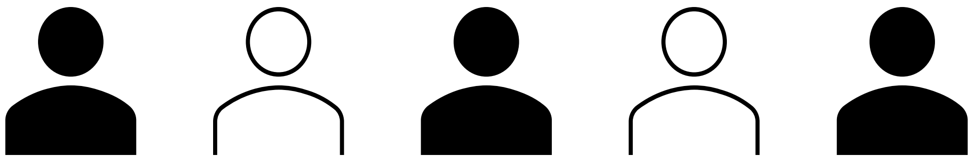
Classifiers

Conditional Mean as a Classifier

Our friend the conditional mean

- What we did
  - What groups allow us to predict our *continuous outcomes* with the **most accurate point estimates** ?
- What we are now doing
  - What groups allow us to predict our *categorical outcomes* with the **greatest probability**?

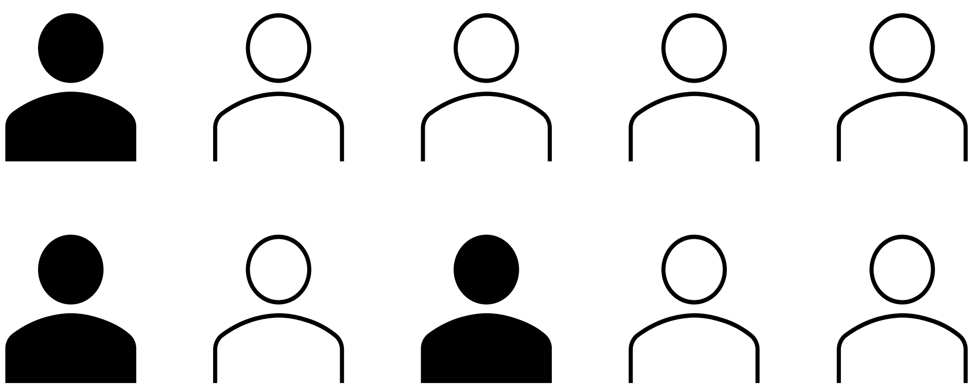
Store 1



3 out of 5

3/5 = .60 or 60%

Store 2



3 out of 10

3/10 = .30 or 30%

Outcomes

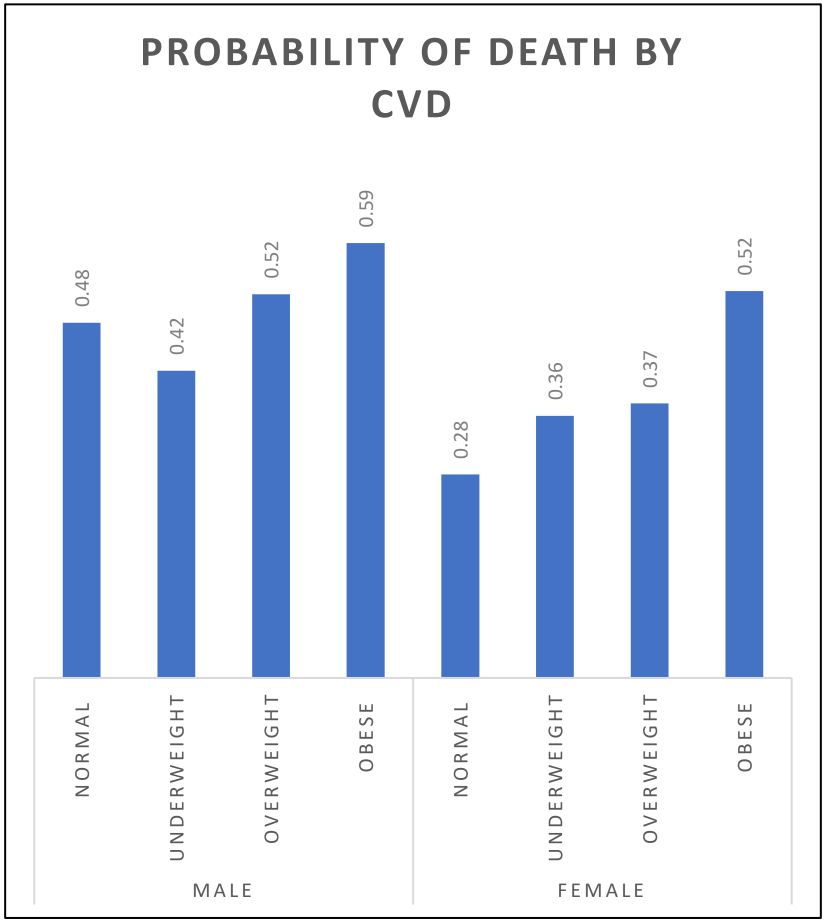
Predictors

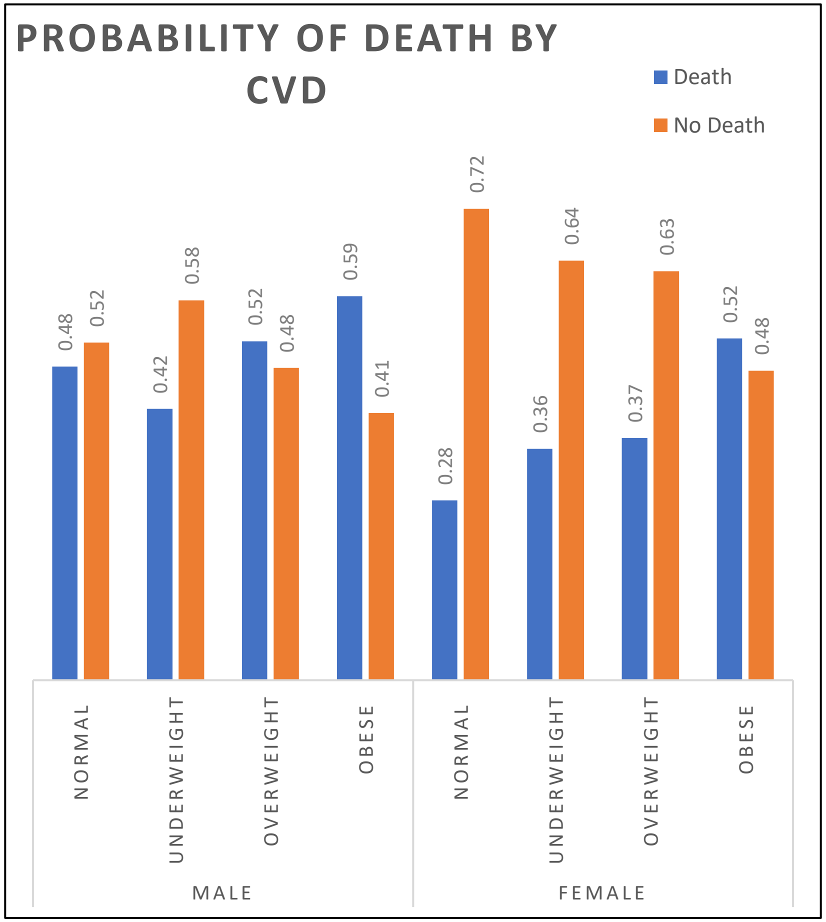
	Death
Male	**0.51**
Normal	0.48
Underweight	0.42
Overweight	0.52
Obese	0.59
Female	**0.34**
Normal	0.28
Underweight	0.36
Overweight	0.37
Obese	0.52
Grand Total	**0.42**

Outcomes

Predictors

	Death b	
	Death	**No Death**
Male	**0.51**	**0.49**
Normal	0.48	0.52
Underweight	0.42	0.58
Overweight	0.52	0.48
Obese	0.59	0.41
Female	**0.34**	**0.66**
Normal	0.28	0.72
Underweight	0.36	0.64
Overweight	0.37	0.63
Obese	0.52	0.48
Grand Total	**0.42**	**0.58**


PROBABILITY OF DEATH BY CVD
0.48
0.42
0.52
0.59
0.28
0.36
0.37
0.52
NORMAL
UNDERWEIGHT
OVERWEIGHT
OBESE
NORMAL
UNDERWEIGHT
OVERWEIGHT
OBESE
MALE
FEMALE


PROBABILITY OF DEATH BY CVD
Death
No Death
0.48
0.52
0.42
0.58
0.52
0.48
0.59
0.41
0.28
0.72
0.36
0.64
0.37
0.63
0.52
0.48
NORMAL
UNDERWEIGHT
OVERWEIGHT
OBESE
NORMAL
UNDERWEIGHT
OVERWEIGHT
OBESE
MALE
FEMALE

Discretizing (Binning)

- Equal-frequency binning
- Equal-width binning
- K-means clustering

Discretizing (Binning)

- Equal-frequency binning
  - n-tiles
    - Medians, quartiles, quintiles, deciles, etc.
  - Equal representation across range
  - Parallels the original distribution
    - Good for model input
- Equal-width binning
- K-means clustering

Discretizing (Binning)

- Equal-frequency binning
- Equal-width binning
  - Each bin is the same size of the range (width)
    - Age, GPA, etc.
  - Convenient for interpretation
  - Must take care when determining the width
- K-means clustering

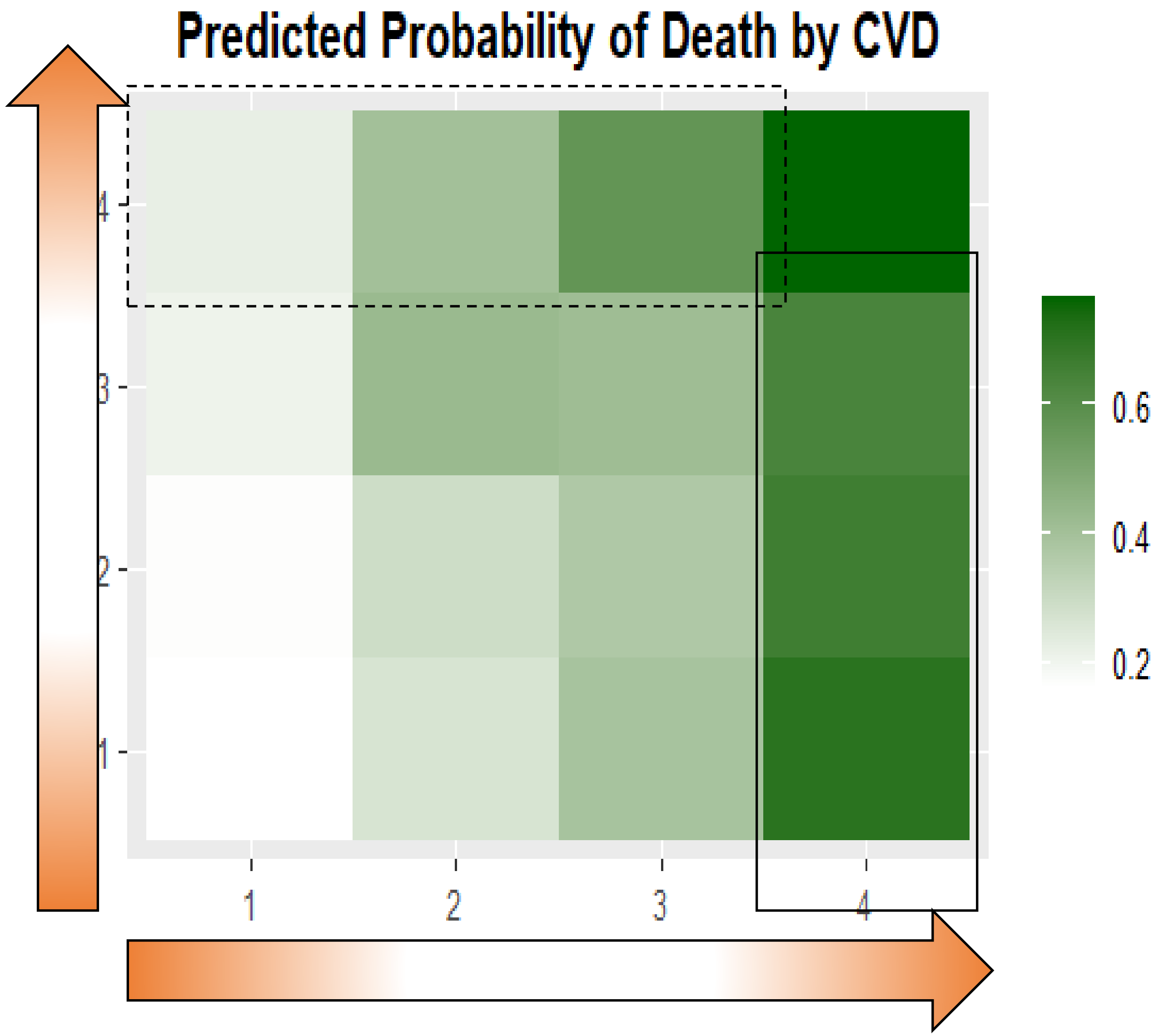
Discretizing (Binning)

- Equal-frequency binning
- Equal-width binning
- K-means clustering
  - Each bin is determined Maximum Likelihood Optimization
    - Cases belong to the “closest mean”
  - Can identify useful profiles/typologies
  - Category labels must be interpreted *post hoc* and can be multidimensional

A Neat Trick

- Outcome (binary) =
  Predictor1 (discretized) + Predictor2 (discretized)

- Heatmap
  - Plot the conditional probability of outcome
    - X-axis: Predictor1
    - Y-axis: Predictor 2
    - Color: Probability



Predicted Probability of Death by CVD
4
3
2
1
1
2
3
4
0.6
0.4
0.2

What did we cover?

- Conditional Mean as a Classifier
  - Probability scores ← Discrete Predictors
- Discretizing Continuous Variables
  - Equal-frequency binning
  - Equal-width binning
  - K-means clustering

- *Next up:*
  - Assessing the conditional mean as a classifier
    - Does the model work well as a Classifier

Classifiers

Evaluating Classifiers: Sensitivity and Specificity

	Reality										**Acc**	**Sen**	**Spe**
	R	***B***	***R***	***B***	***R***	***B***	***R***	***B***	***R***	***B***	–	–	–
Model 1	**R**	R	**R**	R	**R**	R	**R**	R	**R**	R	0.50	1.00	0.00
Model 2	B	**B**	B	**B**	B	**B**	B	**B**	B	**B**	0.50	0.00	1.00
Model 3	**R**	**B**	B	R	**R**	R	B	R	**R**	**B**	0.50	0.60	0.40

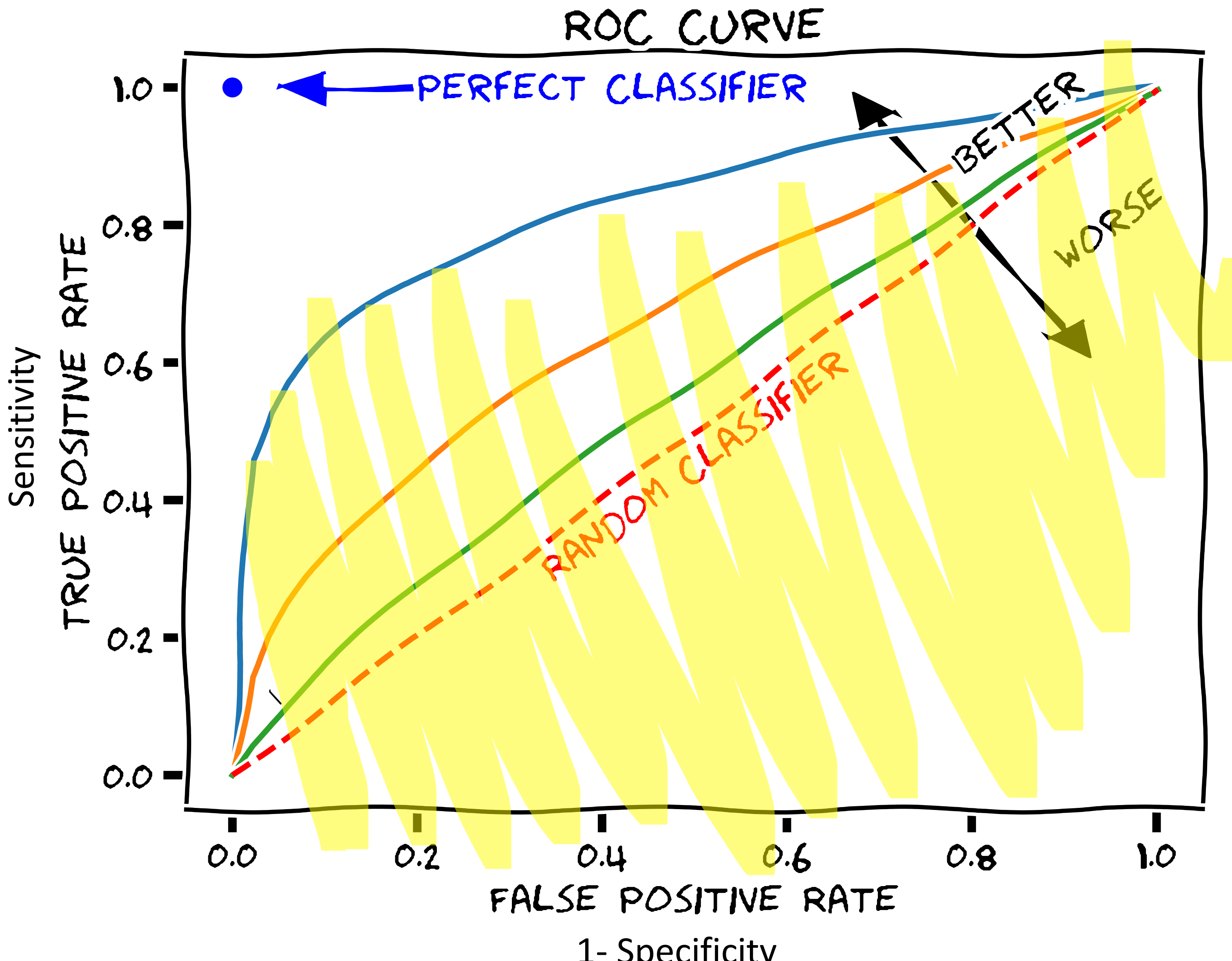
Sensitivity or Specificity?

- Depends…

- Costs of False Positive
  - Squander resources

- Costs of False Negative
  - Miss opportunities

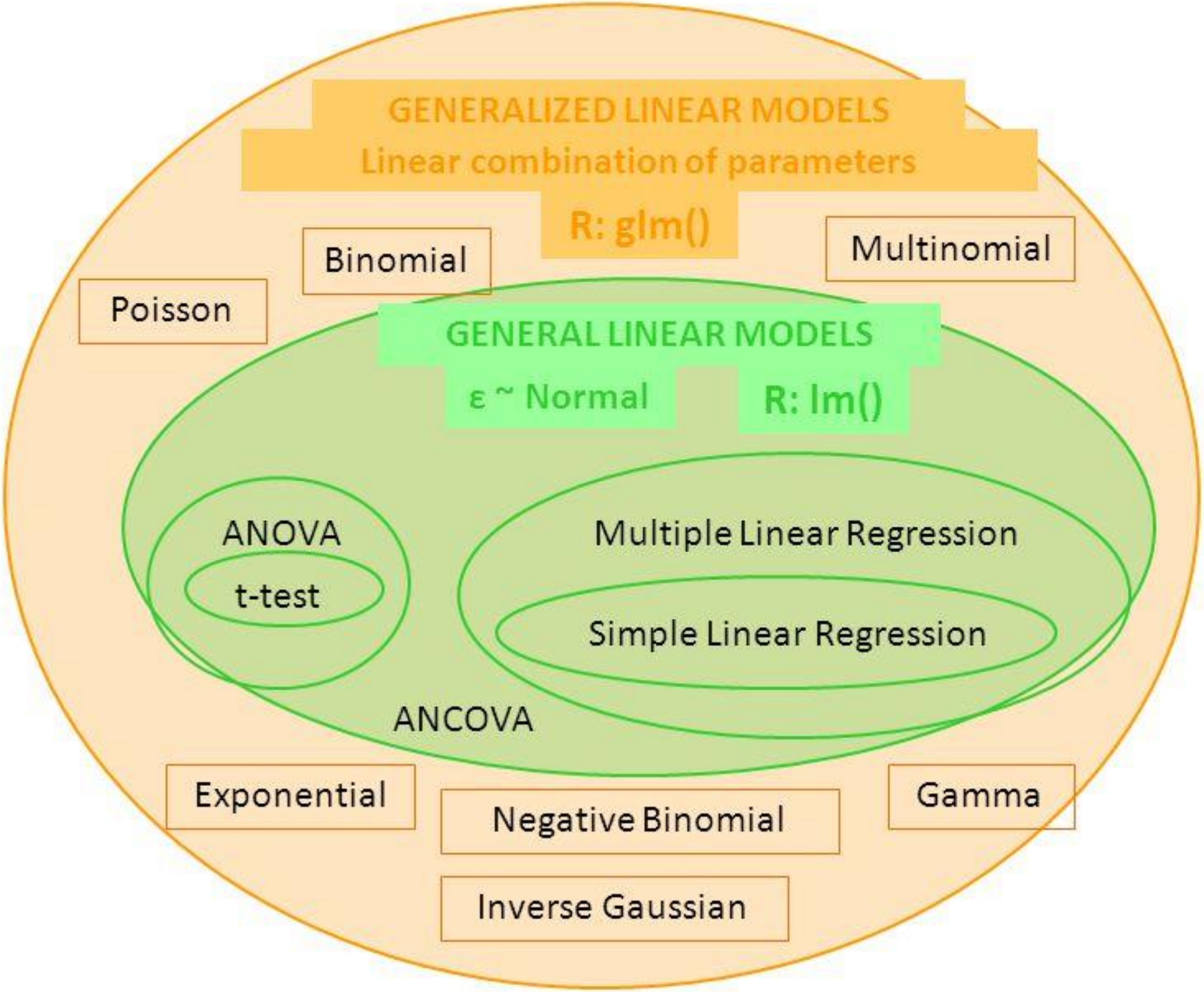
- Trade-offs
  - *Numbersense* Chapter


ROC CURVE
PERFECT CLASSIFIER
BETTER
WORSE
RANDOM CLASSIFIER
TRUE POSITIVE RATE
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
FALSE POSITIVE RATE
1- Specificity

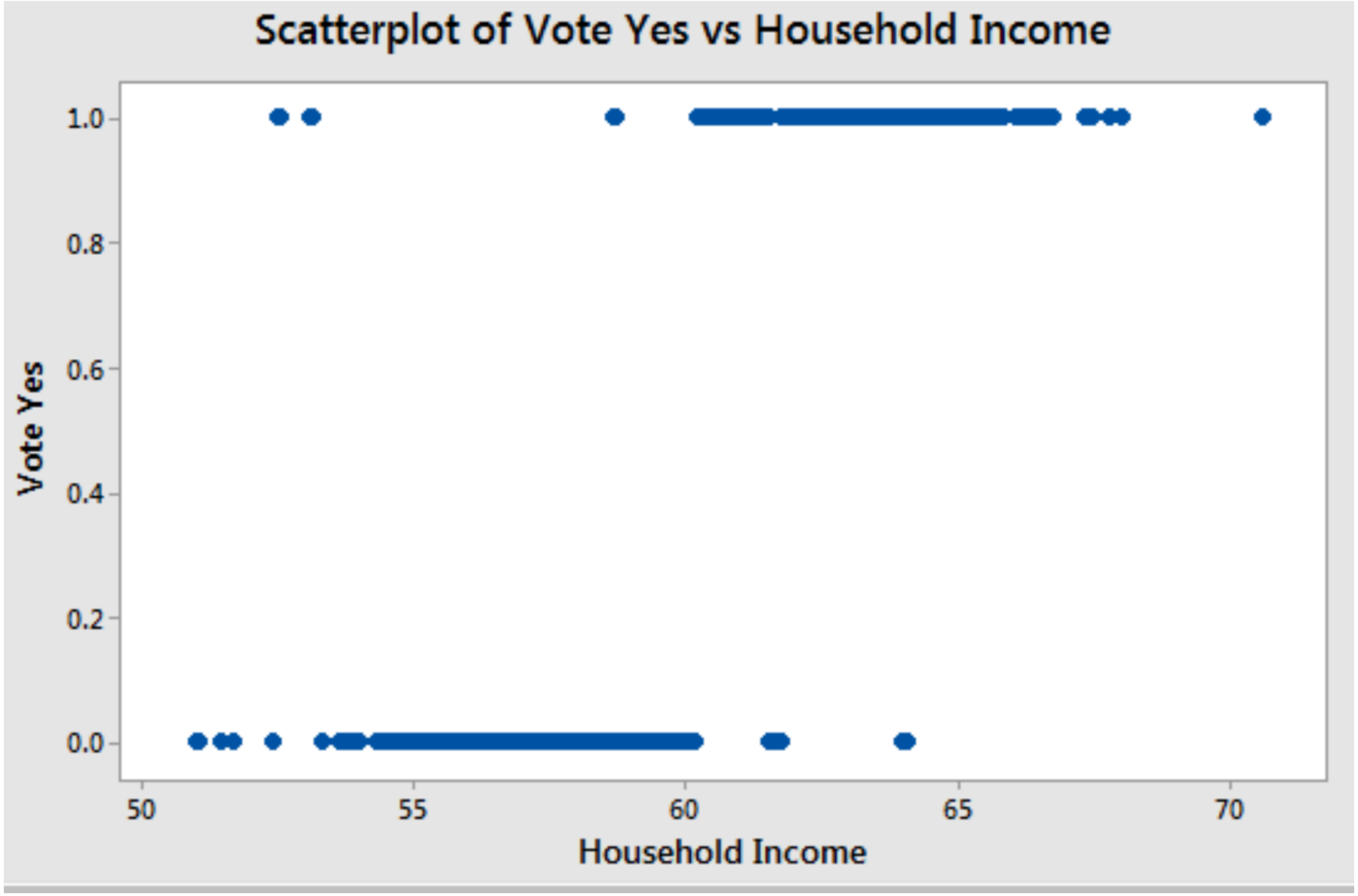
Classifiers

Logistic Regression as a Classifier

What is Logistic Regression?



https://images.slideplayer.com/15/4659936/slides/slide_3.jpg



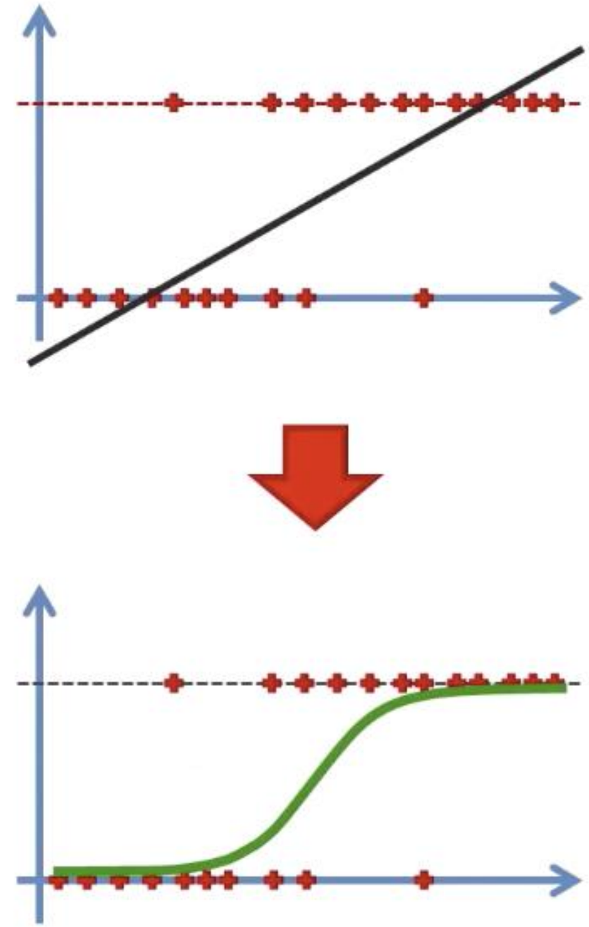
https://online.stat.psu.edu/stat800/book/export/html/711

$$y = b_0 + b_1 * x$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$



https://www.vebuso.com/2020/02/linear-to-logistic-regression-explained-step-by-step/

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.47823    0.21981 -24.923  < 2e-16 ***
AGE          0.09858    0.00428  23.030  < 2e-16 ***
bmicat_X1    0.16434    0.30377   0.541 0.588508
bmicat_X2    0.27592    0.07372   3.743 0.000182 ***
bmicat_X3    0.68385    0.10759   6.356 2.07e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5734.6  on 4220  degrees of freedom
Residual deviance: 5035.5  on 4216  degrees of freedom
  (19 observations deleted due to missingness)
AIC: 5045.5
```

Logistic Regression Coefficients

- Log-odds or the **natural log of the odds** (AKA logit)
- Difficult to interpret.
- We can "exponentiate" logits to create odds ratios (ORs)
  - Easier to understand

  - OR = 1 (no effect)
  - OR < 1 (decrease in odds of outcome=1)
  - OR > 1 (increase in odds of outcome=1)

Interpreting an Odds Ratio

- **Odds Ratio (OR)**
  - Ratio of two values of X (Predictor) that are *one unit* apart
  - *Categorical Predictors*: OR reflects the odds of the Predictor=1 category vs. the Predictor=0 category on the Outcome=1 category
  - *Continuous Predictors*: OR reflects the increase/decrease in odds of the Outcome for a one unit increase in the Predictor

Categorical Example

OR reflects the odds of the Predictor=1 category vs. the Predictor=0 category on the Outcome=1 category

- Predictor = bmicat_X1 (0=~~Underweight~~, **1=Underweight**)
- Outcome = **Died via CVD (1)** vs. did not die via CVD (0)
- **bmicat_X1 OR =** $\boldsymbol{e^{0.16434} = 1.18}$
- **Interpretation:** The odds of **dying via CVD (Outcome=1)** are 1.18 times larger for those **who were Underweight according to the BMI (Predictor=1)** (compared to those who were Normal according to the BMI), holding all other variables in the model constant.

Continuous Example

OR reflects the increase/decrease in odds of the Outcome for a one unit increase in the Predictor

- Predictor = Age
- Outcome = **Died via CVD (1)** vs. did not die via CVD (0)
- **Age OR =** $\boldsymbol{e^{0.09858} = 1.10}$
- **Interpretation:** The odds of **dying via CVD (Outcome=1)** are 1.10 larger for **each additional year of life**, all else being equal.

Another way to report

- (OR - 1) X 100 = percent increase if positive, or decrease if negative, (over reference category of Predictor) in odds of outcome (Outcome)

Categorical Example

- Predictor = bmicat_X1 (0=~~Underweight~~, **1=Underweight**)
- Outcome = **Died via CVD (1)** vs. did not die via CVD (0)
- **bmicat_X1 OR =** $e^{0.16434} = \mathbf{1.18}$
- (1.18 – 1) X 100 = 0.18 X 100 = 18%
- For those **who were Underweight according to the BMI (Predictor=1)** (compared to those who were Normal according to the BMI), the odds of **dying via CVD (Outcome=1) increased by 18%,** all else being equal.

Continuous Example

- Predictor = Age
- Outcome = **Died via CVD (1)** vs. did not die via CVD (0)
- **Age OR =** $e^{0.09858} = \mathbf{1.10}$
- (1.10 – 1) X 100 = 0.10 X 100 = 10%
- **For each additional year of life (1 unit increase on Predictor),** the odds of **dying via CVD (Outcome=1) increase by 10%,** all else being equal.

Now for the intercept

- Linear Regression
  - Value of Outcome when all Predictors equal zero.
- Logistic Regression
  - Probability when all Predictors equal zero.
  - Baseline Probability

$$\frac{(e^{Intercept})}{(1+e^{Intercept})} = \text{Base Probability}$$

Baseline probability

$$\frac{\left(e^{-5.47823}\right)}{\left(1+e^{-5.47823}\right)}$$

$$\frac{0.00417}{(1 + 0.00417\ )}$$

$$= \mathbf{0.0042}$$

The predicted probability that someone will die from CVD is **.42%** when *Age* is zero and they are *Normal according to the BMI* (*Underweight* is zero, *Overweight* is zero, and *Obese* is zero).

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.47823    0.21981 -24.923  < 2e-16 ***
AGE          0.09858    0.00428  23.030  < 2e-16 ***
bmicat_X1    0.16434    0.30377   0.541 0.588508
bmicat_X2    0.27592    0.07372   3.743 0.000182 ***
bmicat_X3    0.68385    0.10759   6.356 2.07e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5734.6  on 4220  degrees of freedom
Residual deviance: 5035.5  on 4216  degrees of freedom
  (19 observations deleted due to missingness)
AIC: 5045.5
```

$$e^{Coefficient} = \text{Odds Ratio} \qquad \frac{(e^{Intercept})}{(1+e^{Intercept})} = \text{Base Probability}$$

Characteristic	OR[1]	95% CI[1]	p-value
(Intercept)	0.00	0.00, 0.01	**<0.001**
AGE	1.10	1.09, 1.11	**<0.001**
bmicat_X1	1.18	0.64, 2.12	0.6
bmicat_X2	1.32	1.14, 1.52	**<0.001**
bmicat_X3	1.98	1.61, 2.45	**<0.001**

[1]OR = Odds Ratio, CI = Confidence Interval

Model Assessment

- Several possible metrics
  - Loglikelihood (LL); Negative loglikelihood (-LL, deviance); Akaike information criterion (AIC); Bayesian information criterion (BIC); Brier score (analogous to RMSE^2)
  - Accuracy, Sensitivity, Specificity, ROC-AUC

	Estimate
Accuracy	*0.68*
Sensitivity	0.54
Specificity	**0.78**
AUC	**0.73**