LR Advanced Topics

Dr. Goutam Chakraborty

Outline

- Interpretation of LR coefficients
  - Odds Ratio and *Doubling Amounts*
- Why do we need transformation of X (input) variables for LR models?
- How do we handle non-numeric X values in LR models?
- How do we handle non-linearity in LR models?

Logistic Regression Prediction Formula

- $$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 \cdot x_1 + \hat{w}_2 \cdot x_2$$ *logit scores*

...

Odds Ratios and Doubling Amounts

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 \cdot x_1 + \hat{w}_2 \cdot x_2$$ *logit scores*

Δx_i	consequence
1	$\Rightarrow$ *odds* $\times\ \exp(w_i)$
$\frac{0.69}{w_i}$	$\Rightarrow$ *odds* $\times\ \mathbf{2}$

Doubling amount: How much does an input need to change to double the odds?

Odds ratio: Amount odds change with a unit change in input.

...

SAS EM Output

Odds Ratio Estimates

Effect		Point Estimate
DemMedHomeValue		1.000
DemPctVeterans		1.007
GiftAvg36		0.990
GiftCnt36		1.059
GiftTimeLast		0.959
M_DemAge	0 vs 1	1.155
M_GiftAvgCard36	0 vs 1	1.253
PromCntCard12		0.963
StatusCat96NK	A vs S	0.957
StatusCat96NK	E vs S	1.481
StatusCat96NK	F vs S	0.633
StatusCat96NK	L vs S	1.179
StatusCat96NK	N vs S	0.898
StatusCatStarAll	0 vs 1	0.869

- For **GiftAvg36**, the odds ratio estimate equals 0.990. This means that for each additional dollar donated (on average) in the past 36 months, the odds of donation during the 97NK campaign change by a factor of 0.99, a 1% decrease.
- For **GiftCnt36**, the odds ratio estimate equals 1.059. This means that for each additional donation in the past 36 months, the odds of donation during the 97NK campaign change by a factor of 1.059, a 5.9% increase.
- For **M_DemAge**, the odds ratio (0 versus 1) estimate equals 1.155. This means that for cases with a 0 value for **M_DemAge**, the odds of donating are 1.155 times higher than the odds of donating for cases with a 1 value for **M_DemAge**.

Extreme Distributions and Regressions

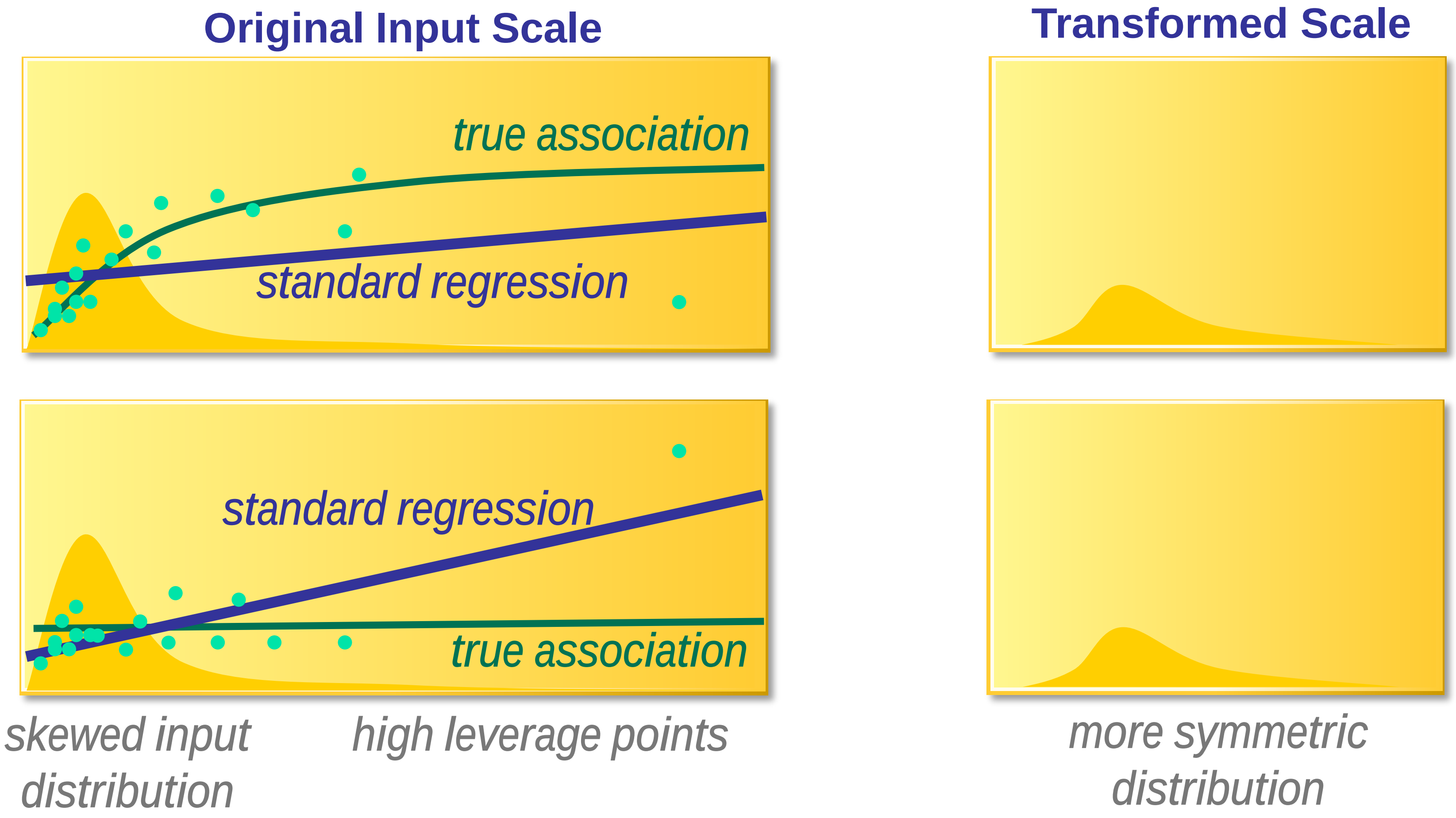
Original Input Scale

skewed input distribution

high leverage points

...

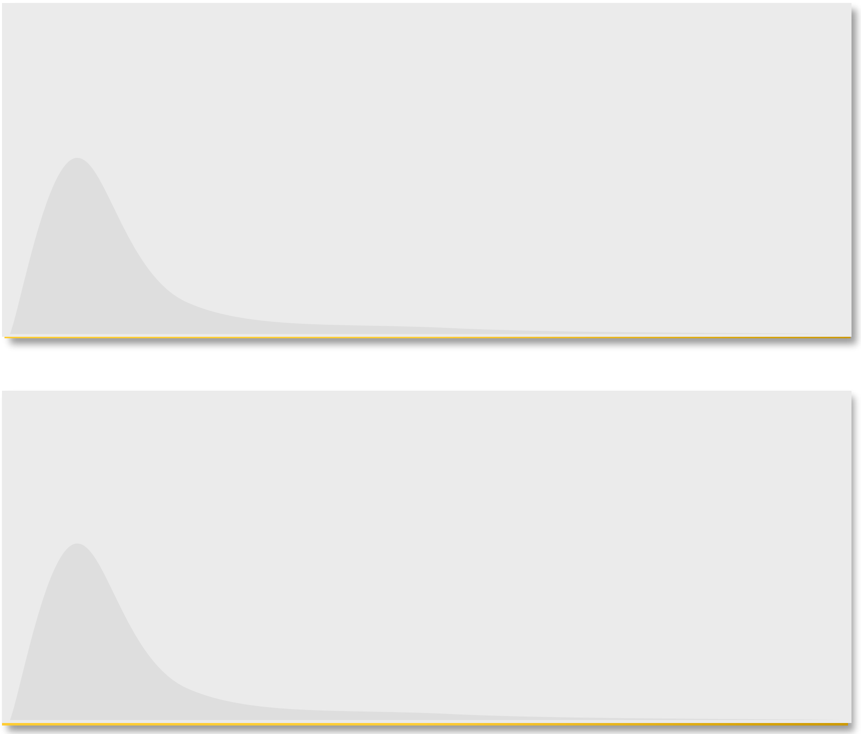
Extreme Distributions and Regressions



...

Input Transformations

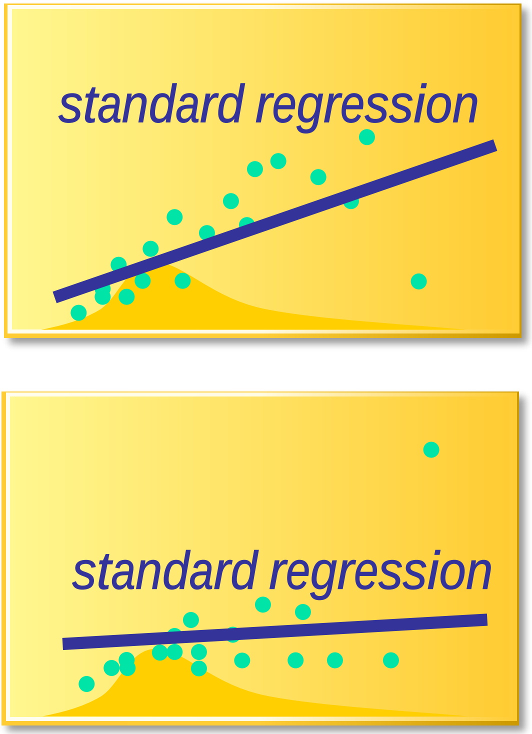
Original Input Scale



skewed input distribution

high leverage points

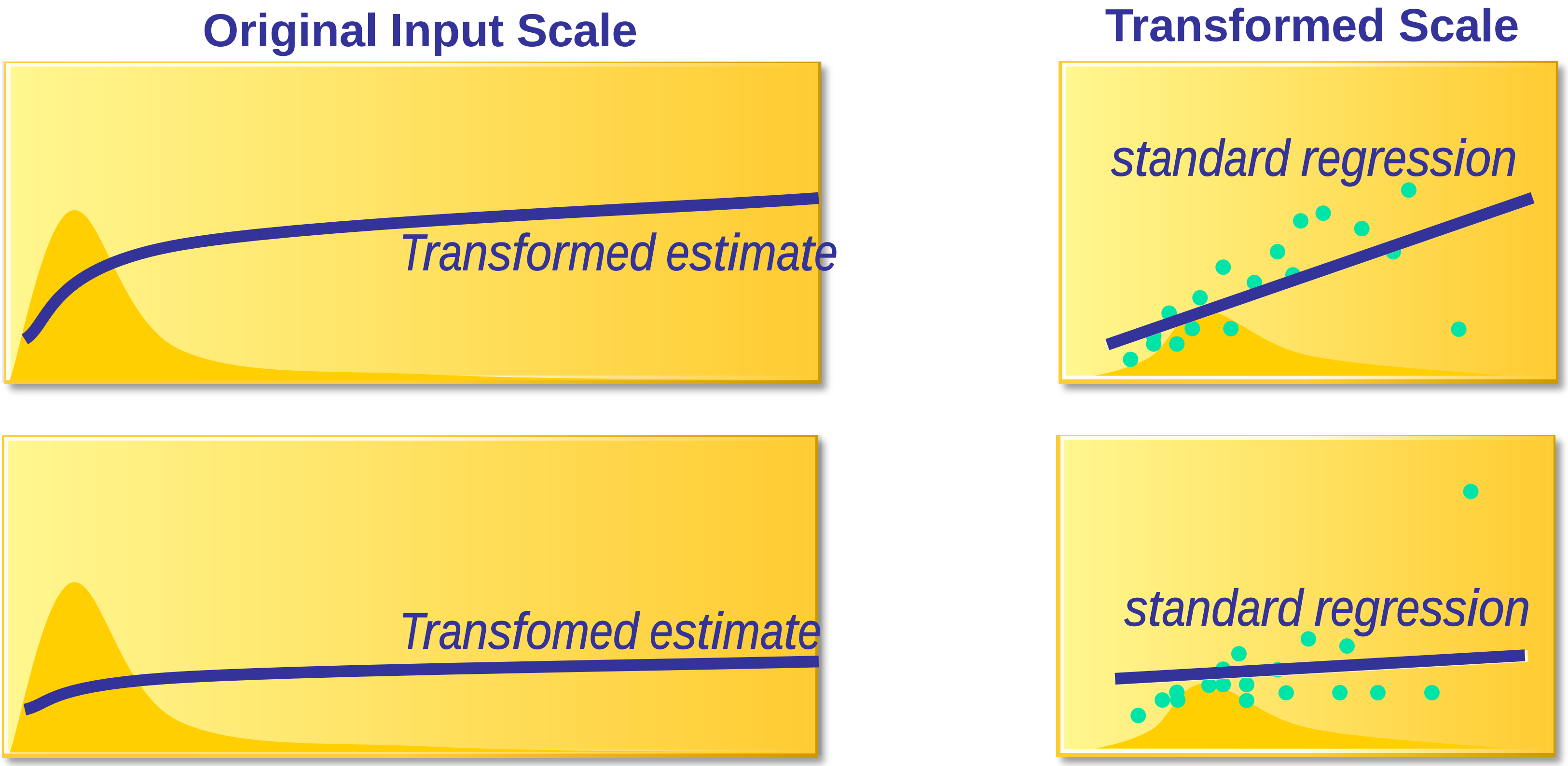
Transformed Scale



more symmetric distribution

...

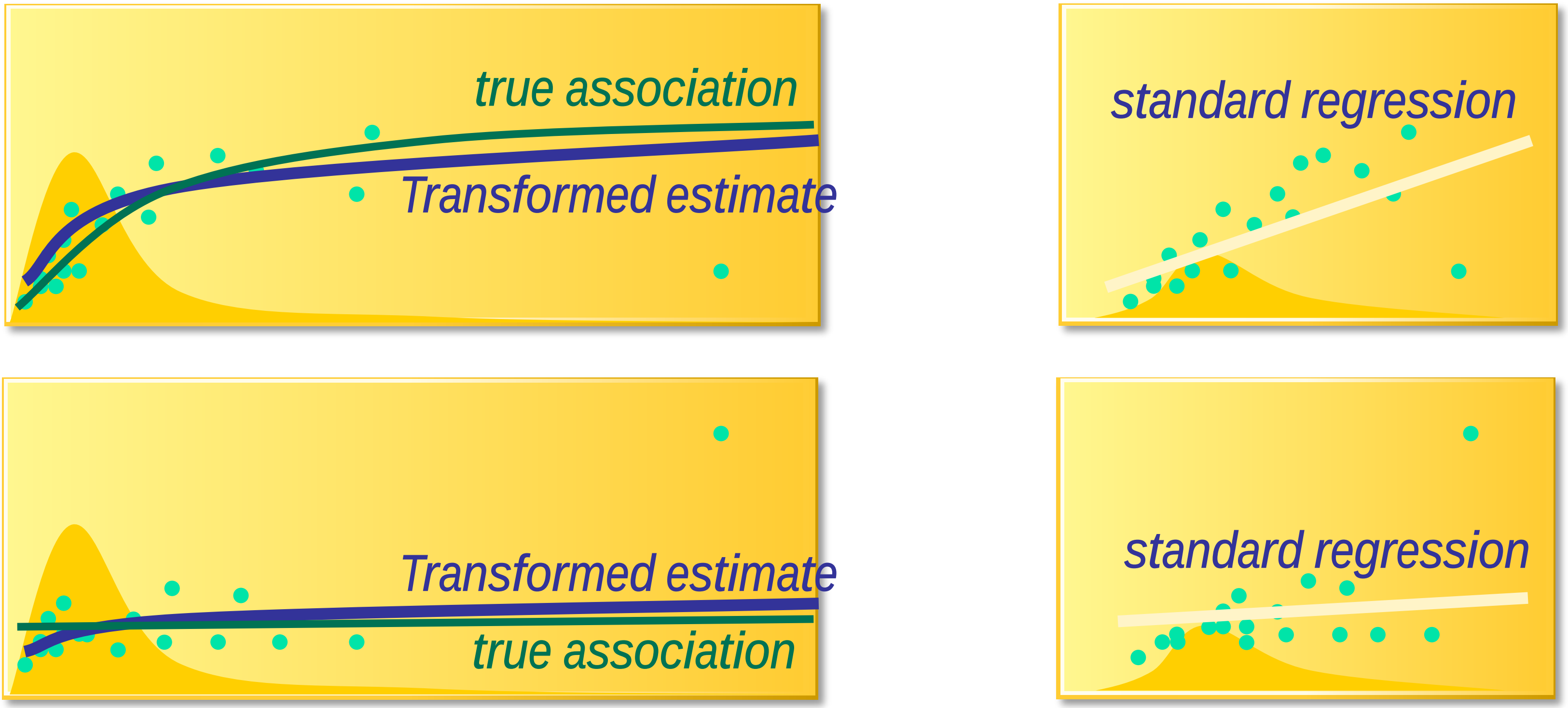
Regularizing Input Transformations



...

Regularizing Input Transformations

Original Input Scale | **Transformed Scale**



...

Nonnumeric Input Coding

Level	D_A	D_B	D_C	D_D	D_E	D_F	D_G	D_H	D_I
A	1	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0
E	0	0	0	0	1	0	0	0	0
F	0	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	1	0	0
H	0	0	0	0	0	0	0	1	0
I	0	0	0	0	0	0	0	0	1

...

Coding Redundancy

Level	D_A	D_B	D_C	D_D	D_E	D_F	D_G	D_H	D_I
A	1	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0
E	0	0	0	0	1	0	0	0	0
F	0	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	1	0	0
H	0	0	0	0	0	0	0	1	0
I	0	0	0	0	0	0	0	0	1

…

Coding Consolidation

Level	D_A	D_B	D_C	D_D	D_E	D_F	D_G	D_H	D_I
A	1	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0
E	0	0	0	0	1	0	0	0	0
F	0	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	1	0	0
H	0	0	0	0	0	0	0	1	0
I	0	0	0	0	0	0	0	0	1

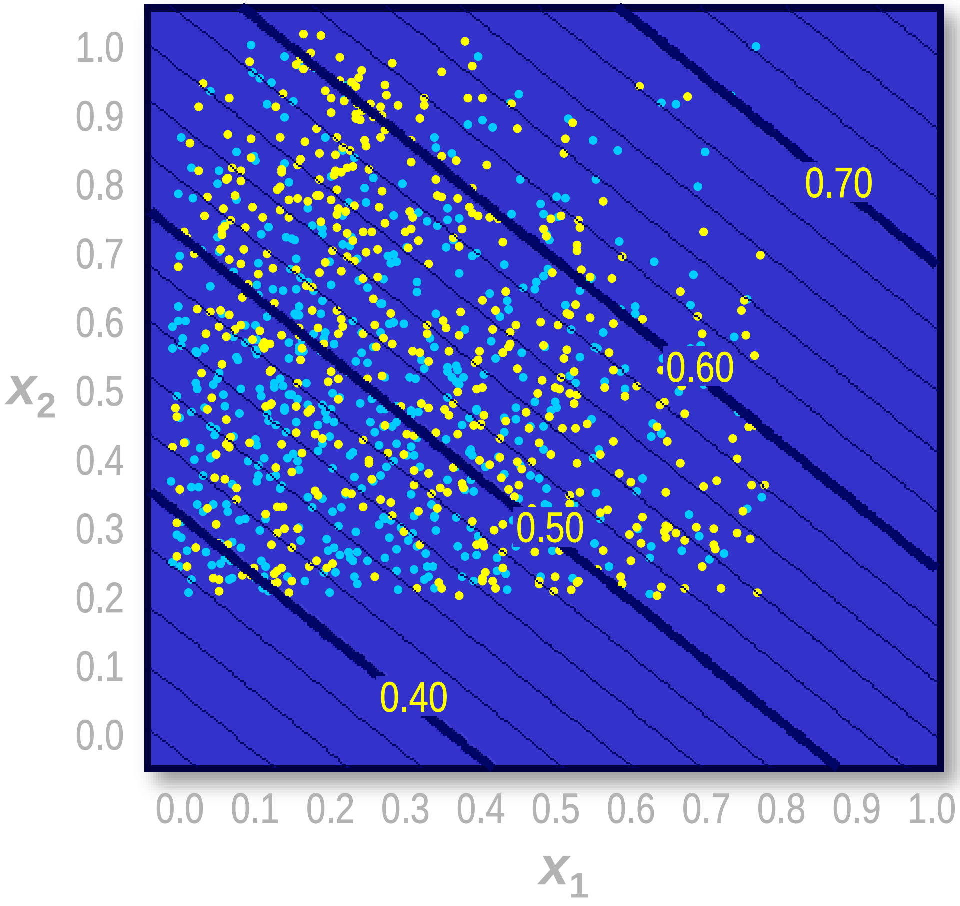
…

Coding Consolidation

Level	D_{ABCD}	D_B	D_C	D_D	D_{EF}	D_F	D_{GH}	D_H	D_I
A	1	0	0	0	0	0	0	0	0
B	1	1	0	0	0	0	0	0	0
C	1	0	1	0	0	0	0	0	0
D	1	0	0	1	0	0	0	0	0
E	0	0	0	0	1	0	0	0	0
F	0	0	0	0	1	1	0	0	0
G	0	0	0	0	0	0	1	0	0
H	0	0	0	0	0	0	1	1	0
I	0	0	0	0	0	0	0	0	1

Standard Logistic Regression

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

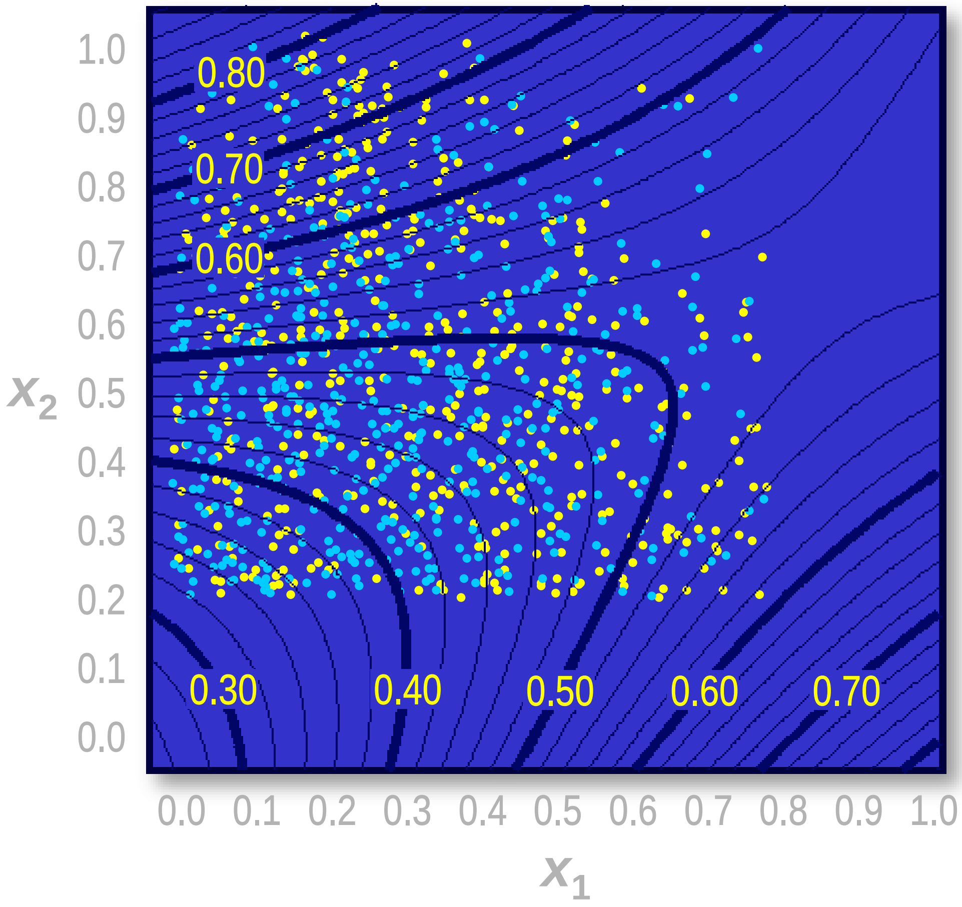


Polynomial Logistic Regression

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

quadratic terms $$+ \hat{w}_3 x_1^2 + \hat{w}_4 x_2^2$$

2-way interaction terms $$+ \hat{w}_5 x_1 x_2$$



...