Introduction to Machine Learning for Social Scientists

Class 6: Classification

Edgar Franco Vivanco

Stanford University
Department of Political Science

edgarf1@stanford.edu

Summer 2018

Where are you struggling?

Mini survey results:

- Functions
- Subsetting (using [])
- Difference between linear regression and logistic regression

Where are you struggling?

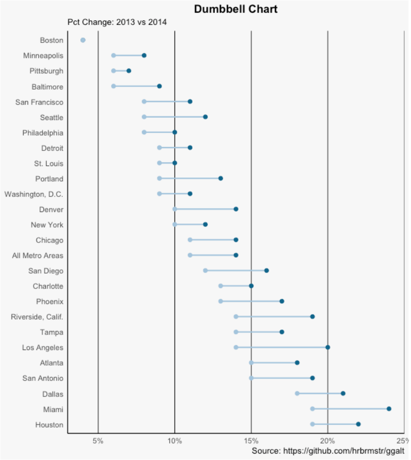
Mini survey results:

- Functions
- Subsetting (using [])
- Difference between linear regression and logistic regression

Tutorials available before midterm

Extra workshops

- ggplot!!
- data manipulation
- text analysis



Other petitions

Mini survey results:

- Connection with Machine Learning:

Other petitions

Mini survey results:

- Connection with Machine Learning:
- Next class will study an application of these methods
- Other fields:

Other petitions

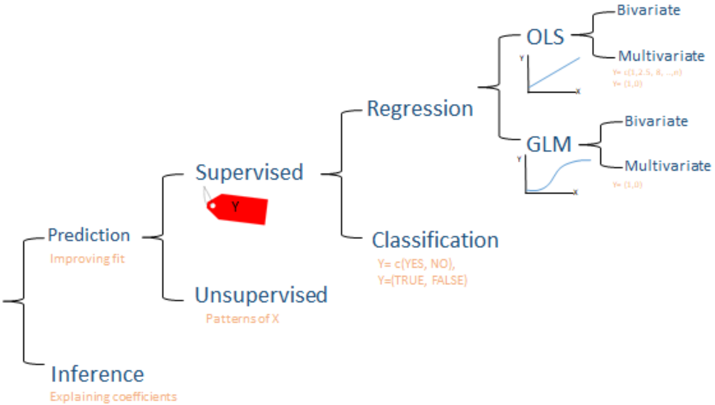
Mini survey results:

- Connection with Machine Learning:
- Next class will study an application of these methods
- Other fields:
- Fake news, Psychology, Sociology, etc.

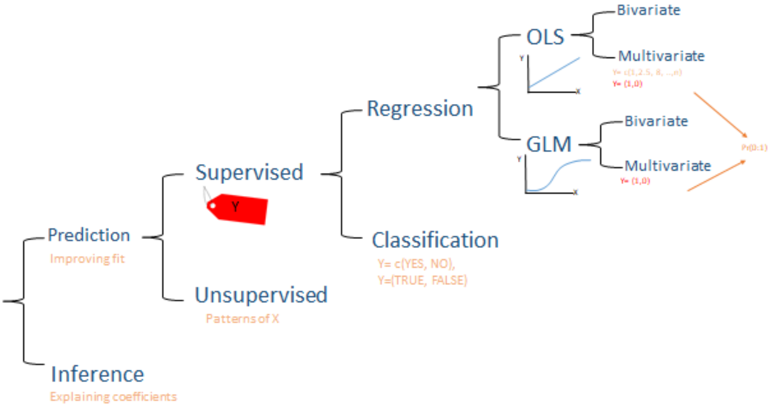
Today's Goals

1. Key concepts:
   - Linear Probability Model vs. Generalized Linear Model
   - Classification
   - Confusion Matrix
   - Performance measures
2. Key techniques and R functions:
   - ifelse
   - table

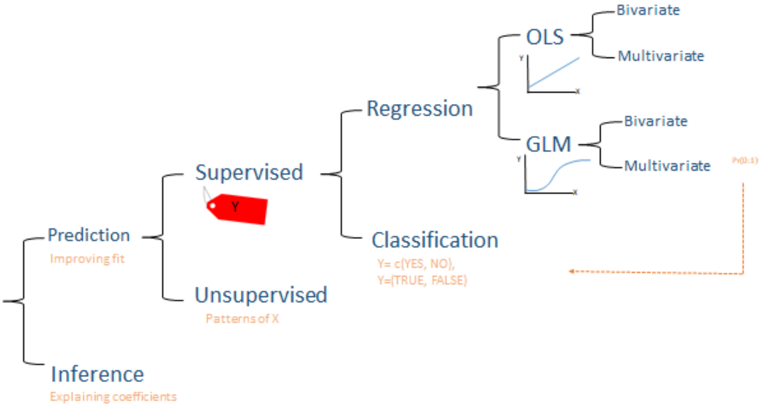
Our Mental Map: OLS and GLM



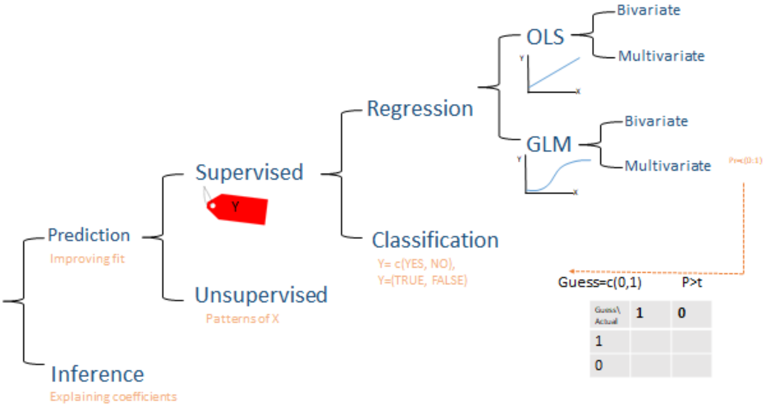
Our Mental Map: Predicting probabilities



Our Mental Map: Classify

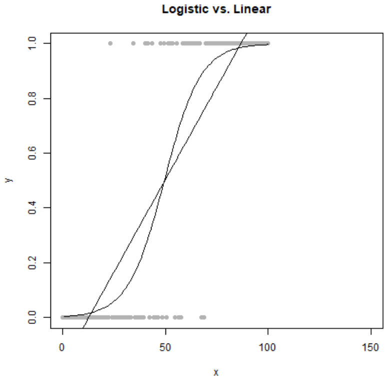


Our Mental Map: Test our model



Overview

From Prediction to classification



- If we have a qualitative outcome (Y=0 or Y=1) we can predict probabilities using a linear or a logistic model.

From Prediction to classification

Filter

	y	state.abb	name	rep	state.name	gorevote
1	1	AL	SESSIONS (R AL)	TRUE	Alabama	41.59
2	1	AL	SHELBY (R AL)	TRUE	Alabama	41.59
3	1	AK	MURKOWSKI (R AK)	TRUE	Alaska	27.67
4	1	AK	STEVENS (R AK)	TRUE	Alaska	27.67
5	1	AZ	KYL (R AZ)	TRUE	Arizona	44.67
6	1	AZ	MCCAIN (R AZ)	TRUE	Arizona	44.67
7	1	AR	HUTCHINSON (R AR)	TRUE	Arkansas	45.86
8	1	AR	LINCOLN (D AR)	FALSE	Arkansas	45.86
9	0	CA	BOXER (D CA)	FALSE	California	53.45
10	1	CA	FEINSTEIN (D CA)	FALSE	California	53.45
11	1	CO	ALLARD (R CO)	TRUE	Colorado	42.39
12	1	CO	CAMPBELL (R CO)	TRUE	Colorado	42.39
13	1	CT	DODD (D CT)	FALSE	Connecticut	55.91
14	1	CT	LIEBERMAN (D CT)	FALSE	Connecticut	55.91
15	1	DE	BIDEN (D DE)	FALSE	Delaware	54.96
16	1	DE	CARPER (D DE)	FALSE	Delaware	54.96
17	0	FL	GRAHAM (D FL)	FALSE	Florida	48.84
18	1	FL	NELSON (D FL)	FALSE	Florida	48.84

Showing 1 to 19 of 100 entries

- If we have a qualitative outcome (Y=0 or Y=1) we can predict probabilities using a linear or a logistic model.
- In our example:
  - Y: Vote for Iraq War (YES=1, NO=0)
  - rep: Senator is Republican
  - gorevote: Percentage of vote for Al Gore in Senator's state

From Prediction to classification

```
> fit <- lm(y ~ rep + gorevote, data = iraqVote)
> summary(fit)

Call:
lm(formula = y ~ rep + gorevote, data = iraqVote)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7654 -0.1533  0.0509  0.2904  0.5707

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.174458   0.236256   4.971 2.87e-06 ***
repTRUE      0.316933   0.080493   3.937 0.000155 ***
gorevote    -0.012376   0.004715  -2.625 0.010072 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3603 on 97 degrees of freedom
Multiple R-squared:  0.2888,    Adjusted R-squared:  0.2742
F-statistic:  19.7 on 2 and 97 DF,  p-value: 6.617e-08
```

- We can run a linear model
- $p(Y = 1|X) = \beta_0 + \beta_1 rep + \beta_2 gorevote$

From Prediction to classification

```
> fit <- lm(y ~ rep + gorevote, data = iraqVote)
> summary(fit)

Call:
lm(formula = y ~ rep + gorevote, data = iraqVote)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7654 -0.1533  0.0509  0.2904  0.5707

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.174458   0.236256   4.971 2.87e-06 ***
repTRUE      0.316933   0.080493   3.937 0.000155 ***
gorevote    -0.012376   0.004715  -2.625 0.010072 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3603 on 97 degrees of freedom
Multiple R-squared:  0.2888,    Adjusted R-squared:  0.2742
F-statistic:  19.7 on 2 and 97 DF,  p-value: 6.617e-08
```

- We can run a linear model
- $p(Y = 1|X) = \beta_0 + \beta_1 rep + \beta_2 gorevote$
- And calculate predictions:
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 rep + \hat{\beta}_2 gorevote$

From Prediction to classification

```
> fit <- lm(y ~ rep + gorevote, data = iraqVote)
> summary(fit)

Call:
lm(formula = y ~ rep + gorevote, data = iraqVote)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7654 -0.1533  0.0509  0.2904  0.5707

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.174458   0.236256   4.971 2.87e-06 ***
repTRUE      0.316933   0.080493   3.937 0.000155 ***
gorevote    -0.012376   0.004715  -2.625 0.010072 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3603 on 97 degrees of freedom
Multiple R-squared:  0.2888,    Adjusted R-squared:  0.2742
F-statistic:  19.7 on 2 and 97 DF,  p-value: 6.617e-08
```

- We can run a linear model
- $p(Y = 1|X) = \beta_0 + \beta_1 rep + \beta_2 gorevote$
- And calculate predictions:
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 rep + \hat{\beta}_2 gorevote$
- $\hat{Y} = 1.144 + 0.3169 rep - 0.0123 gorevote$

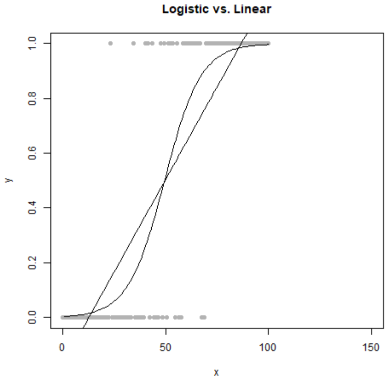
From Prediction to classification

Filter

	y	state.abb	name	rep	state.name	gorevote	pred_prob_lm
1	1	AL	SESSIONS (R AL)	TRUE	Alabama	41.59	0.9766924
2	1	~~AL~~	~~SHELBY (R AL)~~	~~TRUE~~	~~Alabama~~	~~41.59~~	0.9766924
3	1	AK	MURKOWSKI (R AK)	TRUE	Alaska	27.67	1.1489597
4	1	AK	STEVENS (R AK)	TRUE	Alaska	27.67	1.1489597
5	1	AZ	KYL (R AZ)	TRUE	Arizona	44.67	0.9385758
6	1	AZ	MCCAIN (R AZ)	TRUE	Arizona	44.67	0.9385758
7	1	AR	HUTCHINSON (R AR)	TRUE	Arkansas	45.86	0.9238490
8	1	AR	LINCOLN (D AR)	FALSE	Arkansas	45.86	0.6069163
9	0	CA	BOXER (D CA)	FALSE	California	53.45	0.5129861
10	1	CA	FEINSTEIN (D CA)	FALSE	California	53.45	0.5129861
11	1	CO	ALLARD (R CO)	TRUE	Colorado	42.39	0.9667920
12	1	CO	CAMPBELL (R CO)	TRUE	Colorado	42.39	0.9667920
13	1	CT	DODD (D CT)	FALSE	Connecticut	55.91	0.4825424
14	1	CT	LIEBERMAN (D CT)	FALSE	Connecticut	55.91	0.4825424
15	1	DE	BIDEN (D DE)	FALSE	Delaware	54.96	0.4942991
16	1	DE	CARPER (D DE)	FALSE	Delaware	54.96	0.4942991
17	0	FL	GRAHAM (D FL)	FALSE	Florida	48.84	0.5700373
18	1	FL	NELSON (D FL)	FALSE	Florida	48.84	0.5700373
19	1	GA	CLELAND (D GA)	FALSE	Georgia	42.98	0.6425578
20	1	GA	MILLER (D GA)	FALSE	Georgia	42.98	0.6425578

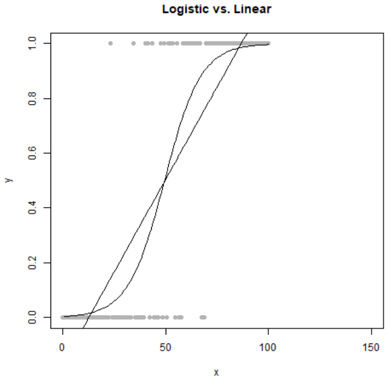
- We can run a linear model
- $p(Y = 1|X) = \beta_0 + \beta_1 rep + \beta_2 gorevote$
- And calculate predictions:
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 rep + \hat{\beta}_2 gorevote$
- $\hat{Y} = 1.1744 + 0.3169 rep - 0.0123 gorevote$
- $0.9766 = 1.1744 + 0.3169 - 0.0123 * 41.59$

From Prediction to classification



- A logistic model will produce predictions between 0 and 1.

From Prediction to classification



- A logistic model will produce predictions between 0 and 1.

From Prediction to classification

```
> rep_reg_glm <- glm(y~rep+gorevote, family = binomial, data = iraqVote)
> summary(rep_reg_glm)

Call:
glm(formula = y ~ rep + gorevote, family = binomial, data = iraqVote)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.12054   0.07761   0.19676   0.59926   1.59277

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.87859    2.27506   2.584  0.00977 **
repTRUE      3.01881    1.07138   2.818  0.00484 **
gorevote    -0.11322    0.04508  -2.512  0.01201 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 107.855  on 99  degrees of freedom
Residual deviance:  71.884  on 97  degrees of freedom
AIC: 77.884

Number of Fisher Scoring iterations: 6
```

- A logistic model will produce predictions between 0 and 1.
- Because it models a relationship:

$$p(X) = \frac{1}{1 + exp^{-\beta X}}$$

From Prediction to classification

```
> rep_reg_glm <- glm(y~rep+gorevote, family = binomial, data = iraqVote)
> summary(rep_reg_glm)

Call:
glm(formula = y ~ rep + gorevote, family = binomial, data = iraqVote)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.12054   0.07761   0.19676   0.59926   1.59277

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.87859    2.27506   2.584  0.00977 **
repTRUE      3.01881    1.07138   2.818  0.00484 **
gorevote    -0.11322    0.04508  -2.512  0.01201 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 107.855  on 99  degrees of freedom
Residual deviance:  71.884  on 97  degrees of freedom
AIC: 77.884

Number of Fisher Scoring iterations: 6
```

- A logistic model will produce predictions between 0 and 1.
- Because it models a relationship:

$$p(X) = \frac{1}{1 + exp^{-\beta X}}$$

Optimized via Maximum Likelihood

- 4.18 = 5.88+3.021-0.113*41.59

From Prediction to classification

```
> rep_reg_glm <- glm(y~rep+gorevote, family = binomial, data = iraqVote)
> summary(rep_reg_glm)

Call:
glm(formula = y ~ rep + gorevote, family = binomial, data = iraqVote)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.12054   0.07761   0.19676   0.59926   1.59277

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.87859    2.27506   2.584  0.00977 **
repTRUE      3.01881    1.07138   2.818  0.00484 **
gorevote    -0.11322    0.04508  -2.512  0.01201 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 107.855  on 99  degrees of freedom
Residual deviance:  71.884  on 97  degrees of freedom
AIC: 77.884

Number of Fisher Scoring iterations: 6
```

- A logistic model will produce predictions between 0 and 1.
- Because it models a relationship:

$$p(X) = \frac{1}{1 + exp^{-\beta X}}$$

Optimized via Maximum Likelihood

- 4.18 = 5.88+3.021-0.113*41.59
- 0.985$= \frac{1}{1 + exp^{-4.18}}$

Call $p_i = \Pr(\text{Vote}_i = 1 | \boldsymbol{x}_i)$

$$\begin{aligned}
\text{Vote}_i &\sim \text{Bernoulli}(p_i) \\
p_i &= f(\boldsymbol{\beta} \cdot \boldsymbol{x}_i) \\
\log\left(\frac{p_i}{1 - p_i}\right) &= \boldsymbol{\beta} \cdot \boldsymbol{x}_i \\
p_i &= \frac{\exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta} \cdot \boldsymbol{x}_i)} \\
&= \frac{1}{1 + \exp(-\boldsymbol{\beta} \cdot \boldsymbol{x}_i)}
\end{aligned}$$

Important functions:

$$\begin{aligned}
\text{odds}(p) &= \frac{p}{1 - p} \\
\text{log odds or logit}(p) &= \log\left(\frac{p}{1 - p}\right) \\
\text{logistic function or logit}^{-1}(a) &= \frac{1}{1 + \exp(-a)}
\end{aligned}$$

From Prediction to classification

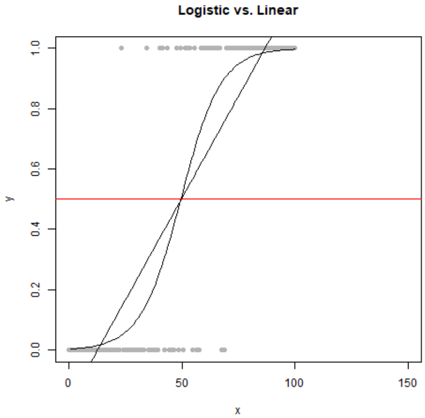
	y	state.abb	name	rep	state.name	gorevote	pred_prob_lm	pred_prob_glm
1	1	AL	SESSIONS (R AL)	TRUE	Alabama	41.59	0.9766924	0.9850607
2	1	AL	SHELBY (R AL)	TRUE	Alabama	41.59	0.9766924	0.9850607
3	1	AK	MURKOWSKI (R AK)	TRUE	Alaska	27.67	1.1489597	0.9968734
4	1	AK	STEVENS (R AK)	TRUE	Alaska	27.67	1.1489597	0.9968734
5	1	AZ	KYL (R AZ)	TRUE	Arizona	44.67	0.9385758	0.9789587
6	1	AZ	MCCAIN (R AZ)	TRUE	Arizona	44.67	0.9385758	0.9789587
7	1	AR	HUTCHINSON (R AR)	TRUE	Arkansas	45.86	0.9238490	0.9759968
8	1	AR	LINCOLN (D AR)	FALSE	Arkansas	45.86	0.6069163	0.6651797
9	0	CA	BOXER (D CA)	FALSE	California	53.45	0.5129861	0.4568942
10	1	CA	FEINSTEIN (D CA)	FALSE	California	53.45	0.5129861	0.4568942
11	1	CO	ALLARD (R CO)	TRUE	Colorado	42.39	0.9667920	0.9836676
12	1	CO	CAMPBELL (R CO)	TRUE	Colorado	42.39	0.9667920	0.9836676
13	1	CT	DODD (D CT)	FALSE	Connecticut	55.91	0.4825424	0.3890361
14	1	CT	LIEBERMAN (D CT)	FALSE	Connecticut	55.91	0.4825424	0.3890361
15	1	DE	BIDEN (D DE)	FALSE	Delaware	54.96	0.4942991	0.4148843
16	1	DE	CARPER (D DE)	FALSE	Delaware	54.96	0.4942991	0.4148843
17	0	FL	GRAHAM (D FL)	FALSE	Florida	48.84	0.5700373	0.5863938
18	1	FL	NELSON (D FL)	FALSE	Florida	48.84	0.5700373	0.5863938

- A logistic model will produce predictions between 0 and 1.
- Because it models a relationship:

$$p(X) = \frac{1}{1 + exp^{-\beta X}}$$

- 4.18 = 5.88+3.021-0.113*41.59
- 0.985$= \frac{1}{1 + exp^{-4.18}}$

How to create classifications?



- We can choose a threshold such as:

$$Pr(Y \hat{=} 1|X) >= t$$

Then clas=1, and 0 otherwise

How to create classifications?

```
> ifelse(pred_prob_glm>=.5, 1,0)
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
  1   1   1   1   1   1   1   1   0   0   1   1   0   0   0   0   0   0   1   1
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
  0   0   1   1   0   1   1   1   1   0   1   1   1   1   1   1   1   1   0   0
 41  42  43  44  45  46  48  49  50  51  52  53  54  55  56  57  58  59  60  61
  0   0   0   0   0   0   1   1   1   1   1   1   1   1   1   1   1   1   0   0
 62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81
  0   1   0   0   1   1   1   1   1   1   1   1   1   1   1   1   1   0   1   1
 82  83  84  85  86  87  88  89  91  92  93  94  95  96  97  98  99 100 101 102
  1   1   1   1   1   1   1   1   0   0   1   1   0   0   1   1   0   0   1   1
```

- We can choose a threshold such as:

  $$Pr(Y \hat{=} 1|X) >= t$$

  Then clas=1, and 0 otherwise
- We can do this by using the function 'ifelse()'

How to create classifications?

Filter

	y	state.abb	name	rep	state.name	gorevote	pred_prob_lm	pred_prob_glm	class_lm	class_glm
1	1	AL	SESSIONS (R AL)	TRUE	Alabama	41.59	0.9766924	0.9850607	1	1
2	1	AL	SHELBY (R AL)	TRUE	Alabama	41.59	0.9766924	0.9850607	1	1
3	1	AK	MURKOWSKI (R AK)	TRUE	Alaska	27.67	1.1489597	0.9968734	1	1
4	1	AK	STEVENS (R AK)	TRUE	Alaska	27.67	1.1489597	0.9968734	1	1
5	1	AZ	KYL (R AZ)	TRUE	Arizona	44.67	0.9385758	0.9789587	1	1
6	1	AZ	MCCAIN (R AZ)	TRUE	Arizona	44.67	0.9385758	0.9789587	1	1
7	1	AR	HUTCHINSON (R AR)	TRUE	Arkansas	45.86	0.9238490	0.9759968	1	1
8	1	AR	LINCOLN (D AR)	FALSE	Arkansas	45.86	0.6069163	0.6651797	1	1
9	0	CA	BOXER (D CA)	FALSE	California	53.45	0.5129861	0.4568942	1	0
10	1	CA	FEINSTEIN (D CA)	FALSE	California	53.45	0.5129861	0.4568942	1	0
11	1	CO	ALLARD (R CO)	TRUE	Colorado	42.39	0.9667920	0.9836676	1	1
12	1	CO	CAMPBELL (R CO)	TRUE	Colorado	42.39	0.9667920	0.9836676	1	1
13	1	CT	DODD (D CT)	FALSE	Connecticut	55.91	0.4825424	0.3890361	0	0
14	1	CT	LIEBERMAN (D CT)	FALSE	Connecticut	55.91	0.4825424	0.3890361	0	0
15	1	DE	BIDEN (D DE)	FALSE	Delaware	54.96	0.4942991	0.4148843	0	0
16	1	DE	CARPER (D DE)	FALSE	Delaware	54.96	0.4942991	0.4148843	0	0
17	0	FL	GRAHAM (D FL)	FALSE	Florida	48.84	0.5700373	0.5863938	1	0
18	1	FL	NELSON (D FL)	FALSE	Florida	48.84	0.5700373	0.5863938	1	0
19	1	GA	CLELAND (D GA)	FALSE	Georgia	42.98	0.6425578	0.7335145	1	1
20	1	GA	MILLER (D GA)	FALSE	Georgia	42.98	0.6425578	0.7335145	1	1
21	0	HI	AKAKA (D HI)	FALSE	Hawaii	55.79	0.4840274	0.3922702	0	0

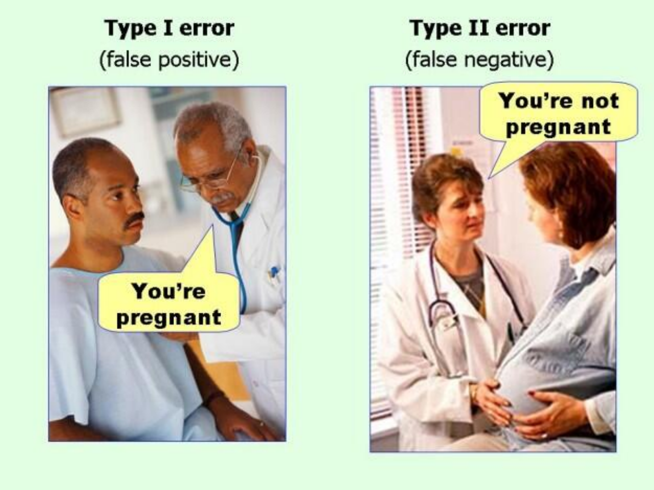
- We can choose a threshold such as:

$$Pr(Y \hat{=} 1|X) >= t$$

Then clas=1, and 0 otherwise

- We can do this by using the function 'ifelse()'
- And now we can start comparing our models with the observed values

Errors


Type I error
(false positive)
You're pregnant
Type II error
(false negative)
You're not pregnant

Confusion Matrix

To asses the quality of our data we compare our classifications with the real data or the "gold standard".

Actual \ Guess	Yes	No
Yes		
No		

Confusion Matrix

Actual \ Guess	Yes	No
Yes	True positive	False Negative
No	False Positive	True Negative

Confusion Matrix:

Code approach:
'ifelse()' function: ifelse(condition, yes, no)

```
# Actual yes and guess yes
tp <- ifelse (y ==1 & predicted==1,1,0)
# Actual no and guess n0
tn <- ifelse(y ==0 & predicted==0,1,0)
# Actual no and guess yes
fp <- ifelse(y ==0 & predicted==1,1,0)
# Actual yes and guess no
fn <- ifelse(y ==1 & predicted==0,1,0)
```

Accuracy

Accuracy is the percentage of observations classified correctly.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalseNegative + FalsePositive}$$

Precision

How many items classified as Yes are correctly classified?

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

It is equal to 1 if all the guesses as Yes are actually Yes.

Recall

How many items **that are actually** as Yes are correctly classified? In other words, is the number of correct results divided by the number of results that should have been returned.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

It is equal to 1 if all the actual Yes are classified as Yes.

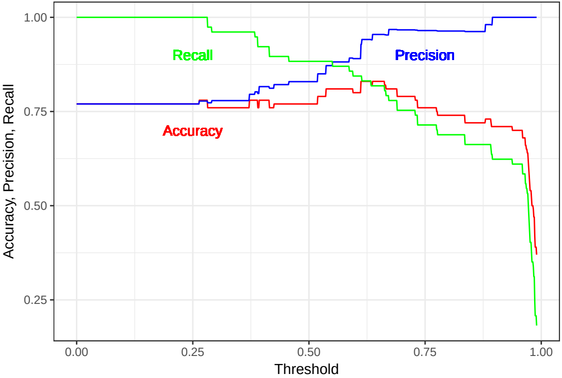
F-score

Harmonic mean of precision and recall:

$$F = \frac{2 * Precision * Recall}{Precision + Recall}$$

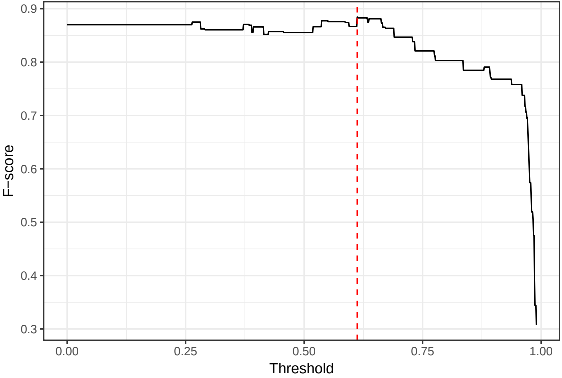
Performance

All these measures are function of the threshold.



F-score

We can find the threshold that optimizes the F-score.



Examples

- **Fraud in bank transactions**: High recall, ie. most of the fraudulent transactions are identified, probably at loss of precision.
- **Twitter:** If we are interested in finding out when a tweet expresses a negative sentiment, we can probably raise precision (to gain certainty).
- **Terrorist attacks:** Given the 800 million average passengers on US flights per year and the 19 (confirmed) terrorists who boarded US flights from 20002017, a very accurate model will predict everyone as non terrorist. Instead, we should focus on recall.

R!

Confusion Matrix: LM

Actual \ Guess	Yes	No
Yes	69	8
No	15	8

Confusion Matrix: GLM

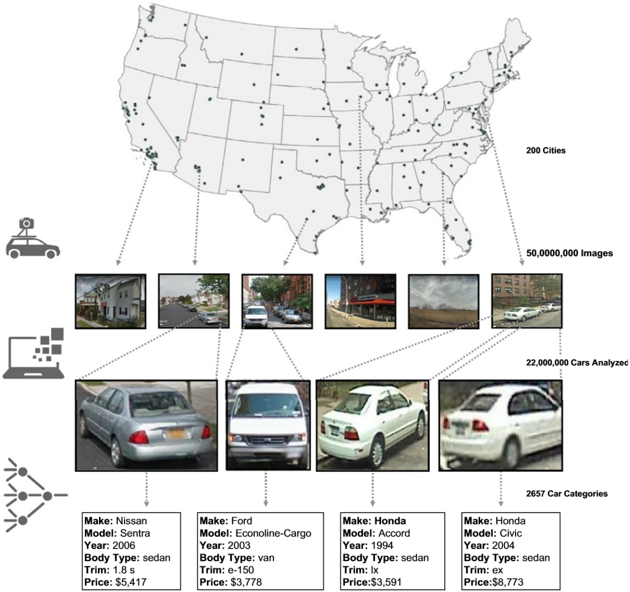
Actual \ Guess	Yes	No
Yes	68	9
No	14	9

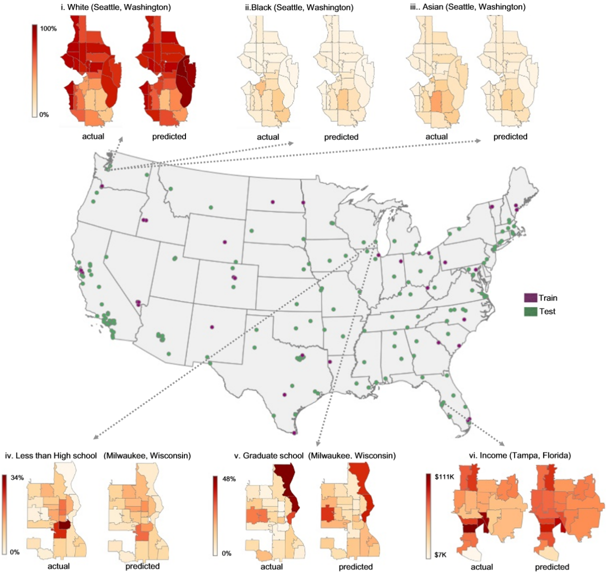
Results

- **LM:**
  - Accuracy: 0.77
  - Precision: 0.8214
  - Recall: 0.8961
  - F-score: 0.8571
- **Logistic**
  - Accuracy: 0.77
  - Precision: 0.8293
  - Recall: 0.8831
  - F-score: 0.8554

NEXT

- Resampling methods (Crossvalidation)
  - Training
  - Test
  - Validation
- Midterm guidelines
- Article


200 Cities
50,0000,000 Images
22,000,000 Cars Analyzed
2657 Car Categories
Make: Nissan
Model: Sentra
Year: 2006
Body Type: sedan
Trim: 1.8 s
Price: $5,417
Make: Ford
Model: Econoline-Cargo
Year: 2003
Body Type: van
Trim: e-150
Price: $3,778
Make: Honda
Model: Accord
Year: 1994
Body Type: sedan
Trim: lx
Price:$3,591
Make: Honda
Model: Civic
Year: 2004
Body Type: sedan
Trim: ex
Price:$8,773


i. White (Seattle, Washington)
ii.Black (Seattle, Washington)
iii.. Asian (Seattle, Washington)
100%
0%
actual
predicted
actual
predicted
actual
predicted
Train
Test
iv. Less than High school (Milwaukee, Wisconsin)
v. Graduate school (Milwaukee, Wisconsin)
vi. Income (Tampa, Florida)
34%
0%
48%
0%
$111K
$7K
actual
predicted
actual
predicted
actual
predicted