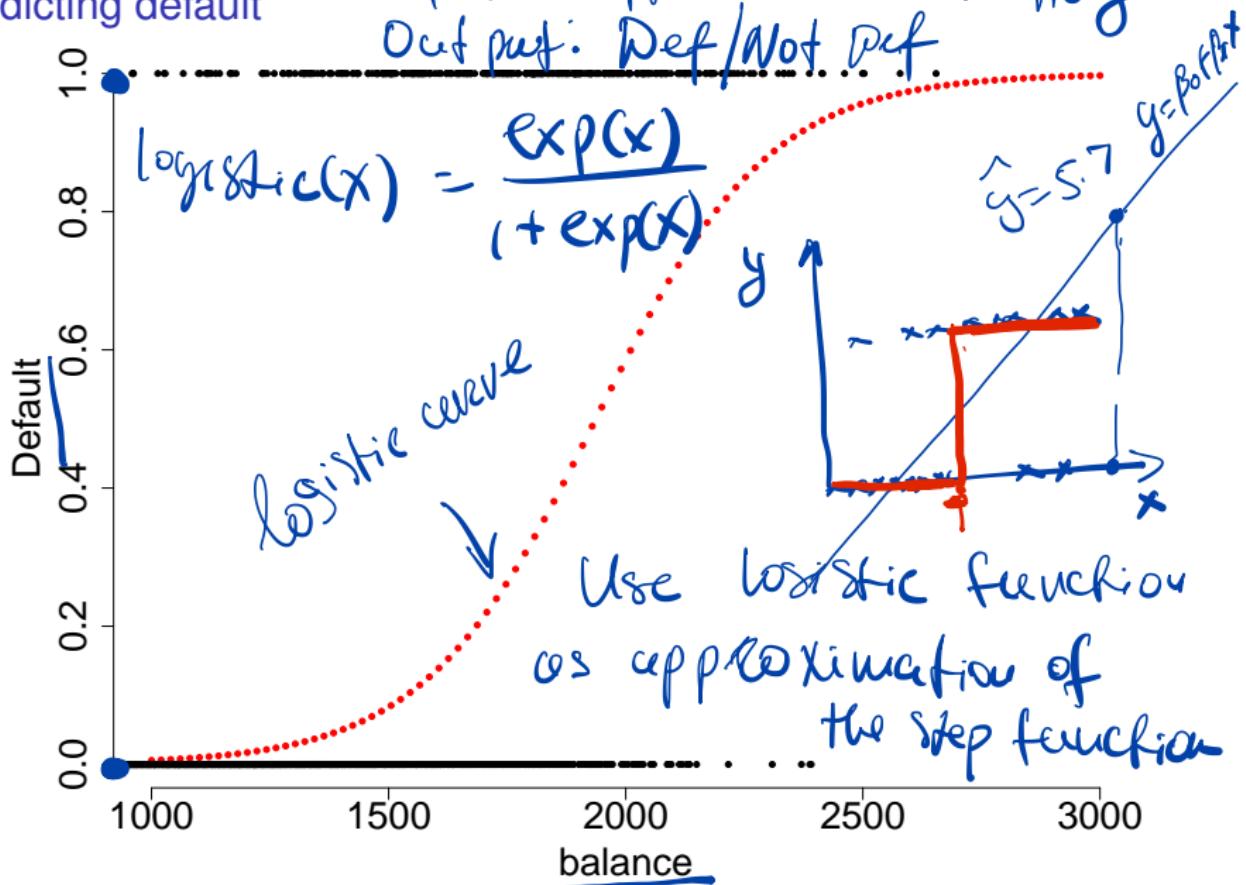
Predicting default

by def $\beta_0 + \beta_1 X$

```
> predict.glm(glm.fit,newdata = list(balance=1000))  
1  
-5.152414  
  
> -1.065e+01 + 5.499e-03*1000  
  
[1] -5.151  
  
> predict.glm(glm.fit,newdata = list(balance=1000), type="response")  
1  
0.005752145 =  $P(y=1 | bal=1000)$   
  
> exp(-1.065e+01 + 5.499e-03*1000)/(1+exp(-1.065e+01 + 5.499e-03*1000))  
  
[1] 0.005760236
```

Predicting default

Inputs: Appl. + Credit history
Out put: Def/Not Def.



$X \rightarrow y$ & y is 0 or 1 Logistic
Regression

1) $\underline{\beta_0} + \underline{\beta_r X} = \mu$; μ takes any value
- ∞ to $+\infty$

2) $P = \frac{e^{\mu}}{1 + e^{\mu}} = \frac{e^{\beta_0 + \beta_r X}}{1 + e^{\beta_0 + \beta_r X}}$

P takes any value between 0 & 1

3) $P = P(y=1|x)$; Decide on threshold
if $P \leq \alpha \Rightarrow \hat{y} = 0$
if $P > \alpha \Rightarrow \hat{y} = 1$

Evaluate the model

$$\underline{\alpha = 0.2}$$

$$\underline{\text{Accuracy} = 0.96} =$$

α - threshold

$P < \alpha \Rightarrow \hat{y} = 0$

$P > \alpha \Rightarrow \hat{y} = 1$

of correct predictions

$\alpha = 0$ A lot of FP
 $\alpha = 1$ A lot of FN

n

	Predicted: YES	Predicted: NO
Actual: YES	TPR = 0.6	FNR = 0.4
Actual: NO	FPR = 0.03	TNR = 0.97

more costly

1

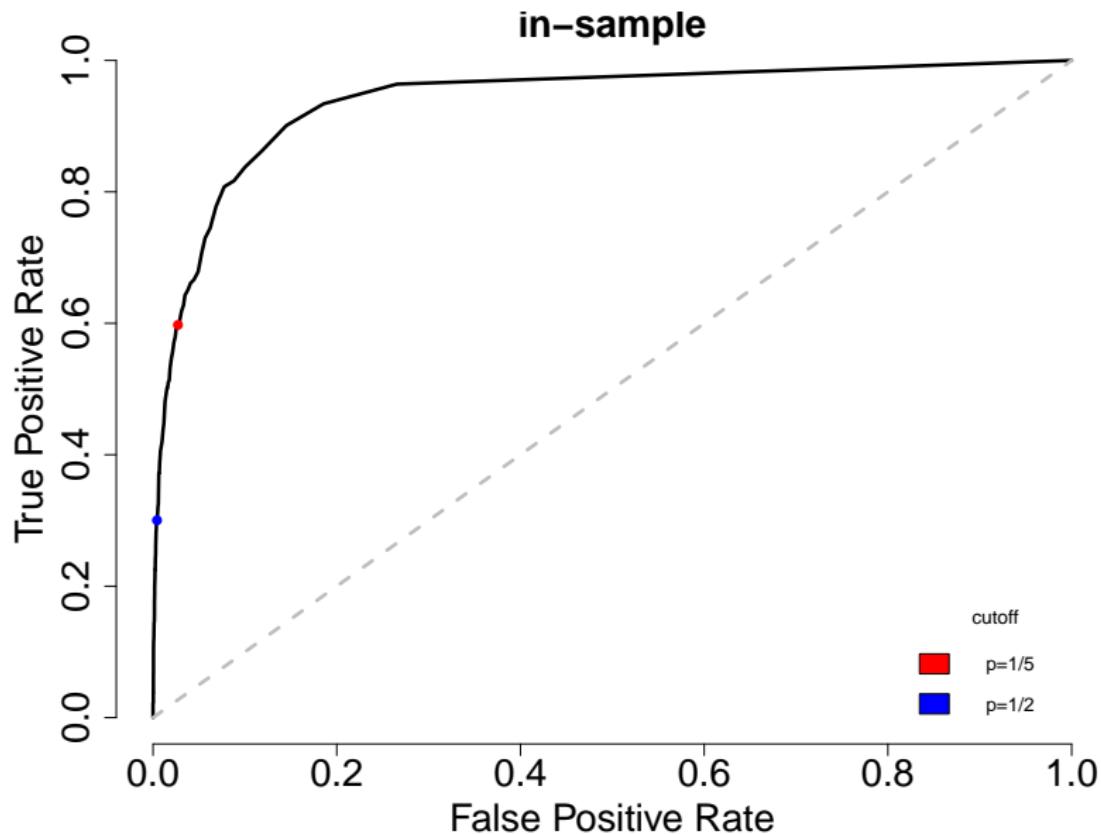
I used $\alpha = 0.2$ as a cut-off. What if I use smaller or larger p , e.g. $p = 0$?

True positive Rate =

of correct Default predictions

of obs. where $y = \text{Default}$

ROC Curve Shows what happens for different cut-off values



Look at other predictors

Call:

```
glm(formula = default ~ balance + income + student, family = binomial,  
    data = Default)
```

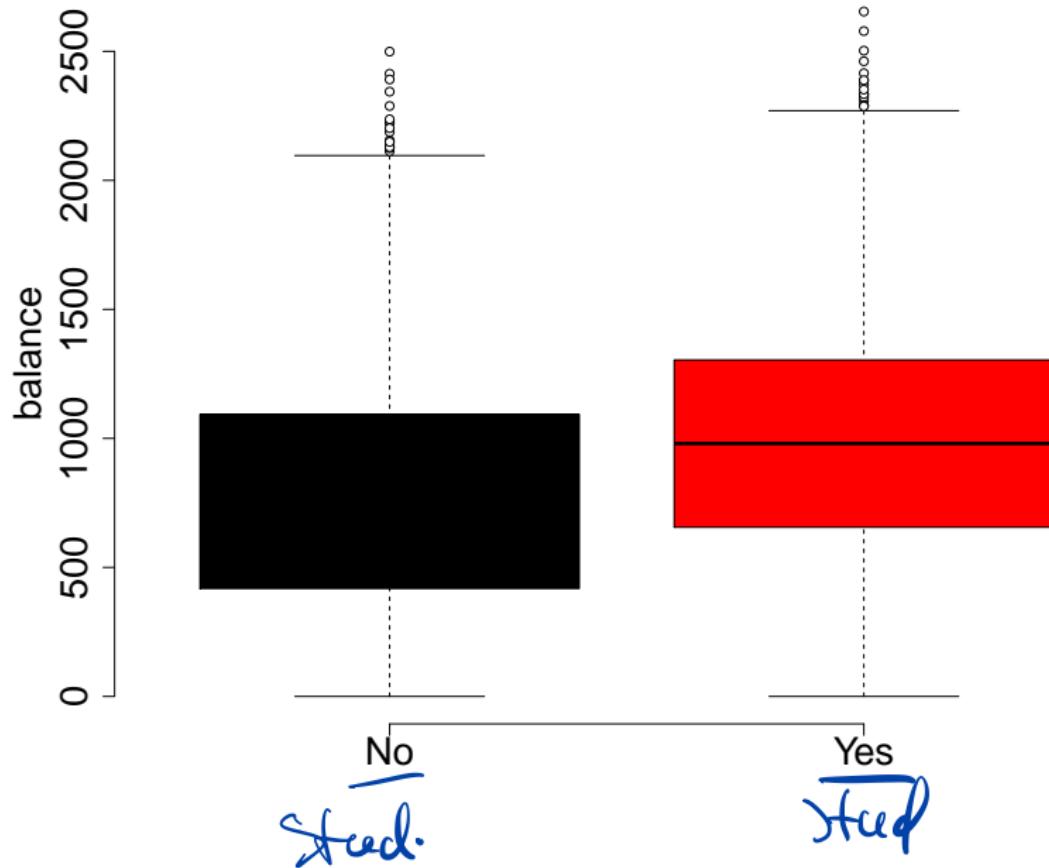


Coefficients:

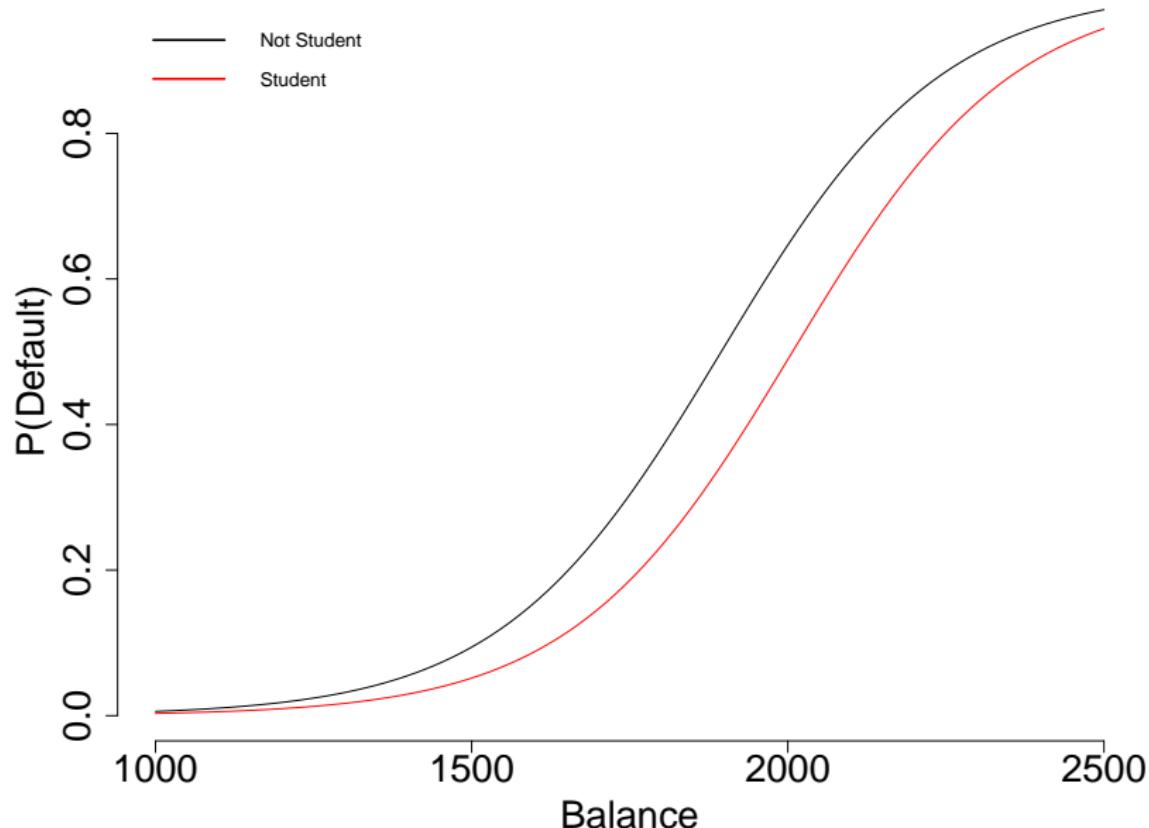
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.087e+01	4.923e-01	-22.080	< 2e-16 ***	
balance	5.737e-03	2.319e-04	24.738	< 2e-16 ***	✓
income	3.033e-06	8.205e-06	0.370	0.71152	
studentYes	-6.468e-01	2.363e-01	-2.738	0.00619 **	✓

Student is significant!?

Student vs Balance



Let's adjust for balance



Data Science in Tennis

Break point is when

game → set → match
6 3

You win game

Opponent is serving.

Data science plays a major role in tennis

- ▶ IBM (major sponsor of grand slams) has developed an AI toolbox
- ▶ We will analyze the Tennis Major Tournament Match Statistics Data Set
- ▶ Each row is a game from four major Tennis tournaments in 2013 (Australia Open, French Open, US Open, and Wimbledon). Let's load the data and familiarize ourselves with it

How important are the breakpoints in tennis?

$y = \text{Win/Loss by Player}$

1

```
d = read.csv("~/book/bookmd/data/tennis.csv")
dim(d)
## [1] 943 44
str(d[,1:5])
```

'data.frame': 943 obs. of 5 variables:
\$ Player1: chr "Lukas Lacko" "Leonardo Mayer" "Marcos Baghdatis" "Dmitry Tu"..
\$ Player2: chr "Novak Djokovic" "Albert Montanes" "Denis Istomin" "Michael "...
\$ Round : int 1 1 1 1 1 1 1 1 1 1 ...
\$ Result : int 0 1 0 1 0 0 0 1 0 1 ...
\$ FNL1 : int 0 3 0 3 1 1 2 2 0 3 ...

We have data for 943 matches and for each match we have 44 columns, including names of the players, their gender, surface type and match statistics

Peak at the data

Let's look at the few columns of the randomly selected five rows of the data

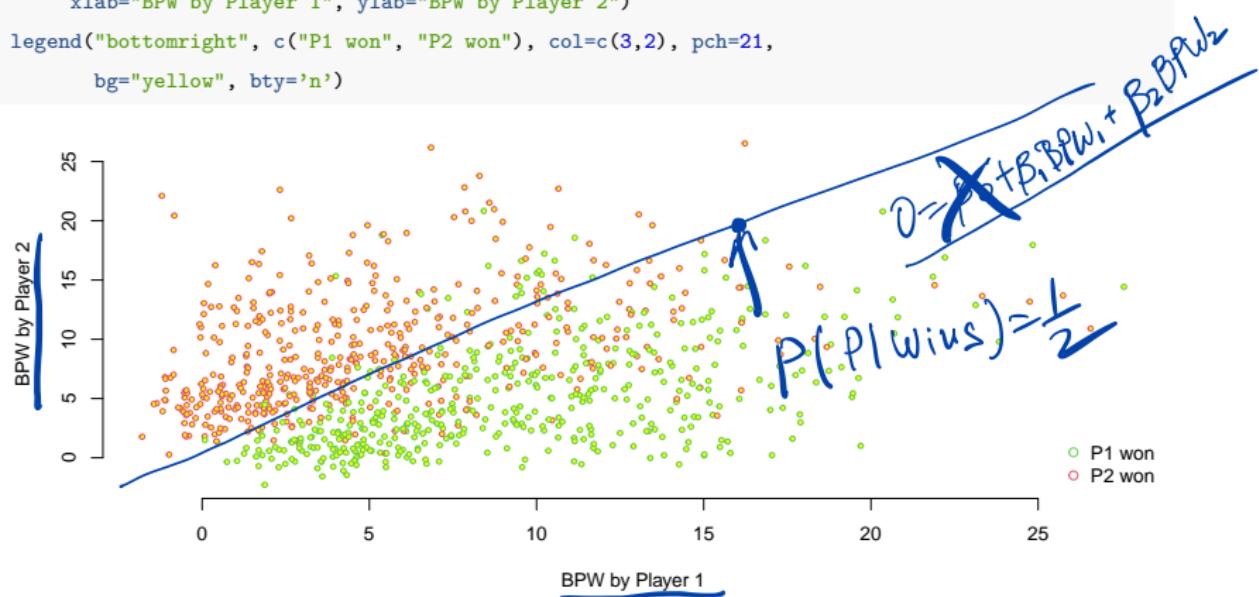
```
d[sample(1:943, size = 5), c("Player1", "Player2", "Round", "Result",
                           "gender", "surf")]
```

```
##          Player1           Player2 Round Result gender surf
## 554    Jurgen Zopp Marcel Granollers     1      0     M Hard
## 112   Fabio Fognini Novak Djokovic      4      0     M Hard
## 39    Thomaz Bellucci Julian Reister     1      1     M Hard
## 669       A Cornet A Tomljanovic      2      1     W Hard
## 744 D.Istomin        A.Seppi         1      0     M Grass
```

Number of break points won by each player

We will plot BPW (break points won) by each player on the scatter plot and will colorize each dot according to the outcome

```
n = dim(d)[1]  
plot(d$BPW.1+rnorm(n), d$BPW.2+rnorm(n), pch=21, bty="n",  
     col=d$Result+2, cex=0.6, bg="yellow", lwd=0.8,  
     xlab="BPW by Player 1", ylab="BPW by Player 2")  
legend("bottomright", c("P1 won", "P2 won"), col=c(3,2), pch=21,  
      bg="yellow", bty='n')
```



There is clearly a pattern! Let's quantify it using logistic regression.

Logistic regression

```
which(is.na(d$BPW.1)) # there is one row with NA value for the BPW.1 value and we remove it

## [1] 171

d = d[-171,]; n = dim(d)[1]
m = glm(Result ~ BPW.1 + BPW.2-1, data=d, family = "binomial")
summary(m)

##
## Call:
## glm(formula = Result ~ BPW.1 + BPW.2 - 1, family = "binomial",
##      data = d)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.425  -0.668  -0.055   0.636   3.085
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## BPW.1      0.4019    0.0264   15.2   <2e-16 ***
## BPW.2     -0.4183    0.0277  -15.1   <2e-16 ***
##
```

How well our model captures the pattern?

R output does not tell us how accurate our model is but we can quickly check it by using the `table` function. We will use 0.5 as a threshold for our classification.

```
table(d$Result, as.integer(m$fitted.values>0.5))
```

```
##      T1      0      1  
## 0  416  61 FN  
## 1    65 400 -TP  
##          FP
```

Confusion matrix (2x2)

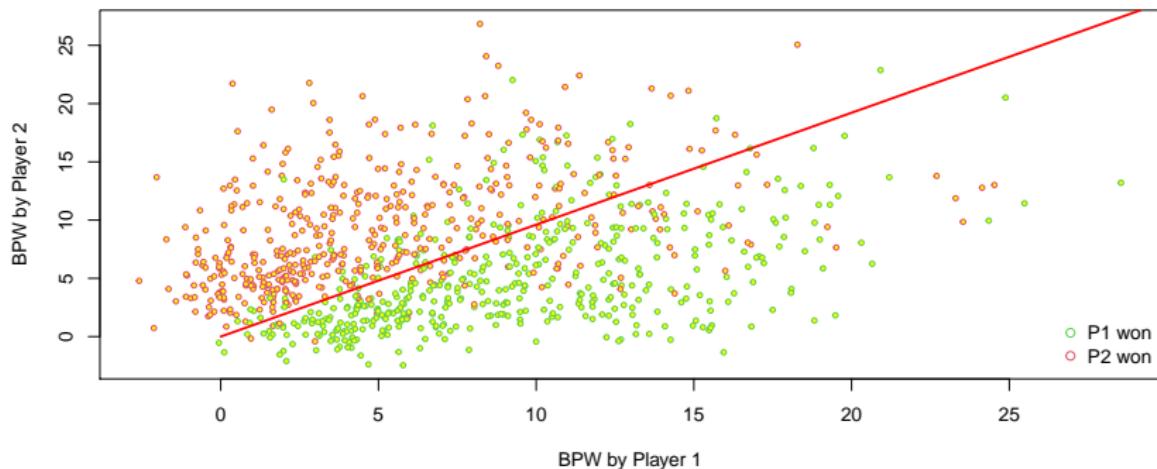
Thus, our model got $(416+400)/942 = 87\%$ of the predictions correctly!

GLM Line

Let's see the line found by the `glm` function

```
plot(d$BPW.1+rnorm(n),d$BPW.2+rnorm(n), pch=21, col=d$Result+2, cex=0.6,
      bg="yellow", lwd=0.8,xlab="BPW by Player 1", ylab="BPW by Player 2")
legend("bottomright", c("P1 won", "P2 won"), col=c(3,2), pch=21,
      bg="yellow", bty='n')

x = seq(0,30,length.out = 200)
y = -m$coefficients[1]*x/m$coefficients[2]
lines(x,y, lwd=2, col="red")
```



What did we find?

- ▶ Effect of a break point on the game outcome is significant
- ▶ It is symmetric, Dah! Effect of loosing break point is the same as the effect of winning one
- ▶ The chances of winning when P1 wins three more break points compared to the opponent:

```
predict.glm(m,newdata = data.frame(BPW.1 = c(0), BPW.2 = c(0)), type="response")
```

```
## 1
```

```
## 0.5
```

```
predict.glm(m,newdata = data.frame(BPW.1 = c(3), BPW.2 = c(0)), type="response")
```

```
## 1
```

```
## 0.77
```

Chances go up by 27%.

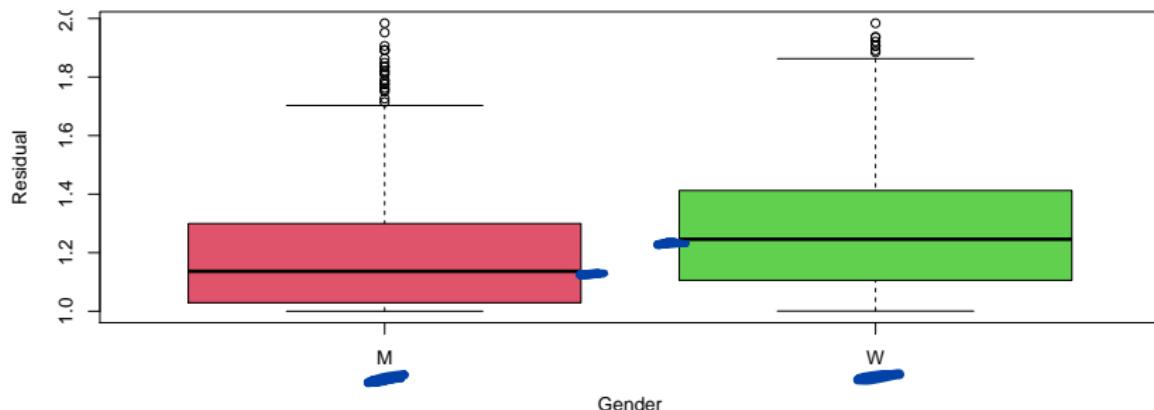
3 extra Break points
lead to 27% increase
in prob. of win.

Are women's matches less predictable?

guyz games

We can test this statement by looking at the residuals. The larger the residual the less predictable the game.

```
d$res = abs(m$residuals)
outlind = which(d$res<2)
boxplot(d$res[outlind] ~ d$gender[outlind], col=c(2,3), xlab="Gender",
        bty='n', ylab="Residual")
```



Looks like the crowd wisdom that Women's matches are less predictable is correct.

LinkedIn Study: How to Become an Executive

$$y = \text{CEO}/\text{not CEO}$$

Analyze the career paths of about 459,000 LinkedIn members who worked at a **Top 10 consultancy** between 1990 and 2010 and became a VP, CXO, or partner at a company with at least 200 employees.

About 64,000 members reached this milestone. $\hat{p} = 0.1394$.

- ▶ Look at their profiles – educational background, gender, work experience, and career transitions.
- ▶ Build a model to predict the probability of becoming an executive.

Conditional on making it into the database

Logistic Regression

Logistic regression with **8 key features** (a.k.a. covariates):

$$\log \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_8 X_8$$

- ▶ p : Probability of “success” – reach VP/CXO/Partner at a company with at least 200 employees.
- ▶ $X_i (i = 1, 2, \dots, 8)$: Features to predict the “success” probability.

Features

Location Features: X_1 Metro region: whether a member has worked in one of the top 10 largest cities in the U.S. or globally.

Personal Features: X_2 Gender: Inferred from member names: 'male', or 'female'.

Education Features: X_3 Graduate education type: whether a member has an MBA from a top U.S. program / a non-top program / a top non-U.S. program / another advanced degree.

X_4 Undergraduate education type: whether a member has attended a school from the U.S. News national university rankings / a top 10 liberal arts college / a top 10 non-U.S. school.

Features

Work Experience:

X_5 Company count: # different companies in which a member has worked.

X_6 Function count: # different job functions in which a member has worked.

X_7 Industry sector count: # different industries in which a member has worked.

X_8 Years of experience: # years of work experience, including years in consulting,
for a member.

$\hat{\beta}$'s of Features¹

1. Location: Metro region: 0.28
2. Personal: Gender(Male): 0.31
3. Education: Graduate education type: 1.16,
Undergraduate education type: 0.22
4. Work Experience: Company count: 0.14,
Function count: 0.26,
Industry sector count: -0.22,
Years of experience: 0.09

MBA from 10 top programs

Main Findings

1. Working across job functions, like marketing or finance, is good. Each additional job function provides a boost that, on average, is equal to three years of work experience. Switching industries has a slight negative impact.
Learning curve? Lost network?
2. MBAs are worth the investment. But pedigree matters.
3. Location matters. NYC helps.

Top five program equivalent to 13 years of work experience!!!

Examples

Person A (p=6%): Male in Tulsa, Oklahoma, Undergraduate degree, 1 job

function for 3 companies in 3 industries, 15-year experience.

Person B (p=15%): Male in London, Undergraduate degree from top international

school, Non-MBA Master, 2 different job functions for 2 companies in 2 industries,

15-year experience.

Person C (p=63%): Female in New York City, Top undergraduate program, Top

MBA program, 4 different job functions for 4 companies in 1 industry, 15-year

experience.

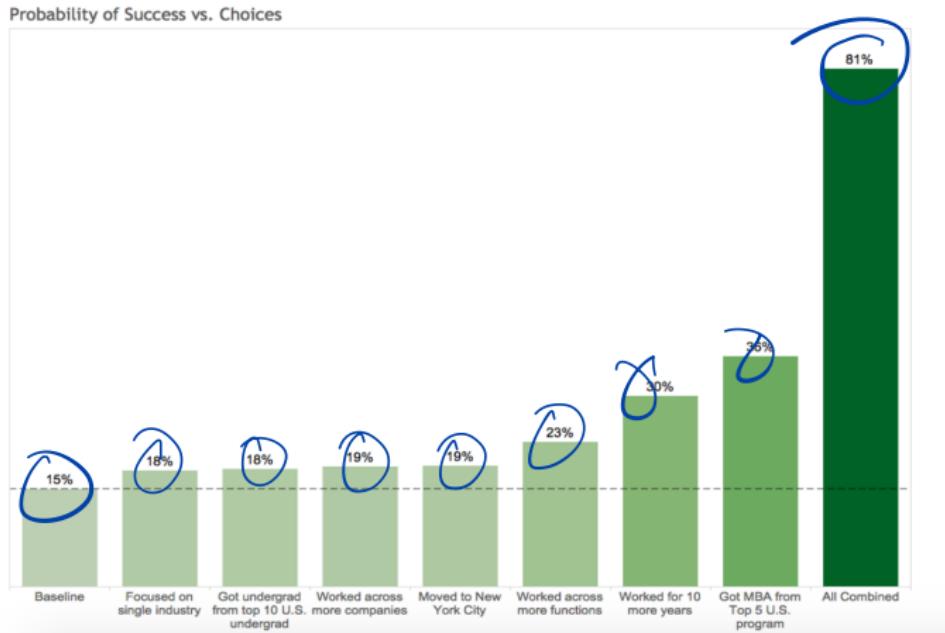
Let's re-design Person B!!

Person B (p=15%): Male in London, Undergraduate degree from top international school, Non-MBA Master, 2 different job functions for 2 companies in 2 industries, 15-year experience.

1. Work in one industry rather than two. Increase 3%
2. Undergrad from top 10 US program rather than top international school. 3%
3. Worked for 4 companies rather than 2. Another 4%
4. Move from London to NYC. 4%
5. Four job functions rather than two. 8%. A 1.5X effect.
6. Worked for 10 more years. 15%. A 2X effect.

NYT article

Choices and Impact (Person B)



Summary

- ▶ Multiple Regression (Newfood study, Golf Analysis)
- ▶ Interactions (how advertisement change price elasticity?)
- ▶ Predictive analytics cases(Target, Walmart, Airbnb, Stitch Fix)
- ▶ Logistic regression (NBA predictions, Horse predictions, LinkedIn)