Class 6a: Predicting Response with Logistic Regression

Professor Blake McShane

MKTG 482: Customer Analytics
Kellogg School of Management

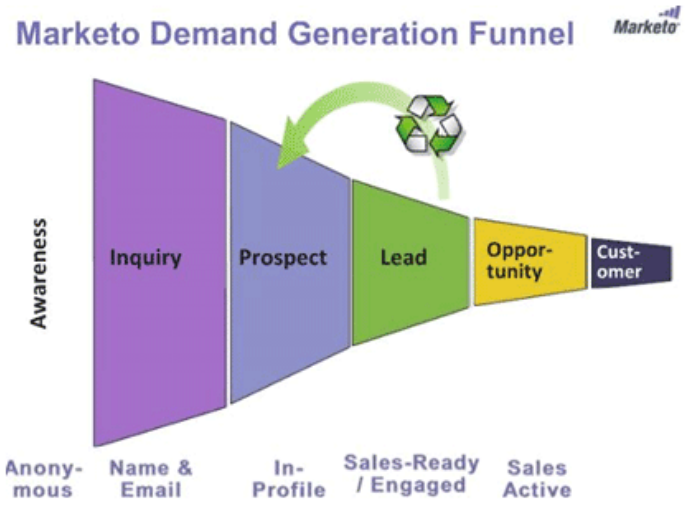
RFM has some major shortcomings

RFM DISADVANTAGES

- Not very 'sophisticated,' i.e. built as rule of thumb
- Does not "scale" well to include other variables
- Predicts on the basis of a **membership in a specific RFM cell**
- Does not predict individual response probabilities based on **individual customer statistics**

Need a more flexible, powerful model to predict response / purchase probability

Lead management is an important part of sales



Many firms monitor the activities of potential customers to determine whether they might be good leads

EXAMPLE: SMARTSTORAGE

- A top object-based storage provider (huge capacity, speed less important)
- Software runs data centers for many large cloud storage providers (e.g. major photo site)
- Have limited number of potential clients
- Identified 14,000 people involved in potentially buying their product
- Can identify most potential decision makers on their site without requiring them to log in

Smartstorage keeps track of behaviors and demographics of potential decision makers

BEHAVIORS

Website

- Visited webpage/blog
- Viewed introductory content
- Viewed mid-stage content
- Viewed late-stage content
- Visited pricing page
- Visited career page
- Watched demos

Other

- Visited at trade-show
- Contacted company
- Provided e-mail
- ...

DEMOGRAPHICS

- High-relevance employer
- High-relevance job title
- Relevant past experience
- Small potential client
- Large potential client
- ...

Many companies use some form of lead scoring to determine "Qualified Leads"

BEHAVIORS

Website

- Visited webpage/blog (+1)
- Viewed introductory content (+5)
- Viewed mid-stage content (+8)
- Viewed late-stage content (+10)
- Visited pricing page (+10)
- Visited career page (-10)
- Watched demos (+5)

Other

- Visited at trade-show (+5)
- Contacted company (+15)
- Provided e-mail (+10)
- ...

DEMOGRAPHICS

- High-relevance employer (+20)
- High-relevance job title (+10)
- Relevant past experience (+6)
- Large potential client (+10)
- ...

Is this predictive analytics?

Which leads should go to sales?

			Behavior Score			
			50•	24-50	0-25	0
			1	2	3	4
Demographic Score	50+	A				
	24-50	B				
	0-25	C				
	0	D				

Source: Marketo: The definitive guide to lead scoring

Predictive Analytics:

Using data **that you have** to predict data that **you don't have**, using statistical or machine learning approaches.

Using "statistical of machine learning approaches" refers to using data to uncover *how* behavior and demographics drive lead quality

How would we implement predictive analytics for lead scoring at Smartstorage? (simplified example)

BEHAVIORS

Website

- Visited webpage/blog
- Viewed introductory content
- Viewed mid-stage content
- Viewed late-stage content
- Visited pricing page
- Visited career page
- Watched demos

Other

- Visited at trade-show
- Contacted company
- Provided e-mail
- ...

DEMOGRAPHICS

- High-relevance employer
- High-relevance job title
- Relevant past experience
- Small potential client
- Large potential client
- ...

Missing?

We use information on 180 leads, including whether they converted to a sale within 150 days of first ID

id	webpageviews	viewedpricing	highrelevancefirm	sale
639	15	0	1	0
272	35	0	1	1
491	7	0	1	0
226	18	1	1	0
7195	13	0	0	0
9080	23	0	0	0
548	14	0	1	0
9605	36	0	0	0
5352	28	0	0	0
4343	35	0	0	0
14971	3	0	0	0
11298	34	0	0	1
317	10	0	1	0
...	...	...	...	...

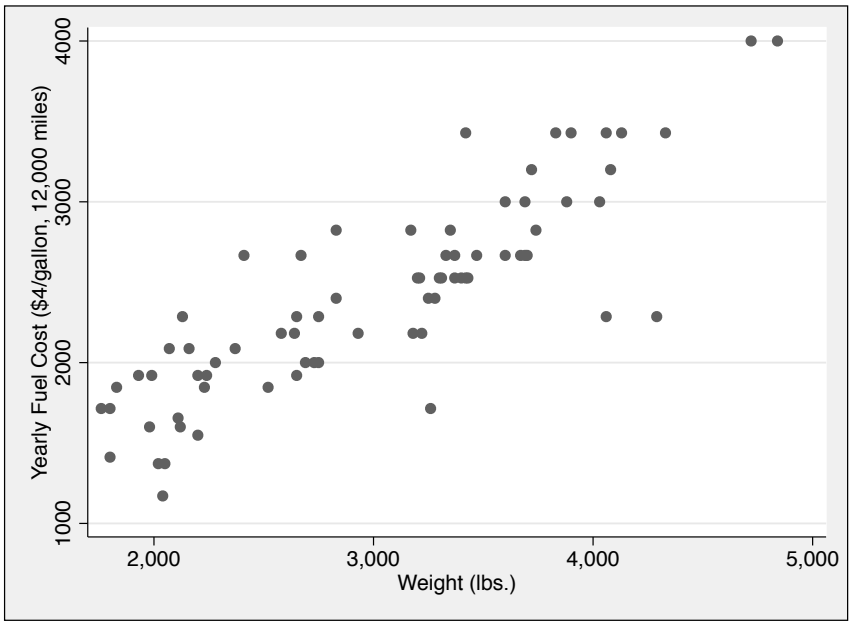
How would we determine the relationship between demographics and behaviors and sales success?

id	webpageviews	viewedpricing	highrelevancefirm	sale
639	15	0	1	0
272	35	0	1	1
491	7	0	1	0
226	18	1	1	0
7195	13	0	0	0
9080	23	0	0	0
548	14	0	1	0
9605	36	0	0	0
5352	28	0	0	0
4343	35	0	0	0
14971	3	0	0	0
11298	34	0	0	1
317	10	0	1	0
...	...	...	...	...

id	webpageviews	viewedpricing	highrelevancefirm	sale
7236	27	0	0	
687	25	1	1	
453	16	0	0	
563	6	0	1	

How do data scientists create a prediction other than RFM?

EXAMPLE: YEARLY FUEL COST AND VEHICLE WEIGHT

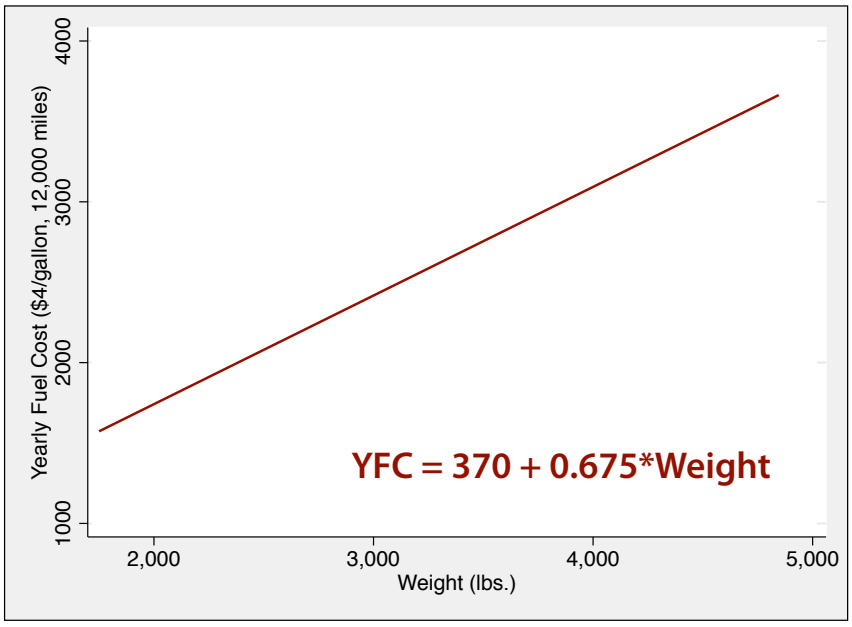


Inputs to calculate yearly fuel cost (YFC):

- MGP of vehicle
- 12,000 miles per year
- $ 4 / gallon

How do data scientists create a prediction using a regression?

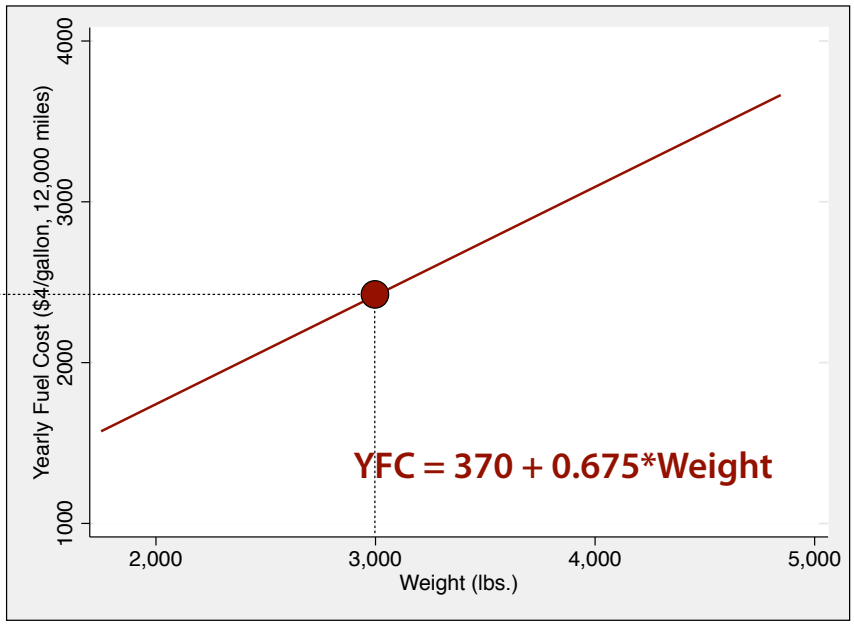
EXAMPLE: YEARLY FUEL COST AND VEHICLE WEIGHT



Weight	Prediction
3000	?
4500	?
2356	?

How do data scientists create a prediction using a regression?

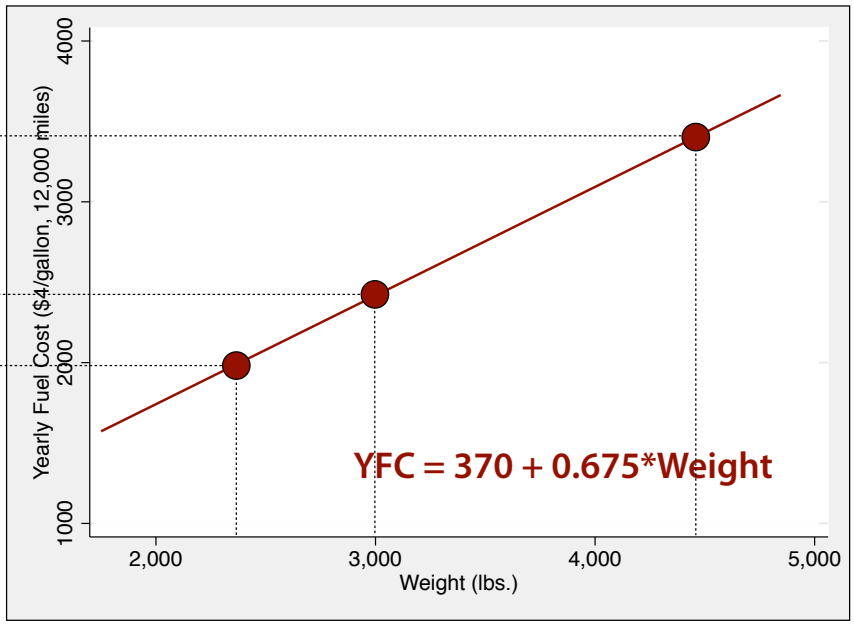
EXAMPLE: YEARLY FUEL COST AND VEHICLE WEIGHT



Weight	Prediction
3000	$2,395.00
4500	?
2356	?

How do data scientists create a prediction using a regression?

EXAMPLE: YEARLY FUEL COST AND VEHICLE WEIGHT



Weight	Prediction
3000	$2,395.00
4500	$3,407.50
2356	$1,960.30

700 lb weight reduction:
0.675*700 = $472.5

Can we use this approach for predicting qualified leads?

PREDICTION APPROACHES

Yearly fuel costs and weight

- After running regression we found that this formula describes the data

YFC = 370 + 0.675*Weight

- Can now predict YFC for any weight

Sales and lead characteristics

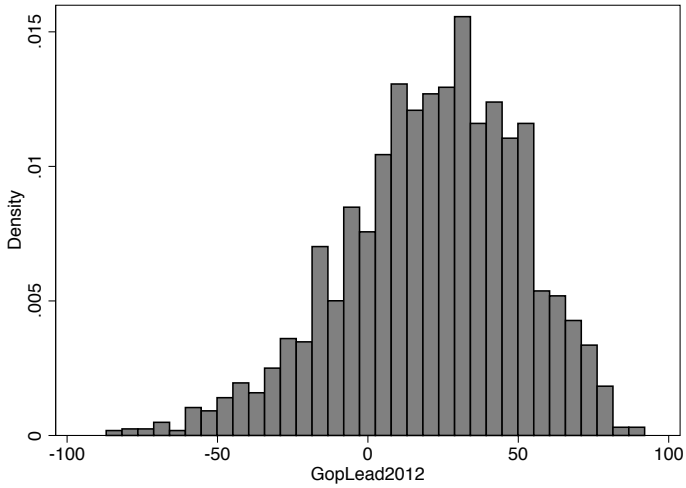
Sale = A + B * (# webpages/blogs visited) + C * ...

- What kind of variable is "sale"
- How do we interpret "predicted sale"?

Let's consider 2016 presidential election

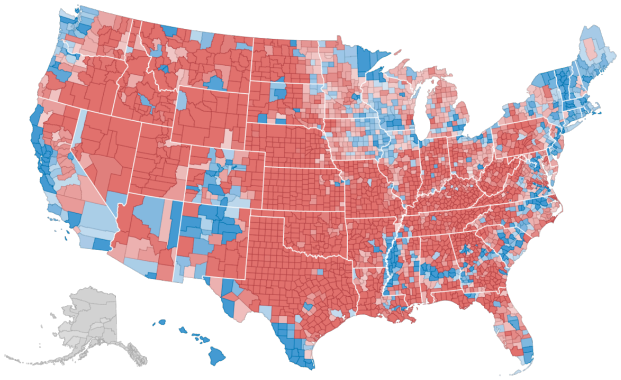
PREDICTING THE PRESIDENTIAL VOTE FOR EACH COUNTY

- **Dependent** variable: **trump** -- "*wins*" (1) or "*loses*" (0) **by county** (3112)
- **Predictor** variable: GOP – DEM 2012 % vote difference ("GopLead2012")

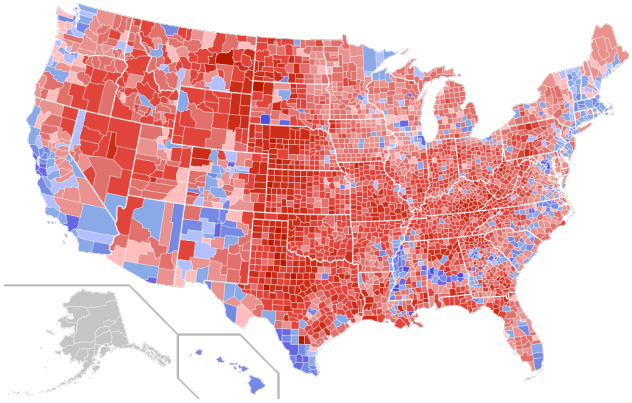


- 0 tie
- – 50 = 50 percentage point lead for **DEM**
- 50 = 50 percentage point lead for **GOP**

2012

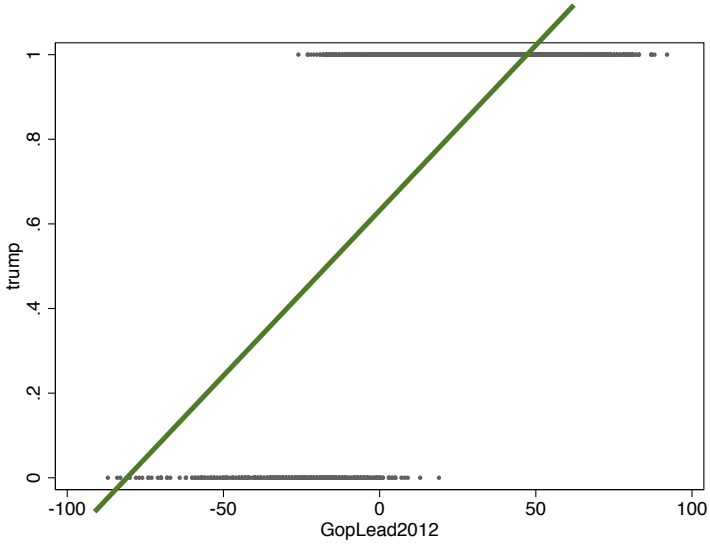


2016



How would we use a regression to predict?

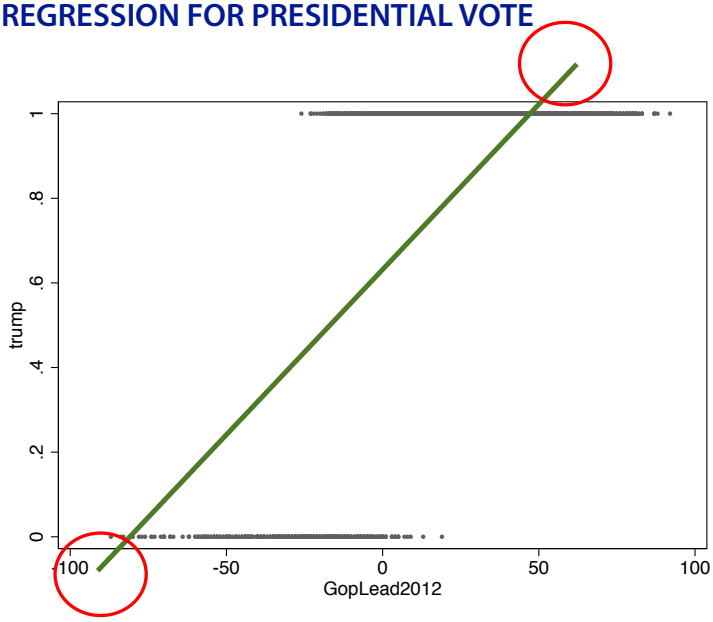
PRESIDENTIAL VOTE FOR EACH COUNTY



trump = A + B • GopLead2012

The regression approach has several problems

REGRESSION FOR PRESIDENTIAL VOTE



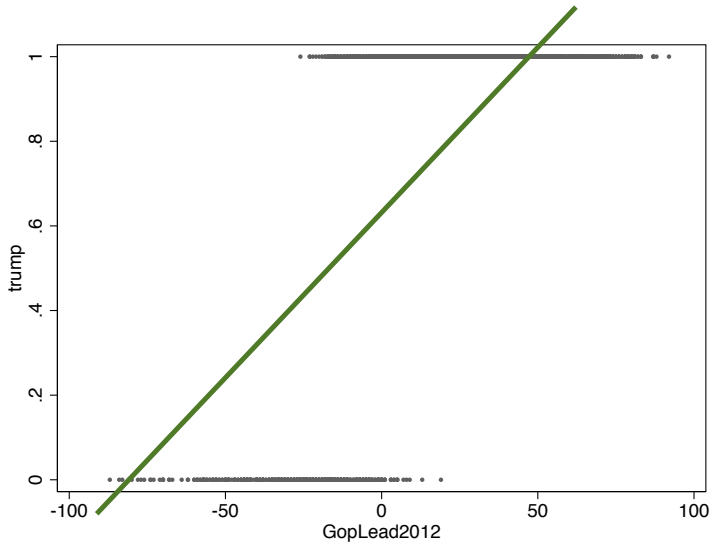
trump = A + B • GopLead2012

Problems:

1. "out-of-range" predictions

The regression approach has several problems

REGRESSION FOR PRESIDENTIAL VOTE



trump = A + B • GopLead2012

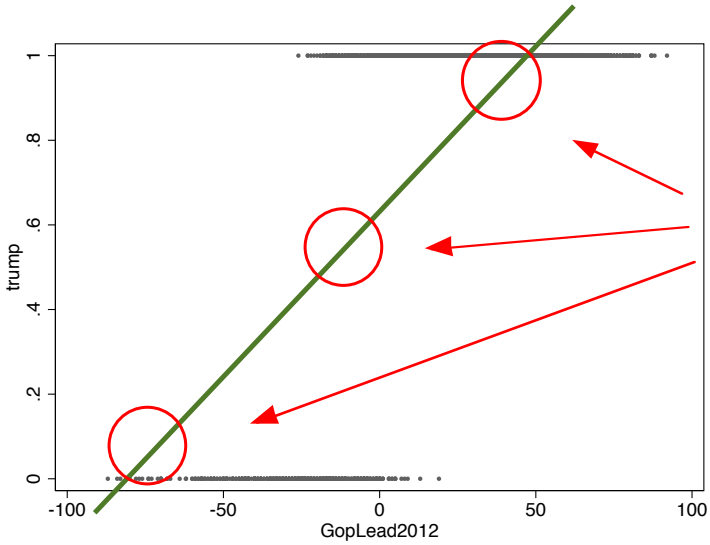
Problems:

1. "out-of-range" predictions

2. What else?

The regression approach has several problems

REGRESSION FOR PRESIDENTIAL VOTE



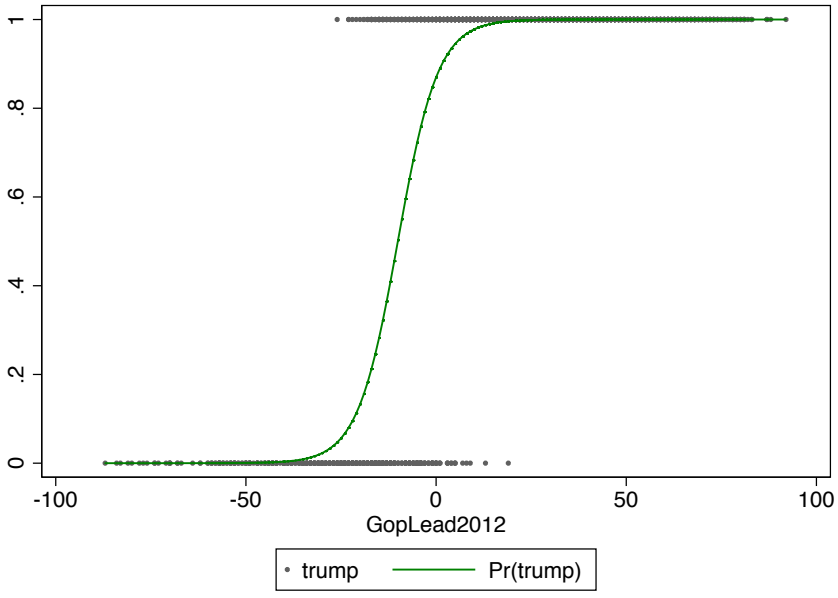
trump = A + B • GopLead2012

Problems:

1. "out-of-range" predictions
2. Linear effect of conservatism

We would like a method that corrects the shortcomings of regression

"IDEAL" PROBABILITY PREDICTION



Logistic regression is a flexible way to predict binary choices

PROPERTIES OF LOGISTIC REGRESSION

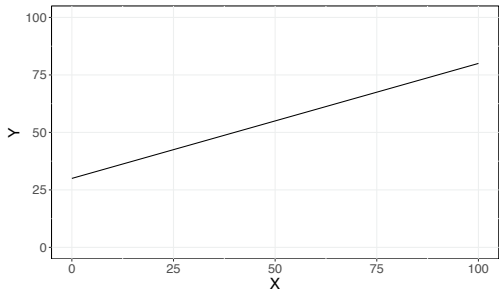
Also known as **logit regression** and the **logit model**

- Used when the dependent variable is binary
  - Buy / do not buy (purchase choice models)
  - Left / stayed (attrition, churn models)
  - Failed / did not fail (predictive maintenance)
- From data science point of view, works similar to regular regression
  - Can include many different variables
  - Fast
  - One of the most popular approaches used in data science

The logistic regression model allows us to easily estimate probabilities

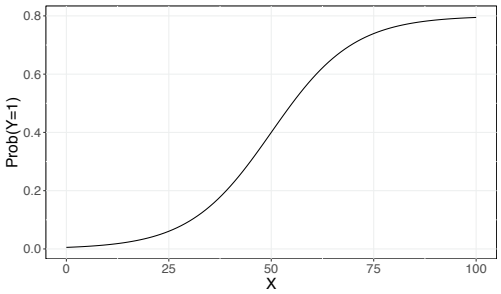
COMPARISON OF REGRESSION APPROACHES

OLS (regular) Regression



$$Y = a + b\,X$$

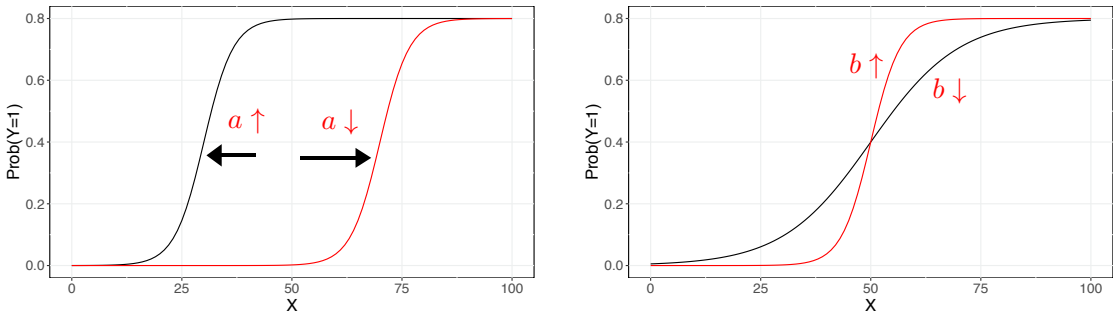
Logistic Regression



$$\text{Prob}(Y = 1) = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

The coefficients change the shape of the logistic regression model

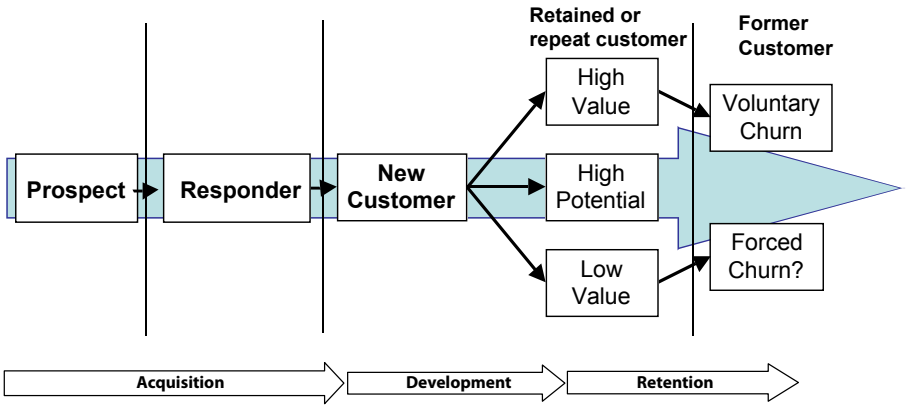
EFFECT OF "a" and "b" COEFFICIENT



$$\text{Prob}(Y = 1) = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

Logistic regression can be used at several stages of the customer lifecycle

APPLICATIONS OF LOGISTIC REGRESSION



- Modeling and predicting response rates
- Modeling and predicting the success of cross/ up-sell attempts
- Modeling and predicting customer attrition
- Modeling and predicting "reactivation" success

In R we use the "glm" command to run a logistic regression

PREDICTING THE PRESIDENTIAL ELECTION BY COUNTY

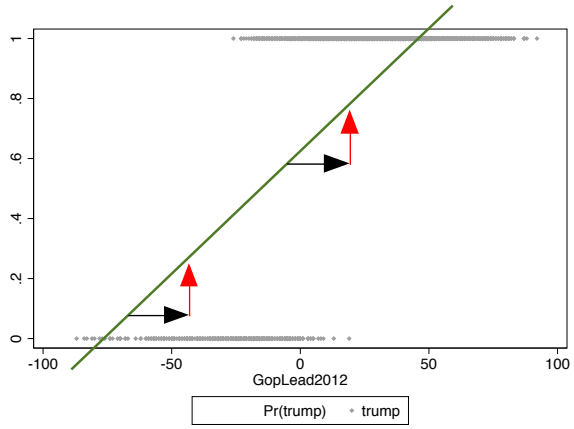
- **Dependent** variable: **trump** -- "*wins*" (1) or "*loses*" (0) **by county** (3112)
- **Predictor** variable: GOP – DEM 2012 % vote difference ("GopLead2012")

```
logit1 <- glm(trump ~ GopLead2012,family=binomial(logit), data=trump2016)
summary(logit1)
```

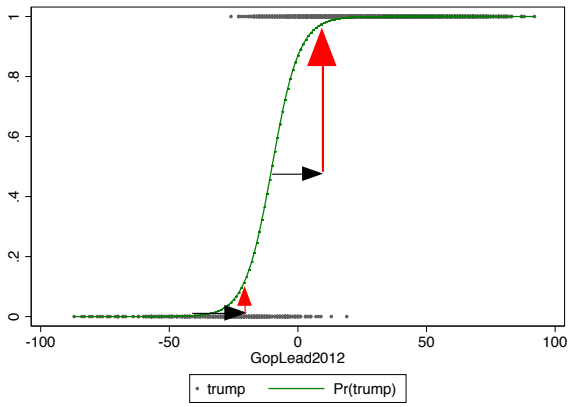
```
Call:
glm(formula = trump ~ GopLead2012, family = binomial(logit),
    data = trump2016)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3147   0.0023   0.0200   0.1096   2.4740

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.89986    0.12101   15.70   <2e-16 ***
GopLead2012  0.18893    0.01007   18.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic coefficients are harder to interpret than "normal" regression coefficients

OLS REGRESSION



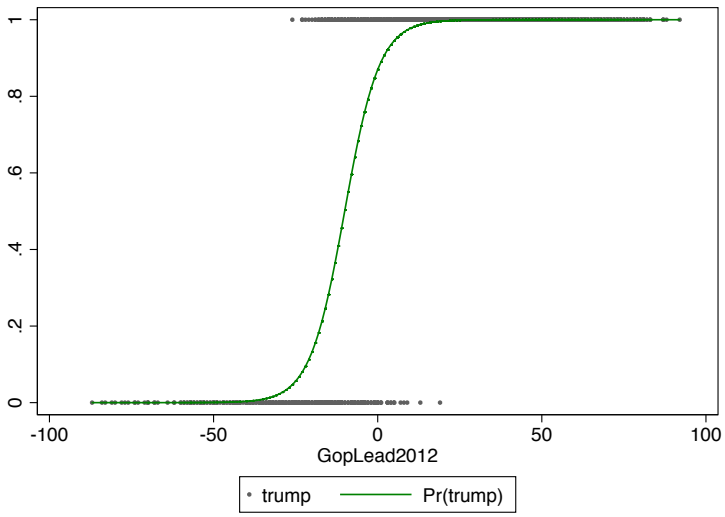
LOGISTIC REGRESSION



- An increase of 1 percentage point in GopLead2012 does not have a constant effect on the dependent variable (predicted probability of Trump winning a county)
- Instead of interpreting coefficients directly, we will plot "marginal effects" (later...)

How do we predict using a logistic regression?

"IDEAL" PROBABILITY PREDICTION



Regression example:

YFC = A + B * Weight

YFC = 370 + 0.675*Weight

Logistic regression example:

$$\text{Prob(vote)} = \frac{e^{A+B\cdot \text{GopLead2012}}}{1 + e^{A+B\cdot \text{GopLead2012}}}$$

$$\text{Prob(vote)} = \frac{e^{1.9+0.19\cdot \text{GopLead2012}}}{1 + e^{1.9+0.19\cdot \text{GopLead2012}}}$$

The prediction can easily be made in Excel

HOW TO PREDICT WITH A LOGISTIC REGRESSION

$$\text{Prob(vote)} = \frac{e^{1.9+0.19\cdot \text{GopLead2012}}}{1 + e^{1.9+0.19\cdot \text{GopLead2012}}}$$

From a statistical program:
R, SAS, Stata, ...

Coefficients	
A	1.899865
B	0.188933
GopLead2012	-50
Logit Formula	=EXP(B2+B3*B5)/(1+EXP(B2+B3*B5))
(evaluated)	**0.001**
	0.1% ← Prediction

In R the prediction is automatically saved in the model

PREDICTED PROBABILITIES IN R

name of saved model

```
logit1 <- glm(trump ~ GopLead2012,family=binomial(logit), data=trump2016)
summary(logit1)
```

Append predictions to dataset

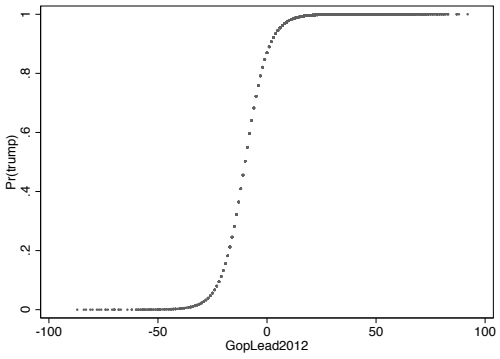
```
trump2016 <- trump2016 %>%
  mutate(pred_prob=predict(logit1, type = "response"))
```

predictions ("response probabilities")

In R the prediction is automatically saved in the model

PREDICTED PROBABILITIES in "trump2016"

	county	trump	GopLead2012	pred_prob
1	Delta County	1	7	9.616684e-01
2	Lipscomb County	1	79	1.000000e+00
3	Walker County	1	53	9.999933e-01
4	Reeves County	0	-16	2.454435e-01
5	Hot Springs County	1	55	9.999954e-01
6	Doniphan County	1	45	9.999696e-01
7	Wake County	0	-10	5.026337e-01
8	Pitt County	0	-7	6.404537e-01
9	Wise County	1	67	9.999995e-01
10	Russell County	1	62	9.999988e-01
11	Grant County	1	52	9.999919e-01
12	Richardson County	1	33	9.997069e-01
13	Monroe County	1	45	9.999696e-01
14	Nassau County	0	-7	6.404537e-01
15	Bullock County	0	-53	2.993539e-04
16	Fulton County	1	57	9.999969e-01



Back to predictive lead scoring ...

id	webpageviews	viewedpricing	highrelevancefirm	sale
639	15	0	1	0
272	35	0	1	1
491	7	0	1	0
226	18	1	1	0
7195	13	0	0	0
9080	23	0	0	0
548	14	0	1	0
9605	36	0	0	0
5352	28	0	0	0
4343	35	0	0	0
14971	3	0	0	0
11298	34	0	0	1
317	10	0	1	0
…	…	…	…	…

id	webpageviews	viewedpricing	highrelevancefirm	sale
7236	27	0	0	
687	25	1	1	
453	16	0	0	
563	6	0	1	

R Demo