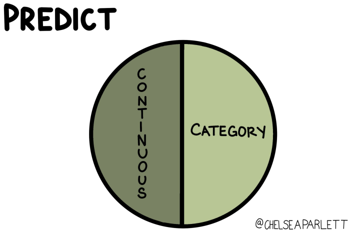
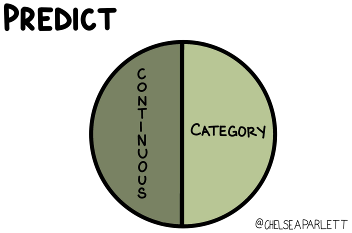


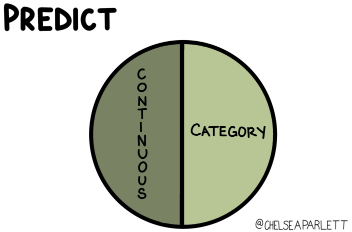
Logistic Regression

Dr. Chelsea Parlett-Pelleriti

Linear Regression in Disguise



Predictions



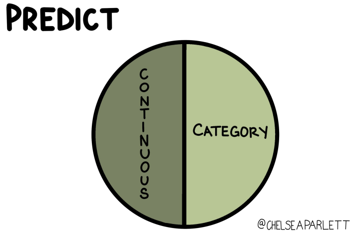
Linear Regression

?

Logistic Regression

Continuous Variable (can be -∞ to ∞)

Binary Categorical Variable (can be 0 or 1)

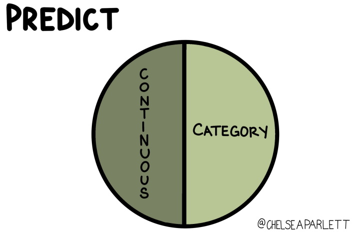


Getting from Binary to Continuous

Linear Models don't predict categorical variables…so what's the closest thing? PROBABILITIES!

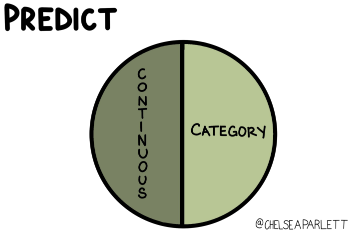
1. We want Probabilities
   - But probabilities are bounded 0-1🙁
2. Convert Probabilities to Odds
   - But Odds only go down to 0, and are not symmetric🙁
3. Convert Odds to Log Odds

Getting from Binary to Continuous

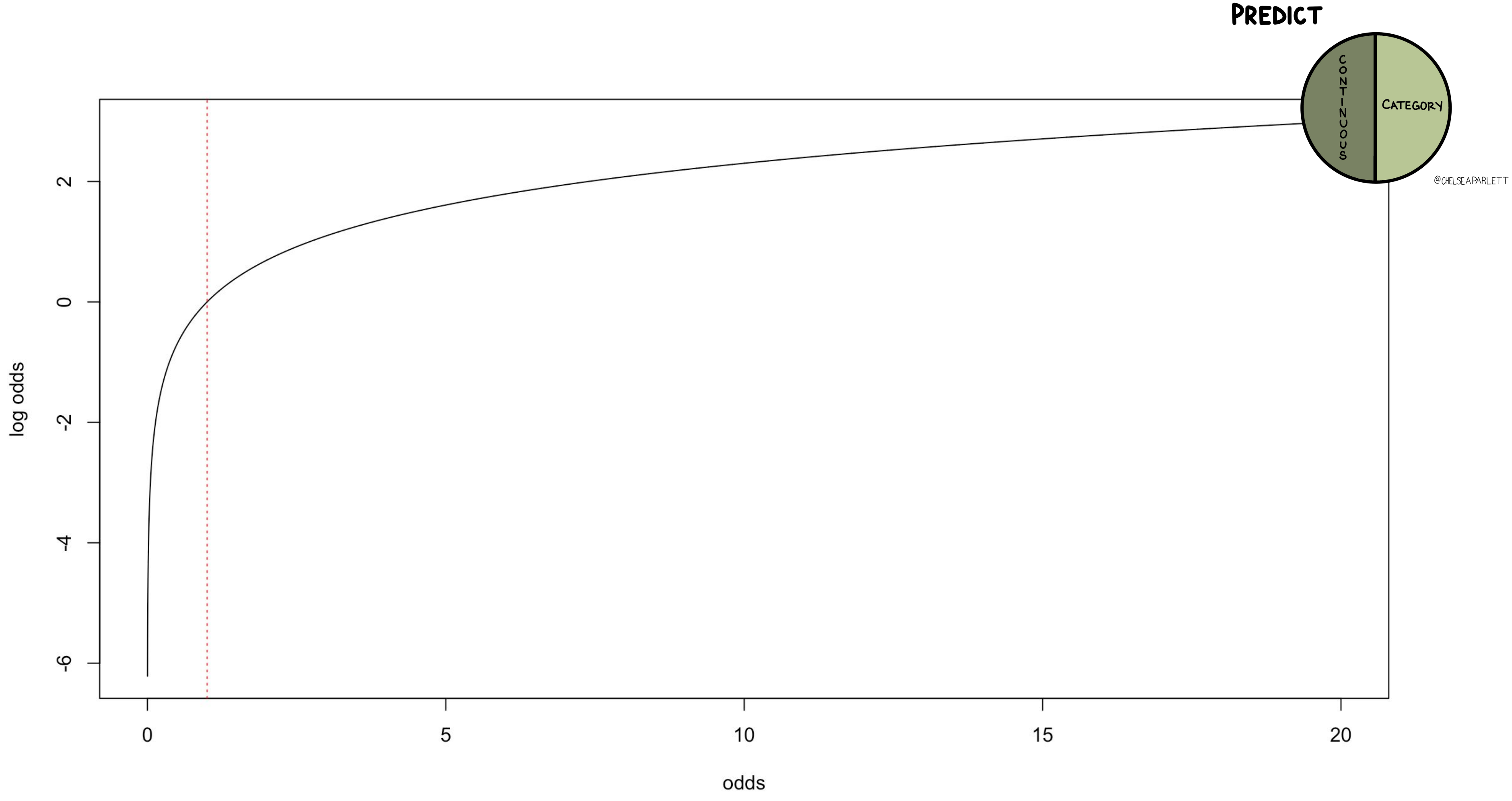


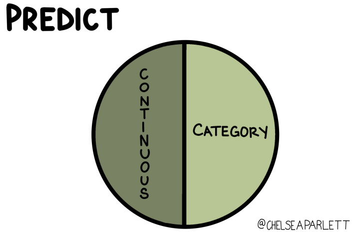
1. We want Probabilities
2. Convert Probabilities to Odds
3. Convert Odds to Log Odds

Getting from Binary to Continuous



1. We want Probabilities
2. Convert Probabilities to Odds
3. Convert Odds to Log Odds
4.

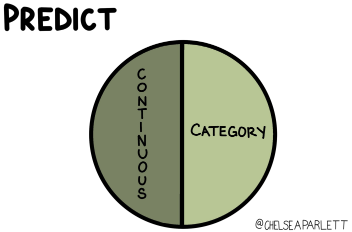

PREDICT
CONTINUOUS
CATEGORY
@CHELSEAPARLETT
log odds
2
0
-2
-4
-6
0
5
10
15
20
odds



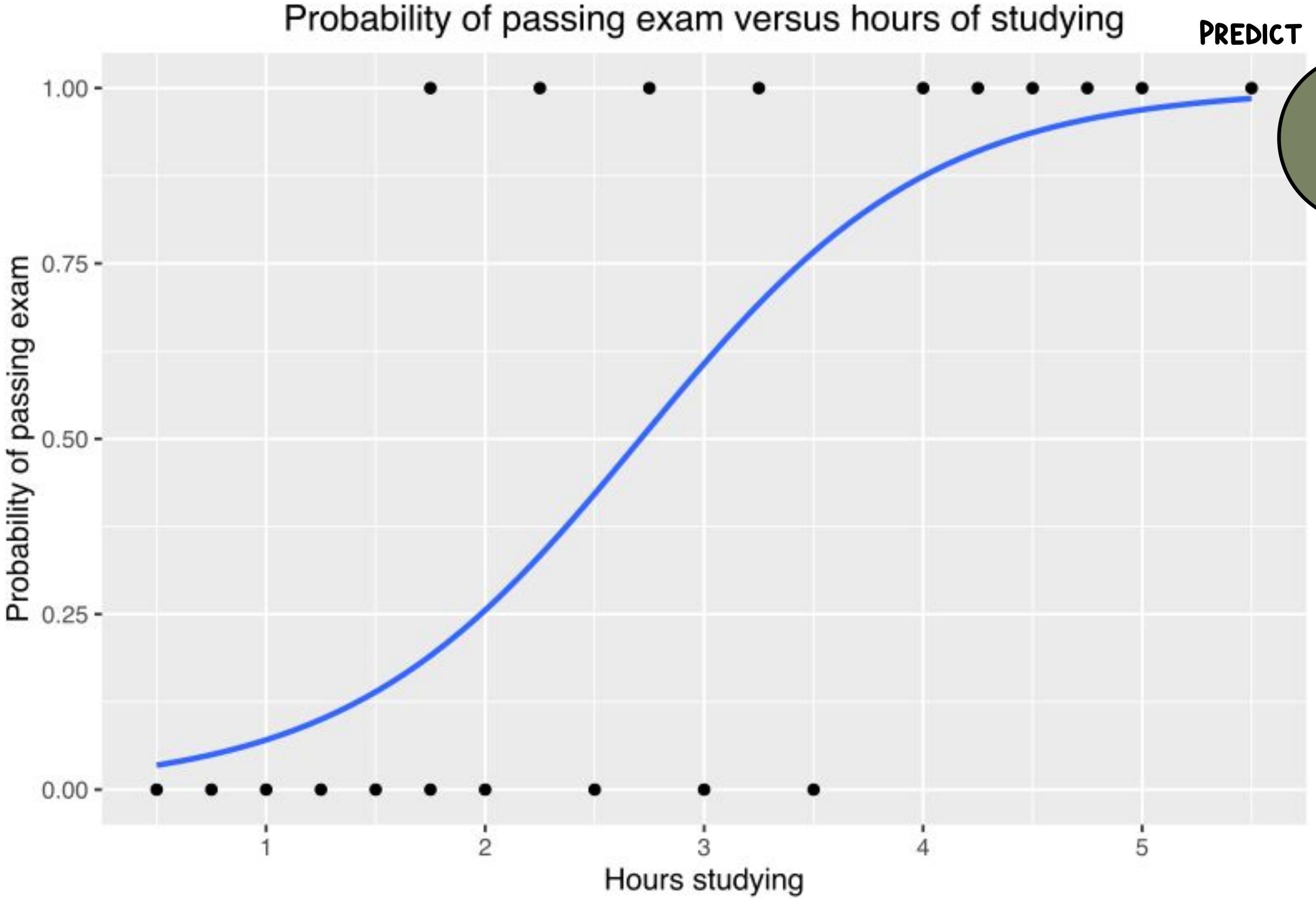
Getting from Binary to Continuous

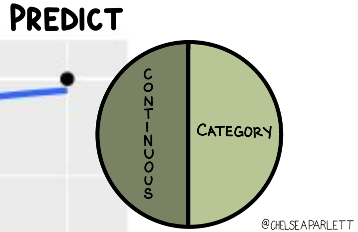
1. We want Probabilities
2. Convert Probabilities to Odds
3. Convert Odds to Log Odds
4.

The Final Formula



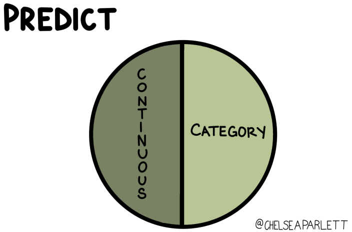
$$\log(p/1-p) = mx + b$$





https://en.wikipedia.org/wiki/Logistic_regression

Assessing Model Performance



- Mainly worried about real world performance: did it make the correct prediction? (accuracy, sensitivity, specificity, F1 score)
- ROC/AUC

Accuracy

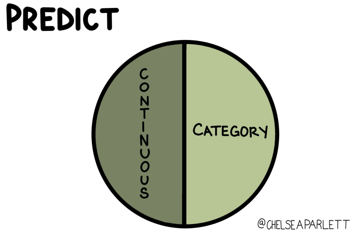
Correct Predictions

True Positive **(TP)** + True Negative **(TN)**

All Predictions

False Negative **(FN)** + False Positive **(FP)** +

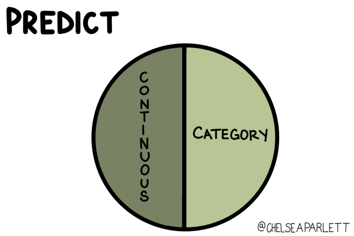
True Positive **(TP)** + True Negative **(TN)**



"How often is the model correct?"

	Predicted Positive	***Predicted*** Negative
Actual Positive	True Positive **(TP)**	False Negative **(FN)**
Actual Negative	False Positive **(FP)**	True Negative **(TN)**

Sensitivity/Recall



Correctly Predicted Positives

True Positive (**TP**)

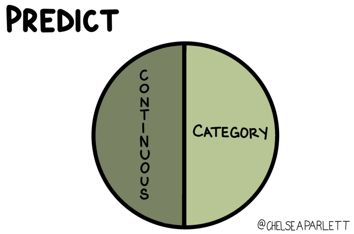
Actual Positives

False Negative **(FN)** + True Positive (**TP**)

"How often is the model correct for Positive Cases?"

	Predicted Positive	***Predicted*** Negative
Actual Positive	True Positive **(TP)**	False Negative **(FN)**
Actual Negative	False Positive **(FP)**	True Negative **(TN)**

Specificity



Correctly Predicted Negatives

True Negative **(TN)**

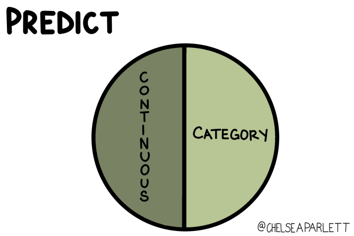
Actual Negatives

True Negative **(TN)** + False Positive **(FP)**

"How often is the model correct for Negative Cases?"

	Predicted Positive	***Predicted*** Negative
Actual Positive	True Positive **(TP)**	False Negative **(FN)**
Actual Negative	False Positive **(FP)**	True Negative **(TN)**

Precision



Correctly Predicted Positives

True Positive (**TP**)

Actual Positives

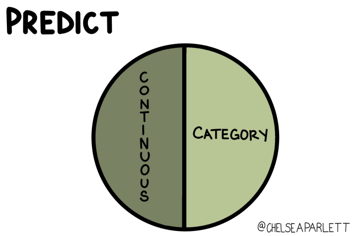
False Positive **(FP)** + True Positive (**TP**)

"How many of the predicted Positives are correct?"

	Predicted Positive	***Predicted*** Negative
Actual Positive	True Positive **(TP)**	False Negative **(FN)**
Actual Negative	False Positive **(FP)**	True Negative **(TN)**

F1 Score

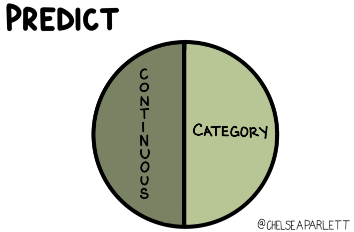
$$\frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$



Combination of Precision (how often predicted positives ARE positive) and Recall (how often we correctly predict actual positives)

	Predicted Positive	***Predicted*** Negative
Actual Positive	True Positive **(TP)**	False Negative **(FN)**
Actual Negative	False Positive **(FP)**	True Negative **(TN)**

ROC AUC



Correctly Predicted Positives

True Positive (**TP**)

Actual Positives

False Negative **(FN)** + True Positive (**TP**)

	Predicted Positive	***Predicted*** Negative
Actual Positive	True Positive (**TP**)	False Negative **(FN)**
Actual Negative	False Positive **(FP)**	True Negative **(TN)**

1 -

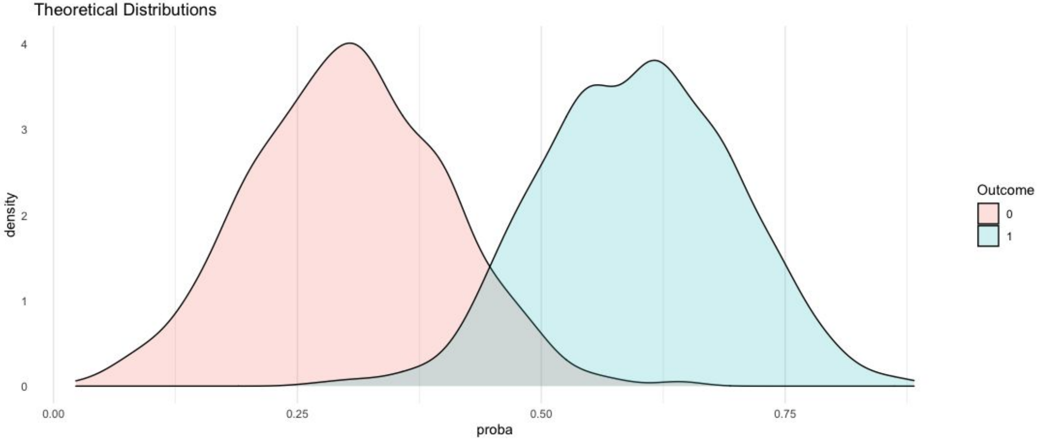
Correctly Predicted Negatives

True Negative **(TN)**

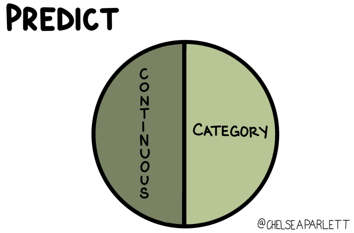
Actual Negatives

True Negative **(TN)** + False Positive **(FP)**

ROC AUC


Theoretical Distributions
density
proba
0.00
0.25
0.50
0.75
0
1
2
3
4
Outcome
0
1

Loss Functions



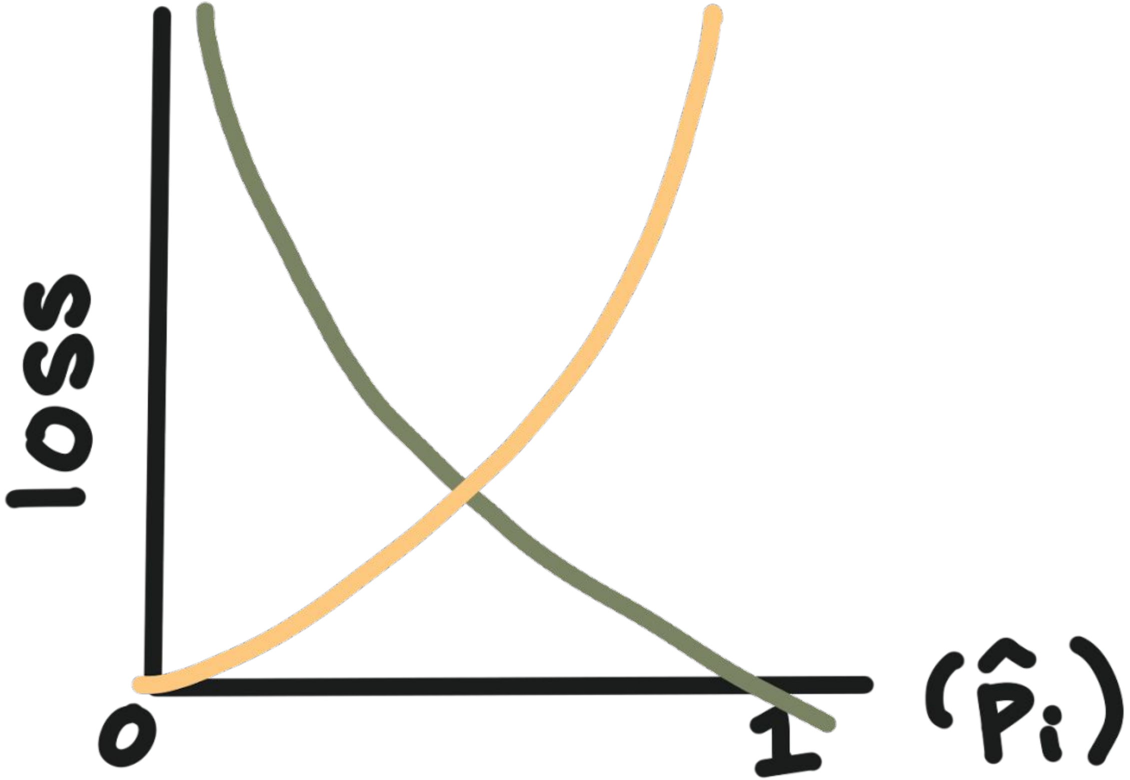
LINEAR:

$$(\hat{y}_i - y_i)^2$$

LOGISTIC:

$$\begin{cases} -\log(\hat{p}_i) & \text{if } y=1 \\ -\log(1-\hat{p}_i) & \text{if } y=0 \end{cases}$$

Loss Functions



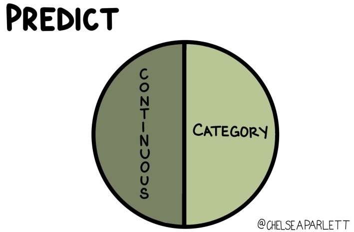
LINEAR:

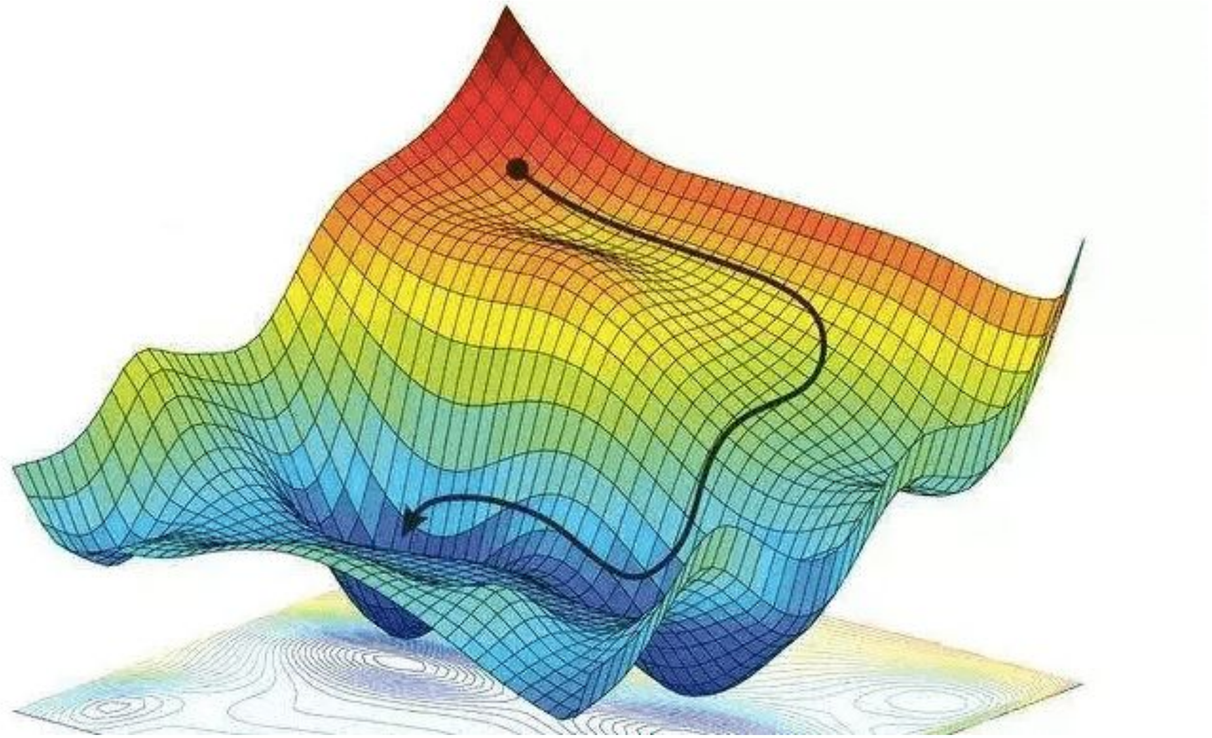
$$(\hat{y}_i - y_i)^2$$

LOGISTIC:

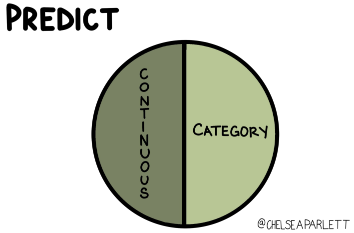
$$\begin{cases} -\log(\hat{p}_i) & \boxed{\text{if } y=1} \\ -\log(1-\hat{p}_i) & \boxed{\text{if } y=0} \end{cases}$$

Gradient Descent





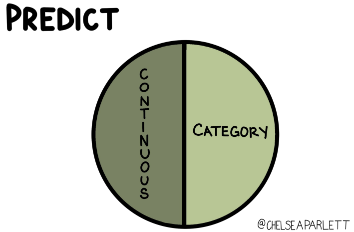
Getting from Binary to Continuous



Linear Models don't predict categorical variables…so what's the closest thing? PROBABILITIES!

1. We want Probabilities
    - But probabilities are bounded 0-1🙁
2. Convert Probabilities to Odds
    - But Odds only go down to 0, and are not symmetric🙁
3. Convert Odds to Log Odds

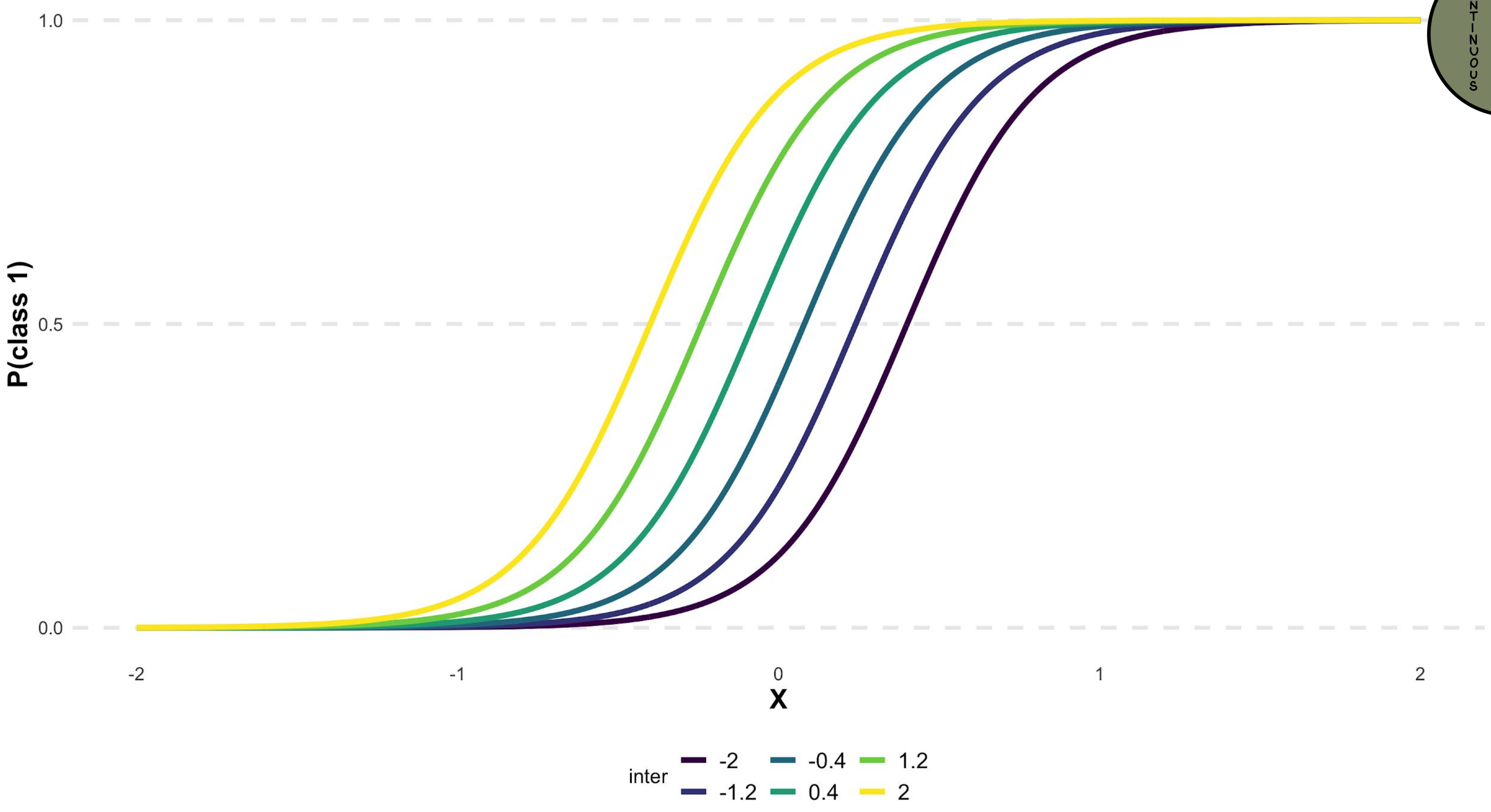
All the Steps

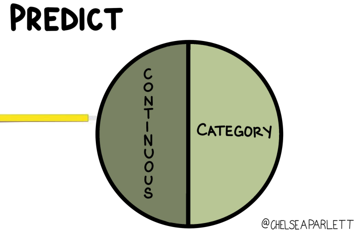


Probability p	**Odds** p/(1-p)	**Log Odds** log(p/1-p)
0.1	0.1111	-2.1972
0.5	**1**	**0**
0.9	9	2.1972

Logistic Curves with different Intercepts

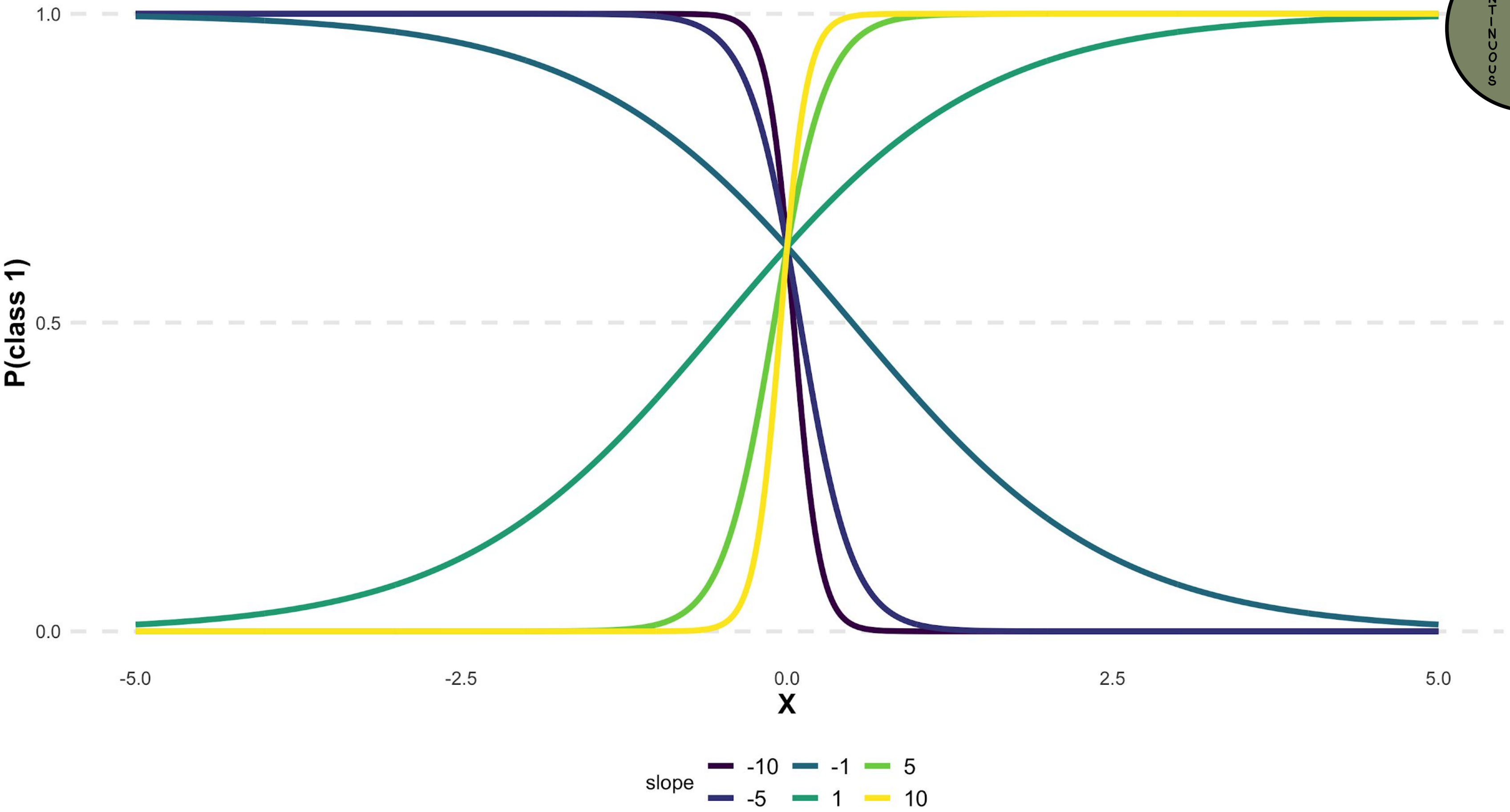
inter + 5x

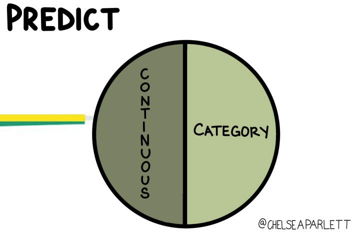


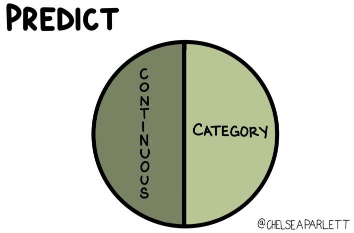


Logistic Curves with different Slopes

0.5 + slope*x



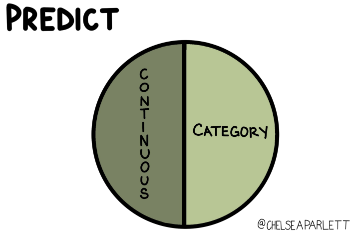




All the Steps

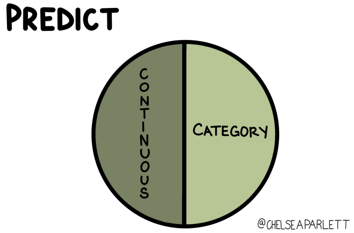
Probability p	**Odds** p/(1-p)	**Log Odds** log(p/1-p)
0.1	0.1111	-2.1972
0.5	**1**	**0**
0.9	9	2.1972

Interpreting Coefficients

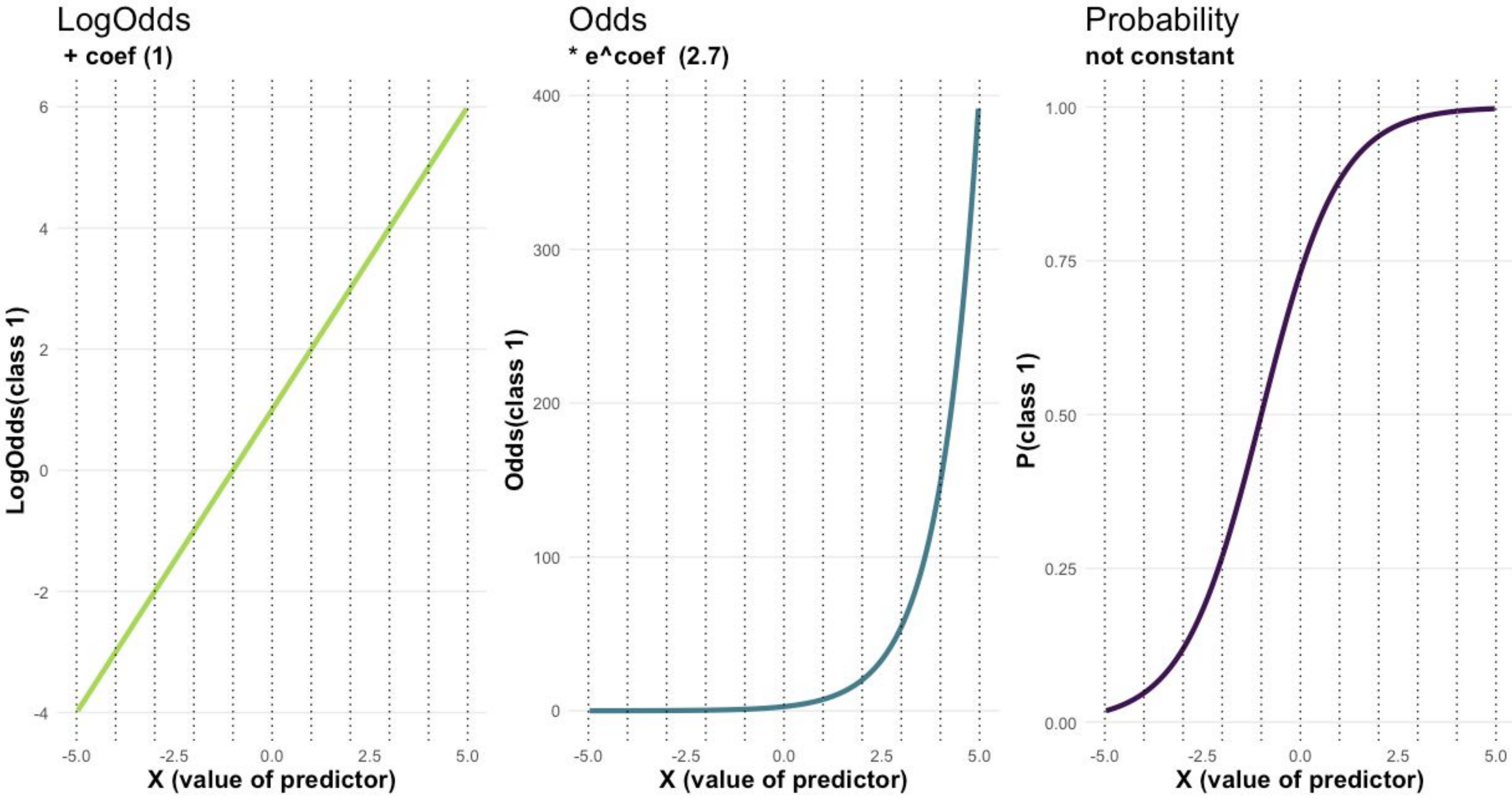


	coef
const	-2.9777
age	0.1445
income	-0.0066
months_subbed	0.0015

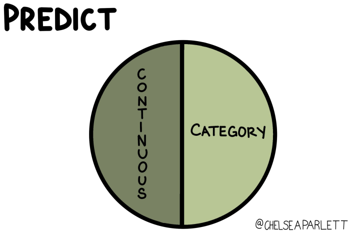
Interpreting Coefficients



	coef	e^{coef}
const	-2.9777	0.05090979
age	0.1445	1.155462
income	-0.0066	0.9934217
months_subbed	0.0015	1.001501


LogOdds
+ coef (1)
LogOdds(class 1)
X (value of predictor)
Odds
* e^coef (2.7)
Odds(class 1)
X (value of predictor)
Probability
not constant
P(class 1)
X (value of predictor)

Interpreting Coefficients



✅ Log odds

✅ Odds

⚠️ Probabilities*