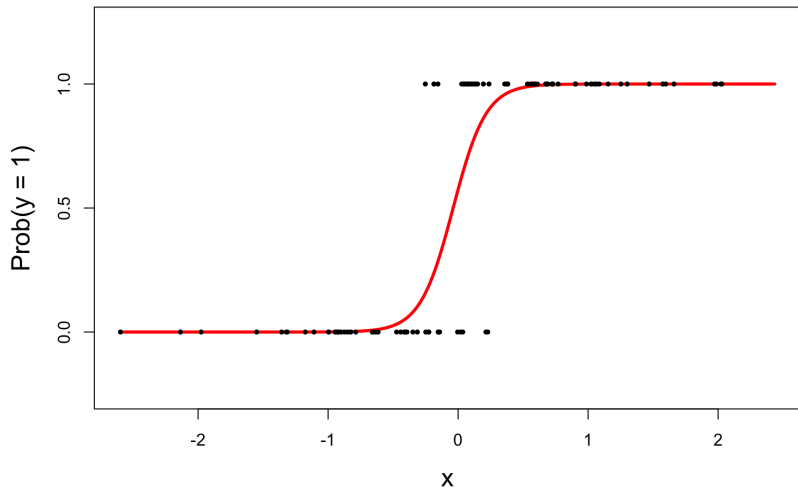


Welcome to STA 101!

Logistic regression: modeling a binary response



We can use this to do classification

Use data to train a *classifier*:

new case (x) \implies Classifier \implies Guess the type (\hat{y})

Examples:

Predictor (x)	Response (y)
Word counts in email	Spam or not?
Medical image	Cancer or not?
Loan application	High risk or not?

The logistic regression model

On the probability scale:

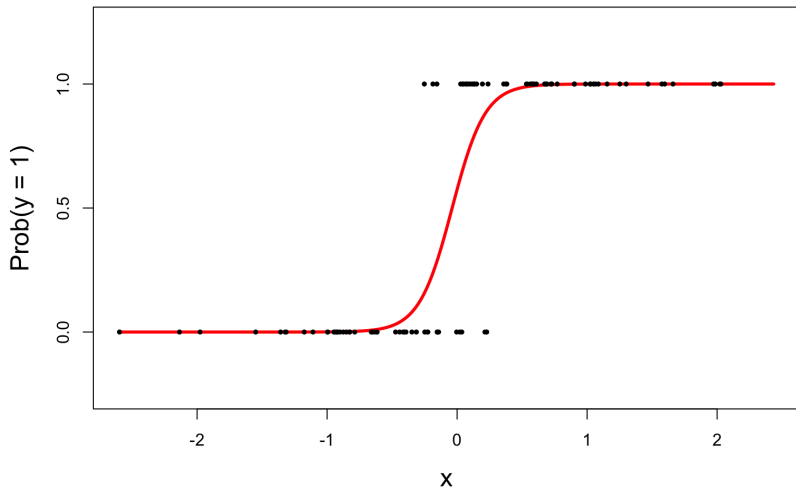
$$\text{Prob}(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m)}}.$$

On the log-odds scale, it's the linear model again:

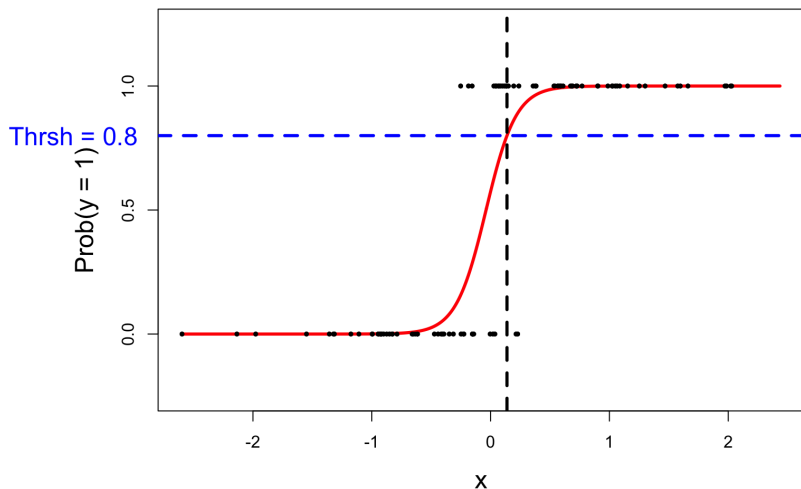
$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m.$$

- Unless you choose to use this for your project, you needn't fret about the details. But just know that...
- The β_j control the shape of the curve;
- We estimate them to get the “best fit” (maximum likelihood).

What can you do once you have this?

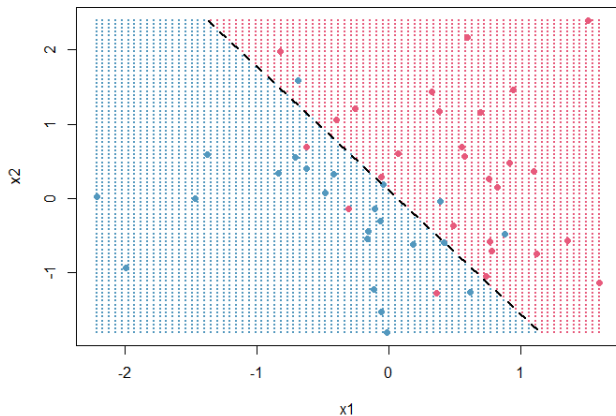


Decision boundaries for classification/prediction



- new x^* to the right of the black dotted line \implies predict $\hat{y} = 1$
- new x^* to the left of the black dotted line \implies predict $\hat{y} = 0$

Decision boundaries for classification/prediction



- new (x_1^*, x_2^*) above the line \implies predict $\hat{y} = 1$
- new (x_1^*, x_2^*) below the line \implies predict $\hat{y} = 0$

Spam filter example

`https://sta101-f24.github.io/computing/
computing-logistic.html`

Data analysis

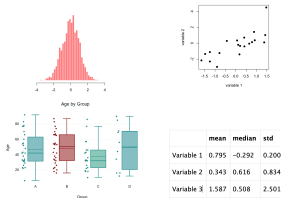
Data analysis

Transforming messy, incomplete, imperfect data into **knowledge**.

What form does that knowledge usually take?

- pictures;
- a concise set of numerical summaries.

subject	variable_1	variable_2
1	-1.65692830	-2.16524631
2	-0.90396488	-2.97993045
3	1.37141732	0.09720280
4	-0.43176527	0.27970110
5	0.40649190	0.69143221
6	1.47092198	4.47233461
7	-0.78625051	-1.24276055
8	0.64835135	-0.06749005
9	0.06363568	0.33517580



“Turn big box of numbers into pictures and small box of numbers.”

The *art* of data science

- How do you trick the human brain into appreciating complex multivariate relationships with flat, static pictures?
- What concise summaries should you look at?
- How should these things work together?

Statistical inference: October 17 and beyond

best-guess \pm margin-of-error.

After data analysis produces an answer to a question, how do you quantify the reliability of that answer? Two considerations:

- How “all over the place” are the data?
- How much of it do you have?

Punchline: meager, noisy datasets typically produce less reliable conclusions than large, less noisy ones.

Next up: how do you properly quantify this?

Food for thought

California Proposition 25 (2020)

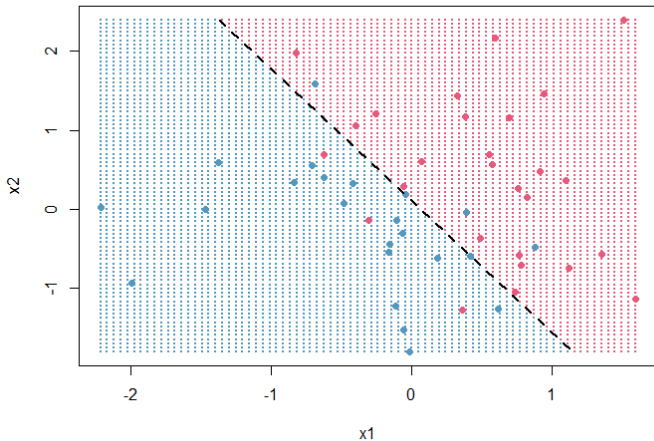
Popular referendum on 2018's Senate Bill 10:

- **YES:** replace cash bail with “risk assessment.”
 - Democratic Party, Governor Gavin Newsom, League of Women Voters of California, California Medical Association, Democracy for America (progressive PAC), etc.
- **NO:** keep the cash bail system.
 - Republican Party, American Bail Coalition, ACLU of Southern California, NAACP, California Asian Pacific Chamber of Commerce, etc.

If passed, each county would be empowered to develop a tool that predicts the risk of a suspect reoffending before trial.

Judges would consult this prediction to make bail decisions.

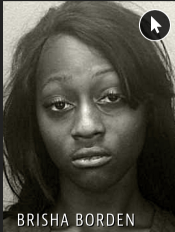

“Risk assessment” might just mean this stuff...



Above the line means high risk means no bail. Is this progress?

What happens when we try something like this?

Two Petty Theft Arrests



VERNON PRATER

LOW RISK **3**

BRISHA BORDEN

HIGH RISK **8**

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Petty Theft Arrests



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK **3**

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK **8**

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

California Proposition 25 did not pass

The cash bail system was retained:

Choice	Votes	Percent
YES	7,232,380	43.59%
NO	9,358,226	56.41%

Note:

- reasonable people can debate if this outcome is good or bad;
- every Californian was invited to decide whether statistics and data science should be deployed to make decisions with major social consequences. They opted out.
- This vote was held in the pre-ChatGPT era. What would the outcome be today? Is the case for YES stronger or weaker?

Under the hood, this might just be logistic regression too...

The Markup

The Secret Bias Hidden in Mortgage-Approval Algorithms

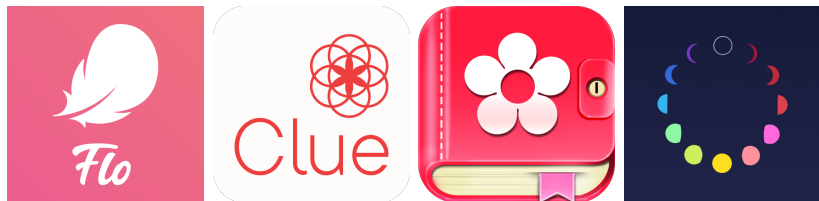
By Emmanuel Martinez and Lauren Kirchner

August 25, 2021 06:50 ET

Armies of PhDs are taking lucrative jobs at financial firms to develop models that assess the creditworthiness of loan applicants and automate the underwriting process. Few of them have spent much time thinking about the ethical implications of deploying statistical methods in the “real world.” It is seldom taught.

Decision-making under uncertainty: period trackers

Data + Model to predict timing of menstrual cycle.



A great example, perhaps, of data and modeling truly improving modern life, and a perfect microcosm of the themes of our course...

...until they sell your data to a third party without you realizing.

Data privacy

AKA The reason Tony Soprano ripped the GPS out of his Escalade:

The New York Times

Automakers Are Sharing Consumers' Driving Behavior With Insurance Companies

LexisNexis, which generates consumer risk profiles for the insurers, knew about every trip G.M. drivers had taken in their cars, including when they sped, braked too hard or accelerated rapidly.



By Kashmir Hill

Kashmir Hill has been writing about technology and privacy for more than a decade.

Published March 11, 2024 Updated March 13, 2024

But I don't want this to be an avalanche of pure negativity...

Faster, more accurate cancer screening

Augmenting doctors' diagnostic capacity so that they make fewer mistakes, treat more people, and focus on other aspects of care:

AAMCNEWS

Is it cancer? Artificial intelligence helps doctors get a clearer picture

AI tools are analyzing images and tissue samples to detect cancer sooner and more precisely. Doctors hope to improve and accelerate patient care.

By Patrick Boyle, Senior Staff Writer

March 28, 2024

October 9, 2024

The Nobel Prize in Chemistry 2024

David Baker

“for computational protein design”



Demis Hassabis

“for protein structure prediction”



John Jumper

“for protein structure prediction”



- AlphaFold2: “predicting 3D structures [of proteins] (y) directly from the primary amino acid sequence (x).”
- “researchers can now better understand antibiotic resistance and create images of enzymes that can decompose plastic.”

The question

How do we get more of the good and less of the bad?

I don't know. But these things couldn't hurt:

- integrating a serious discussion of right and wrong into all levels of the statistics curriculum, from AP Statistics to PhDs at Stanford;
- recognizing that all sorts of careers have a seat at the table besides just eggheads who know how the math goes: lawyers, nurses, school teachers, doctors, social workers, philosophers, ...

There are tremendous opportunities here for all sorts of folks to make contributions toward the good. I hope you do!