Smallest font

Please turn off and put away your cell phone

Calibration slide

Smallest font

Introduction to Data Science

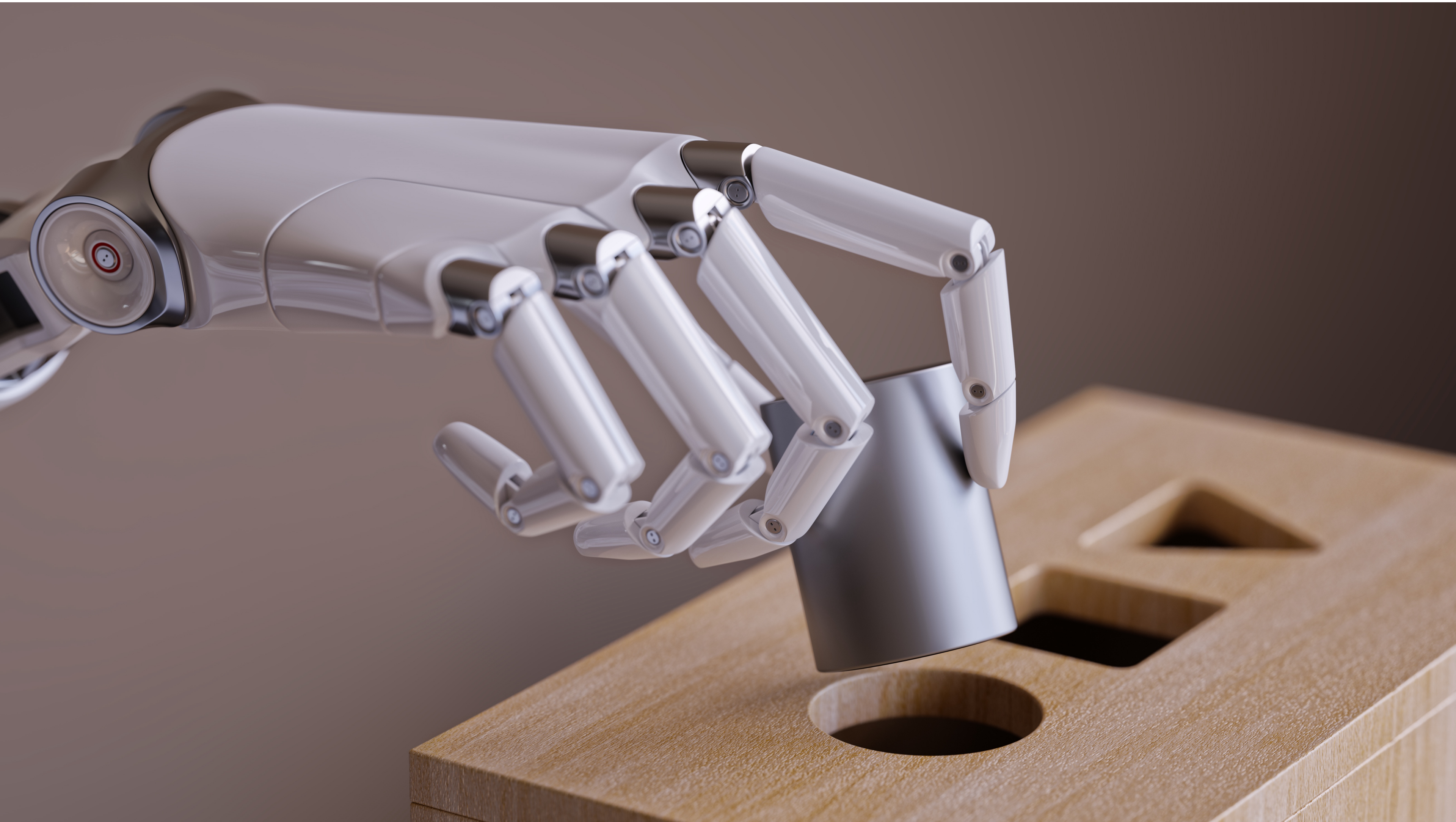
Looking back at growing up

- BF_{10} = How many times more likely is the observed data assuming the alternative hypothesis is true compared to assuming that the null hypothesis is true.
- Maximum likelihood: Allows to identify the hypothesis with the largest Bayes Factor.
- PPV: Posterior probability that the effect is real, if the result of the test is significant.
- Credible interval: Bayesian version of a confidence interval, takes the prior distribution into account.
- High power: Yields higher PPV and makes prior less relevant to determine the posterior.
- When to use Bayesian statistics? If we have good priors or insufficient data.

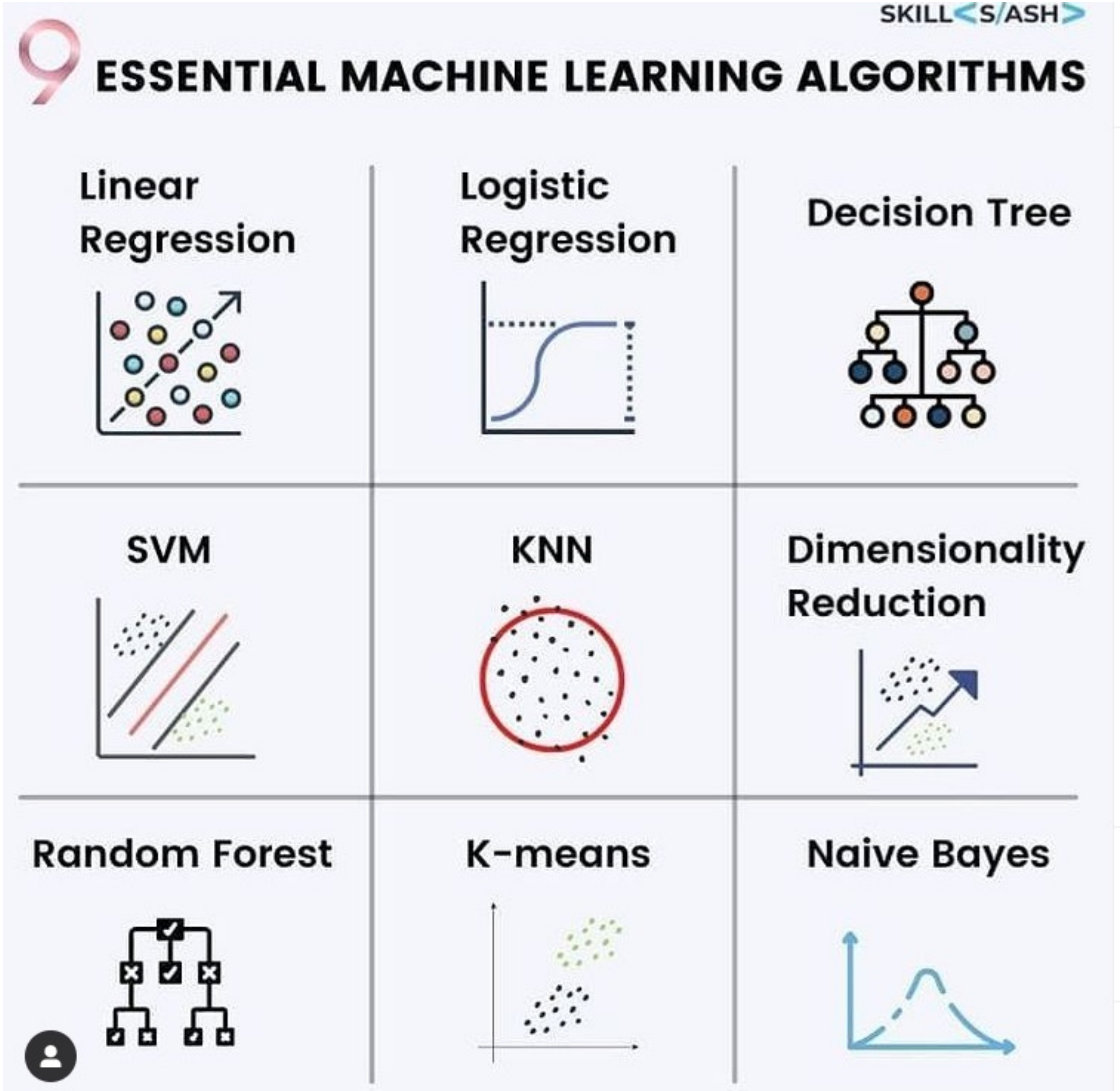
VI

“Machine learning”

Machine Learning



As seen online (in no particular order)

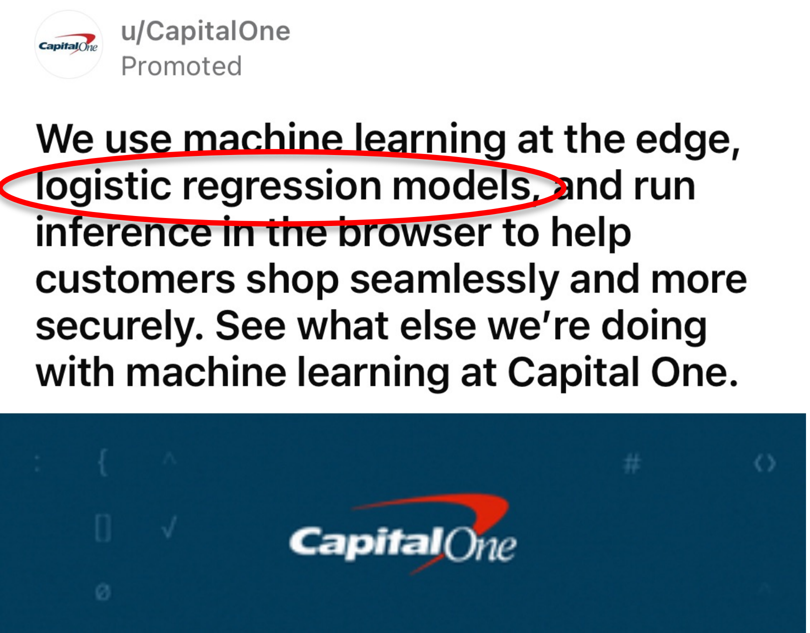


Logistic regression

…4,5,3,2,9,4,7,…

0

1



Logistic regression: When you just need to predict a **binary** outcome and the link between predictor and outcome might be **nonlinear**

This use case is surprisingly common:

Win/loss outcomes (games, elections)

Up / down votes (like or not)

Buy / sell decisions

Getting a college degree - or not

Predicting marriage

Predicting divorce

Epidemiology: Who will get sick?

Medicine: Life vs. death outcomes

Psychology: Fight vs. flight

Housing: Predicting mortgage default

Law: Guilty vs. not guilty verdicts

Getting into grad school – or not

How do you do that?

The toy example: Grad school admissions

- You apply to graduate school.
- What are possible outcomes?
- You get in
- or
- You don’t get in
- What are the odds that one of these outcome will happen?
- 50%?
- Why or why not?

Predictors matter

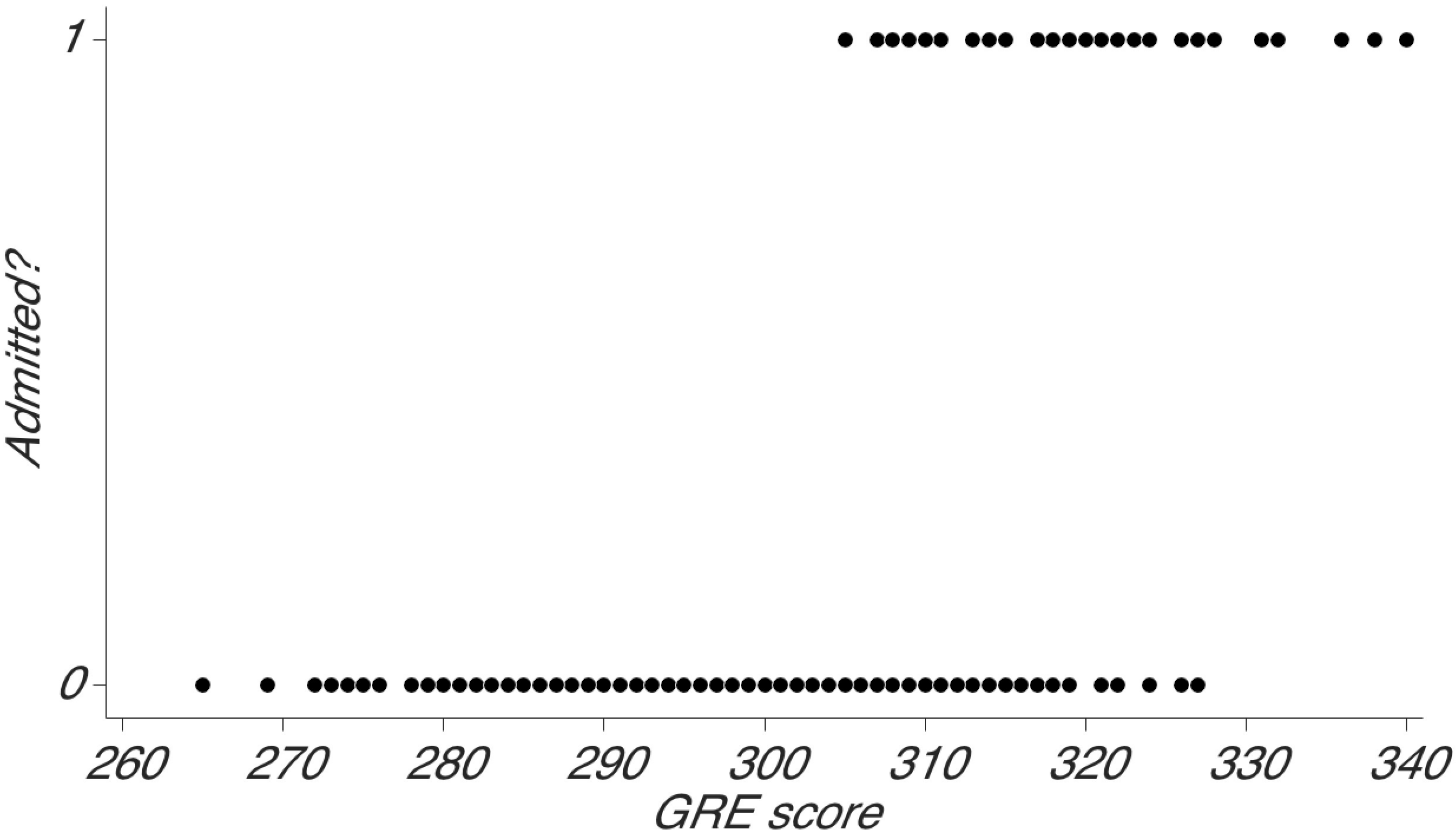
- As a college senior, you are busy applying to graduate schools.
- As part of this process, you take the GRE.
- GRE test scores can range from 260 to 340 (both quantitative and verbal scores range from 130 to 170).
- Graduate schools factor in these scores when deciding to admit or not admit a candidate.
- They also factor in other things, e.g. college GPA, letters of recommendation, etc.
- Your score is 307
- Now what?

You wonder if you should even bother applying

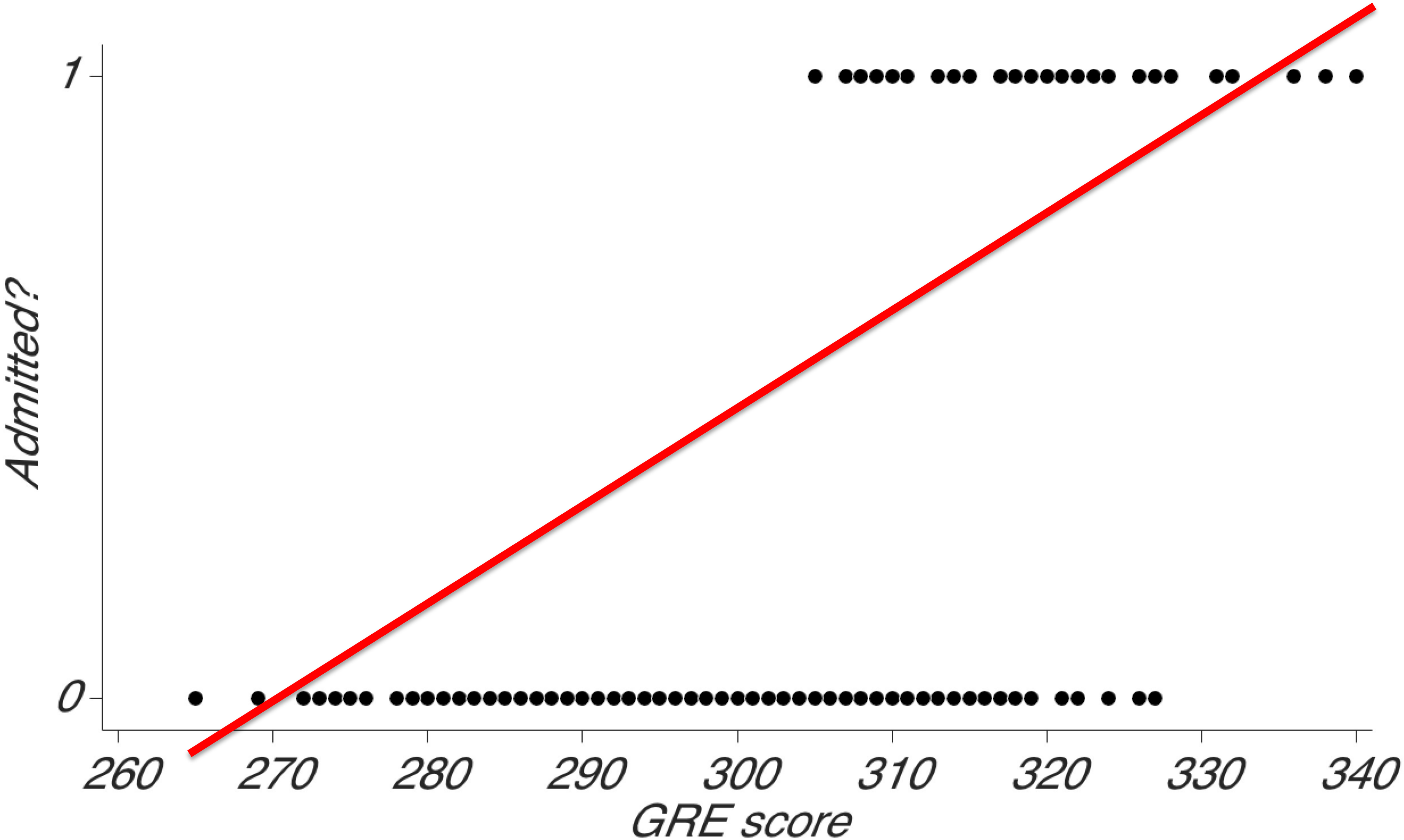
- You email the school and ask what your **odds** are of getting in with that score.
- They email you back saying that they don't know for sure, but that they are happy to send you the raw data from last years admissions process.
- The dataset contains information about GRE scores of 500 applicants and whether they were admitted or not.

Applicant	1	2	3	4	5	…	496	497	498	499	500
GRE score	304	279	338	296	299	…	312	290	319	300	293
Admitted	0	0	1	0	0	…	1	0	0	0	0

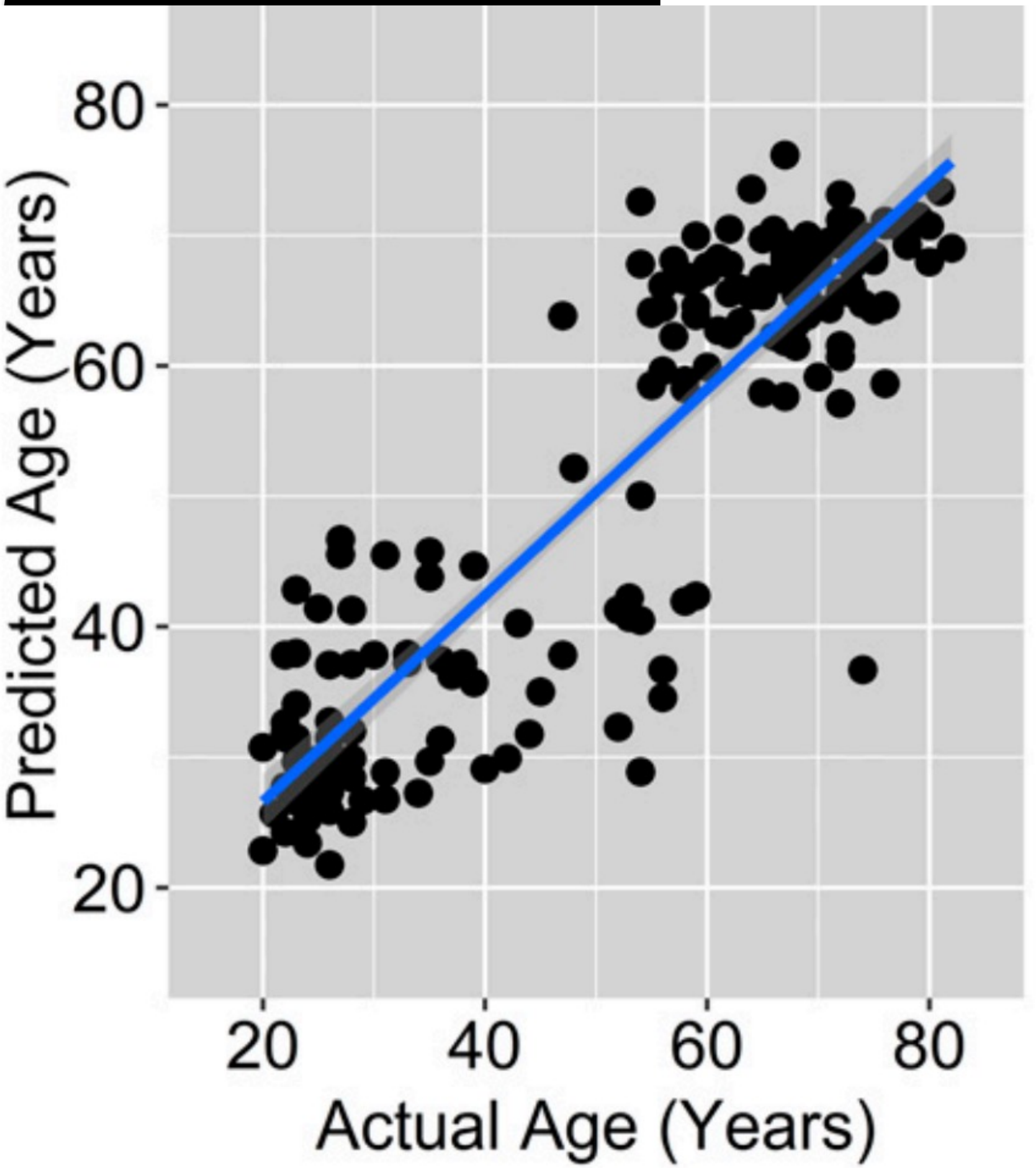
Now what?



How good is the best fit line?



Aside: Just because you can fit a linear model doesn't mean you should



Goyal et al. (2018). Persistent metabolic youth in the aging female brain. PNAS

Sometimes, linear regression makes no sense

- One can still determine a “best fit” line (one always can), but it makes no sense here.
- How would you interpret it?
- Linear regression assumes that data are distributed normally.
- Here, the outcomes are all either 0 and 1 (binary or dichotomous). Definitely not normal.
- Also: Effect of unit change is assumed to be constant.
- Also: Linear regression is unbounded.
- So linear regression gives nonsensical results here.
- We need to do better/something else:
- Logistic regression

Logistic regression

- Provides a **nonlinear** model that links the predictor and the outcomes.
- Gives **odds** that an outcome happens, vs. it not happening, for a given predictor value.
- **Odds**: Probability of an event occurring divided by the probability of the event not occurring.

$$Odds = \frac{p}{\neg p} = \frac{p}{1-p}$$

p = 0.5 Odds = 1:1

p = 0.75 Odds = 3:1

p = 0.33 Odds = 1:2

$p = 0.25$. What are the odds?

A. 1:1

B. 1:2

C. 1:3

D. 3:1

Logs

+ is to ***** as ***** is to **^**

***** Is to **/** as **^** is to **log**

Counting how often to add	**Counting how often to multiply**
$10 + 10 + 10 = 30$	$10 * 10 * 10 = 1000$
$10 * 3 = 30$	$10 \wedge 3 = 1000$
$30 / 10 = 3$	$\log_{10}(1000) = 3$
$2 + 2 + 2 = 6$	$2 * 2 * 2 = 8$
$2 * 3 = 6$	$2 \wedge 3 = 8$
$6 / 2 = 3$	$\log_{2}(8) = 3$

The *logit* function

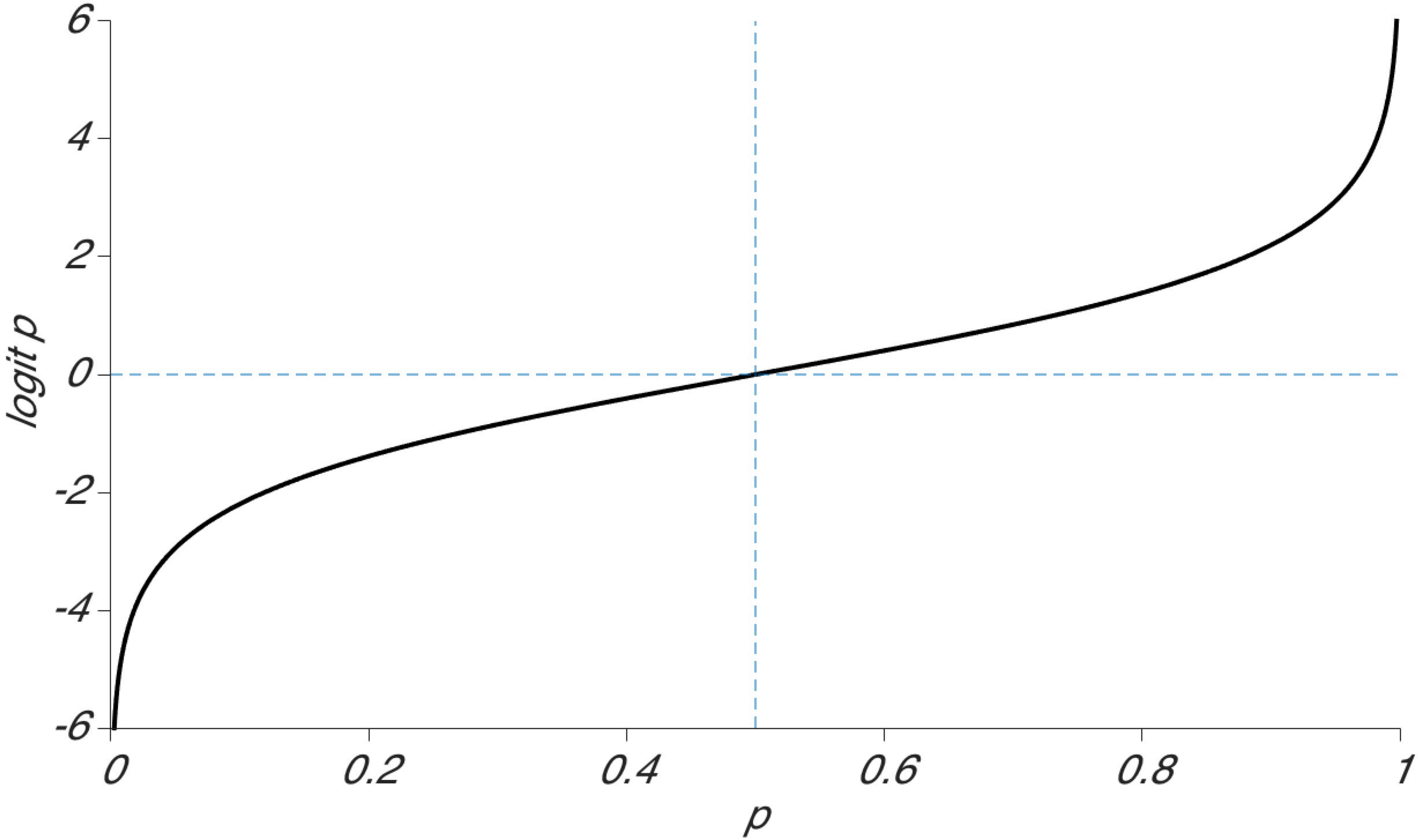
- Links the values in the predictor variable to the probabilities of the outcomes.
- We don't know what those are. Have to estimate them from the data.
- **Logit**: Natural log of the odds
- Natural log: logarithm with e as base: $\log_e x = \ln x$
- ln stands for "logarithmus naturalis"

x = Data
Here: GRE score

$$\operatorname{logit}(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

- Solves the boundary problem - probability is bounded, odds are not – go to infinity. Log odds to negative inf.
- All of this is a bit… abstract.
- What does the logit function look like?

Like that – a plot of log odds



We’re close, but not quite there

- Logit looks interesting.
- But it is kind of the opposite of what we want.
- Probability was on the x-axis. We want to know what the probabilities are (on the y-axis).
- To get that, we need to take the inverse of the logit function. (turning the graph on its side)

$$\log it^{-1}(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$$

Derivation of logit inverse by solving for p

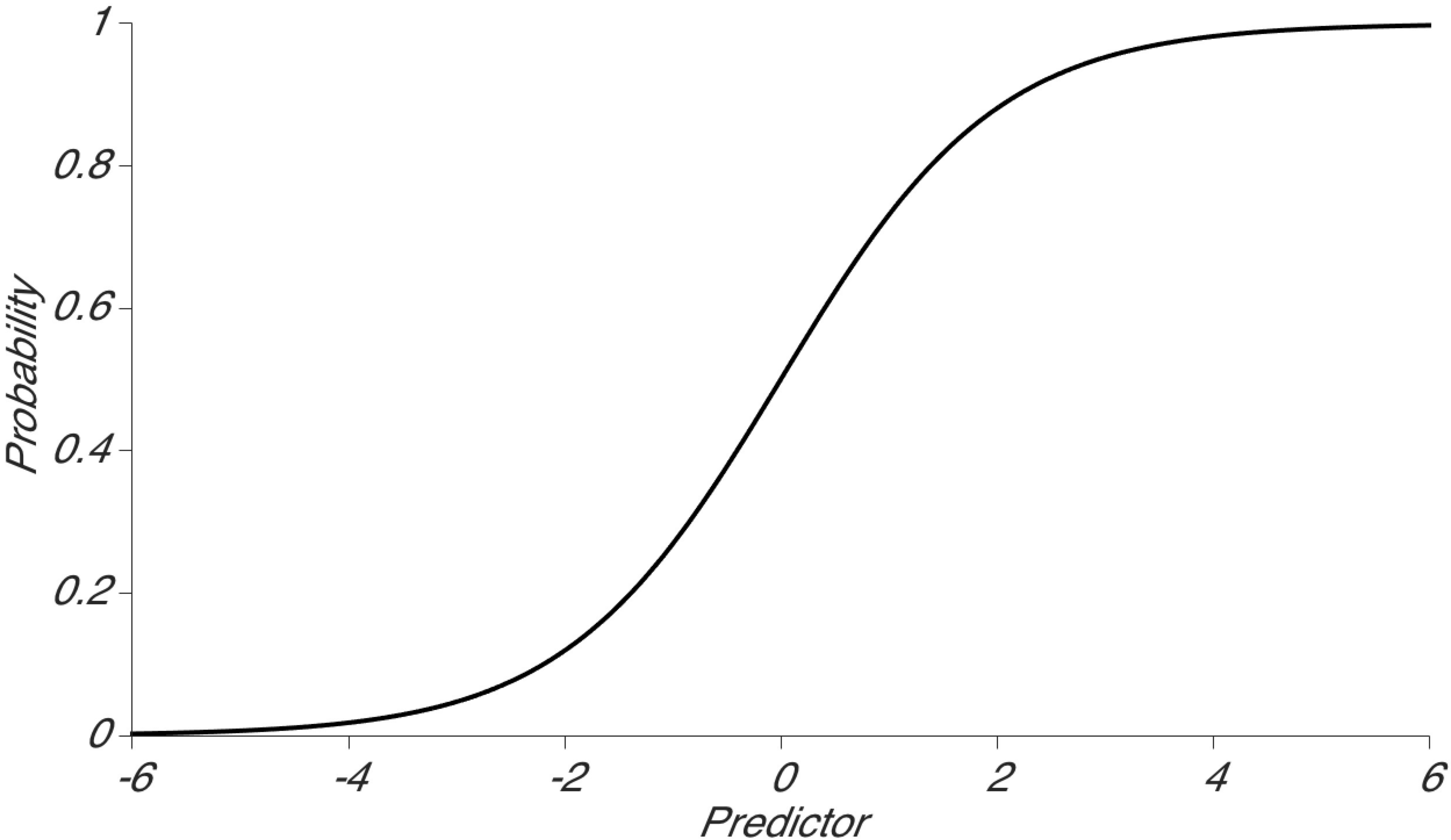
$$\log it(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1} \qquad p = e^{\beta_0 + \beta_1 x_1}\left(1-p\right)$$

$$p = e^{\beta_0 + \beta_1 x_1} - e^{\beta_0 + \beta_1 x_1} * p \qquad p + e^{\beta_0 + \beta_1 x_1} * p = e^{\beta_0 + \beta_1 x_1}$$

$$p(1 + e^{\beta_0 + \beta_1 x_1}) = e^{\beta_0 + \beta_1 x_1} \qquad p = \frac{e^{\beta_0 + \beta_1 x_1}}{(1 + e^{\beta_0 + \beta_1 x_1})}$$

So the inverse logit looks like this



The logistic function

- The logistic function is an inverse logit function.
- It looks S-shaped.
- It’s called a “sigmoidal function”.
- This is the function that underlies logistic regression.
- We fit it to the combination of predictors and outcomes.
- Using “Maximum Likelihood Estimation” (MLE)
- We have already seen MLE when fitting distributions to data when talking about Bayes.

Another intuition on maximum likelihood

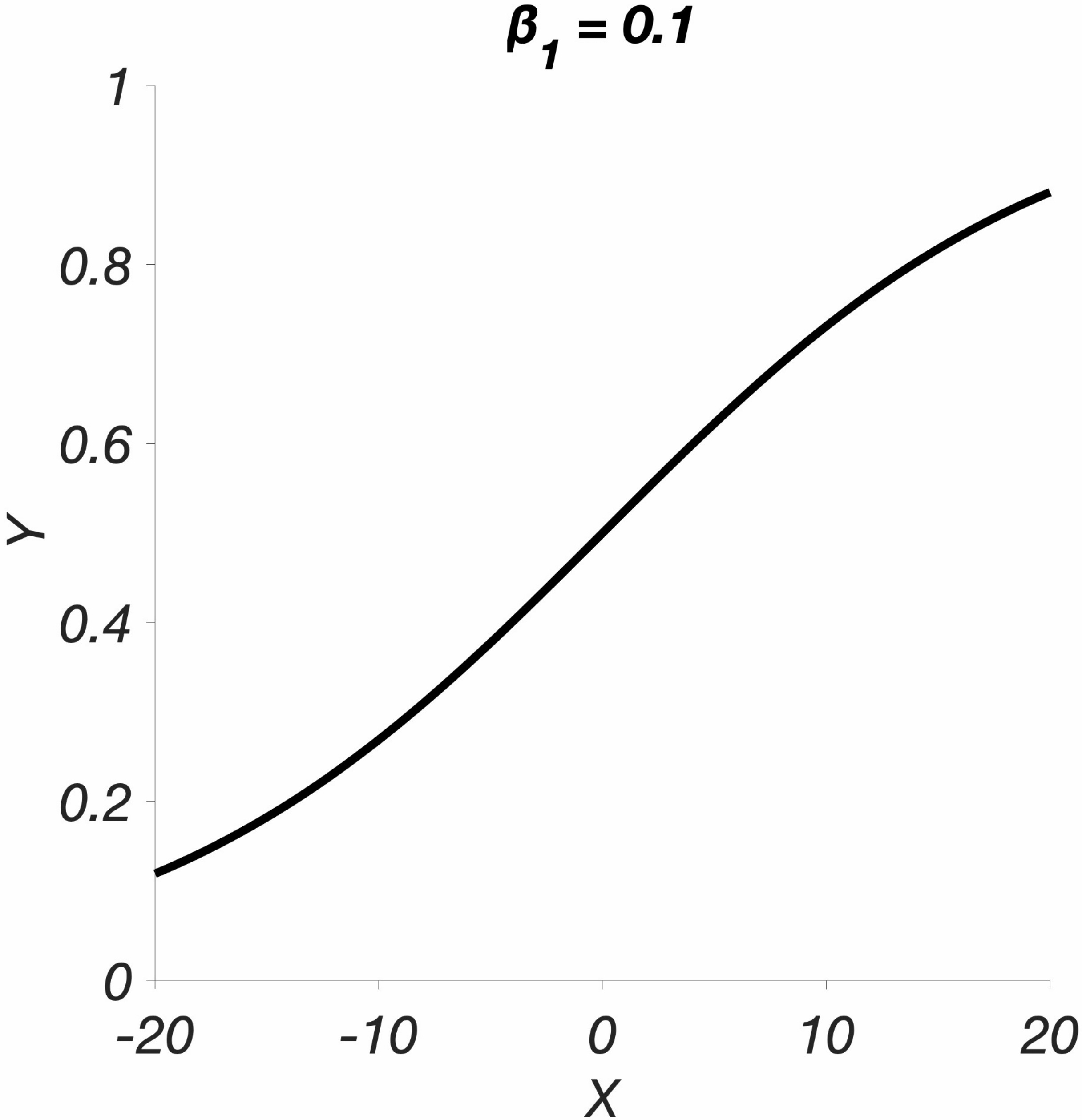
Try $\beta_1 = 0.3$ $\quad\Sigma(|\text{deviations}|) = 1.85$

Applicant	1	2	3	4	5	…	496	497	498	499	500
GRE score	304	279	338	296	299	…	312	290	319	300	293
Admitted	0	0	1	0	0	…	1	0	0	0	0
p (admitted)	0.3	0.05	0.9	0.2	0.22	…	0.75	0.15	0.8	0.23	0.17
Deviation	0.3	0.05	0.1	0.2	0.22	…	0.25	0.15	0.8	0.23	0.17

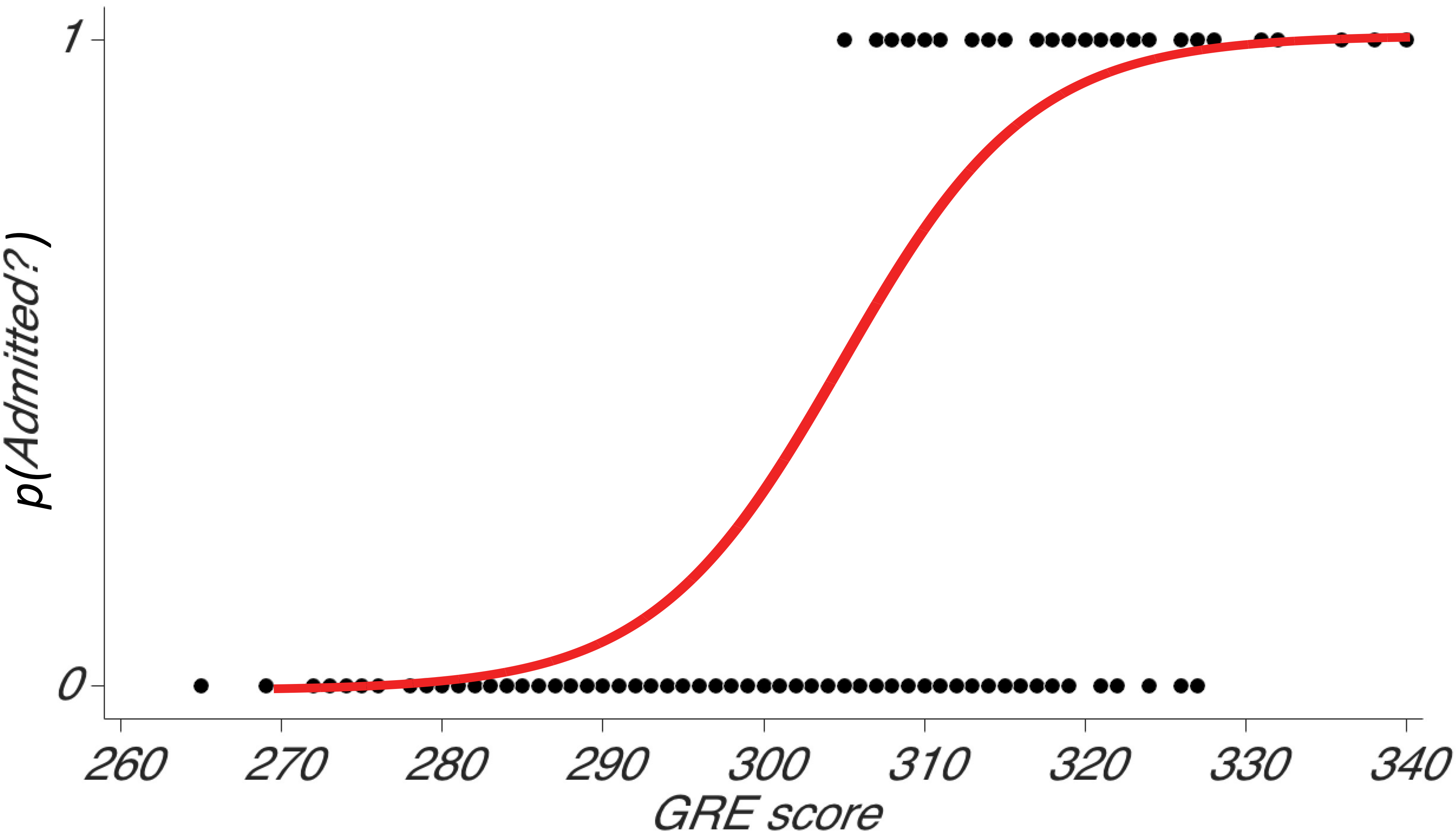
Try $\beta_1 = 0.4$ $\quad\Sigma(|\text{deviations}|) = 2.82$

Applicant	1	2	3	4	5	…	496	497	498	499	500
GRE score	304	279	338	296	299	…	312	290	319	300	293
Admitted	0	0	1	0	0	…	1	0	0	0	0
p (admitted)	0.4	0.1	0.95	0.25	0.27	…	0.8	0.2	0.85	0.28	0.22
Deviation	0.4	0.1	0.05	0.25	0.27	…	0.2	0.2	0.85	0.28	0.22

The impact of varying β_1



Revisiting our scatter plot

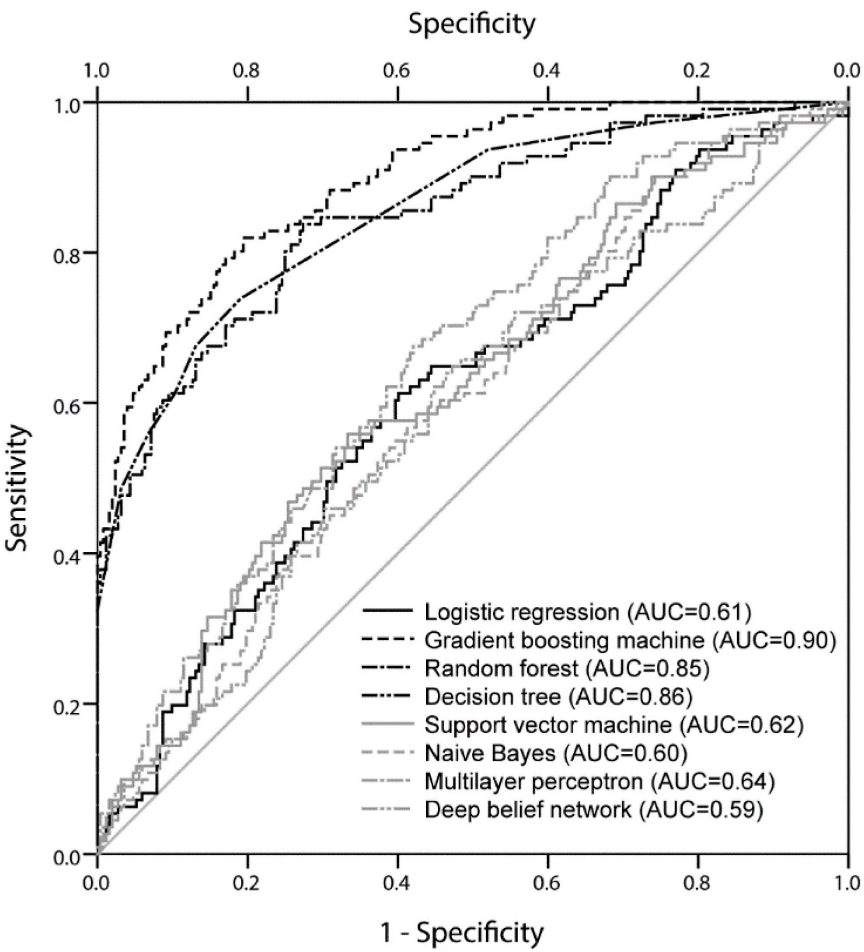


Applying logistic regression

- Logistic regression solves for the beta-weights.
- With the beta weights and the predictor values, we can estimate the probability that we will get admitted.
- Your GRE score was 307. Doing the calculations show that p (getting in) = 0.15.
- More prep, then re-taking: Can expect a 10 point gain on the GRE.
- Plugging the new expected value – 317 – into the equation shows that the new probability of getting in is now 0.7.

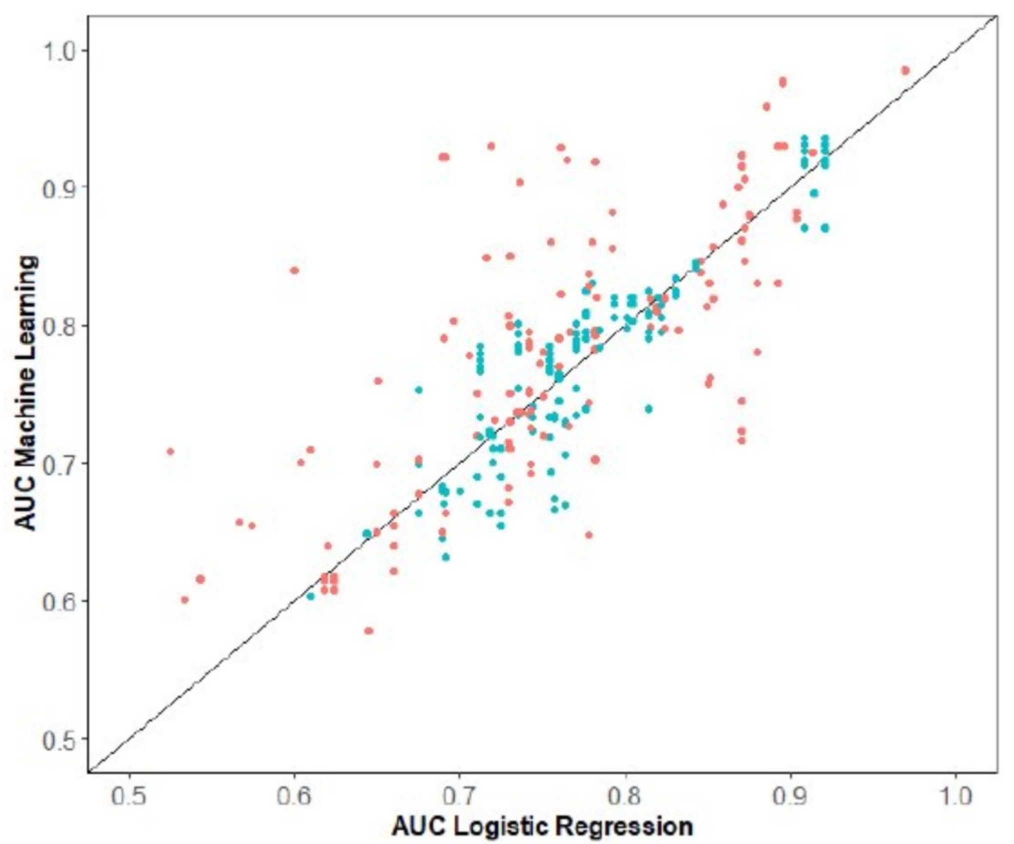
Logistic regression is straightforward, yet powerful so it is commonly used

"When we raise money it's **AI**, when we hire it's **machine learning**, and when we do the work it's **logistic regression.**"



Lee et al., 2018

Figure A.2. Scatter plot of the area under the ROC curve (AUC) for LR vs ML for all 282 comparisons. Comparisons with low risk of bias are shown in green, comparisons with high risk of bias in red.



Van Calster et al., 2019

Logistic Regression – the sweet spot?

Method/Feature	Linear Regression	Logistic Regression	Advanced machine learning methods
Assumes normal data distribution	Yes	No	No
Assumes equal impact of unit change	Yes	No	No
Bounded predictions?	No	Yes	Yes
Readily interpretable?	Yes	Yes	No
Easy to build?	Yes	Yes	Sometimes
Risk of overfitting	Low	Low	High