NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 14: Lecture 1
Logistic Regression

How can we modify linear regression to predict qualitative variables?

DS-UA 112
Introduction to Data Science

Week 14: Lecture 1
Logistic Regression

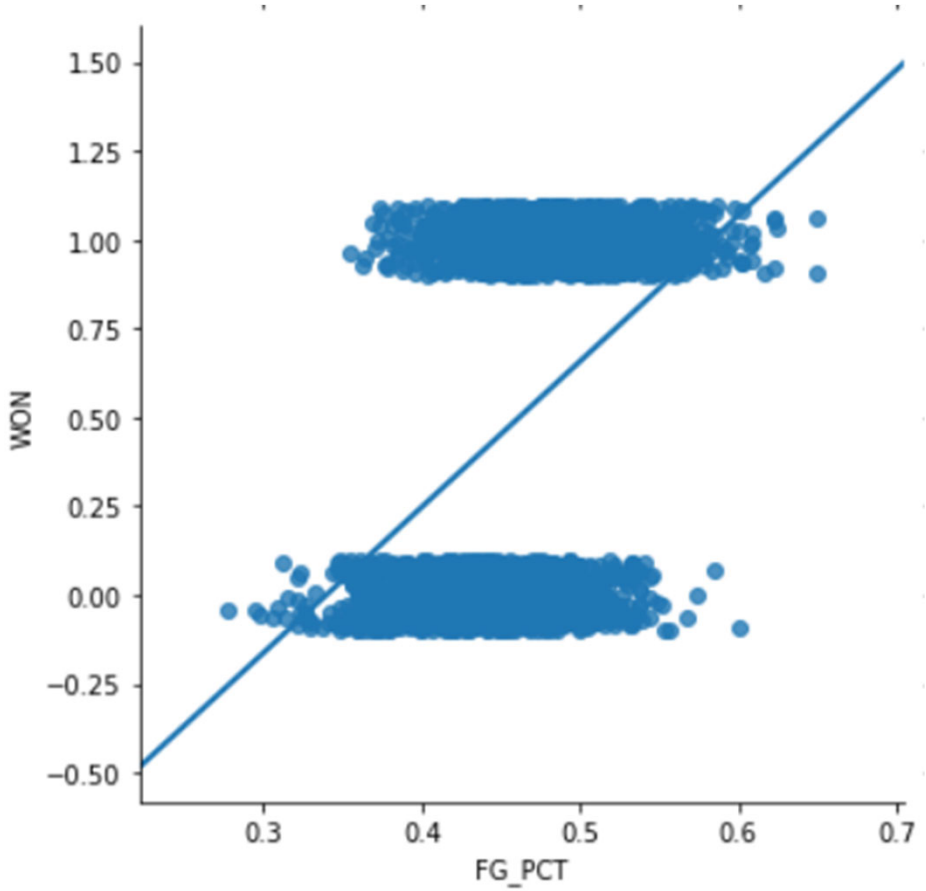
Adapted from Nolan, Speed, Gonzalez, Lau

Announcements

- Please check Week 14 agenda on NYU Classes
  - Lab 13
    - Due on Friday May 1 at 11:59PM EST
  - Project 2
    - Due on Tuesday May 12 at 11:59PM EST

Review

- We use linear regression to predict a quantitative response variable.
- We use logistic regression to predict a qualitative response variable.
- Usually we encode the categories with numbers like 0 and 1



Review

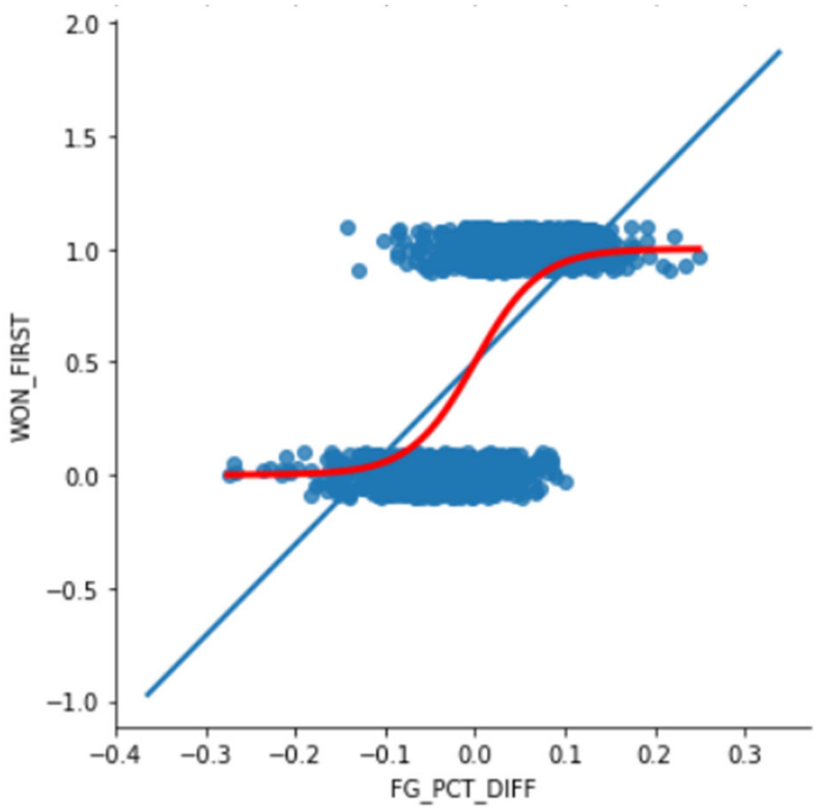
- We can fit a curve to the data with the sigmoid function

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

- If we replace t with

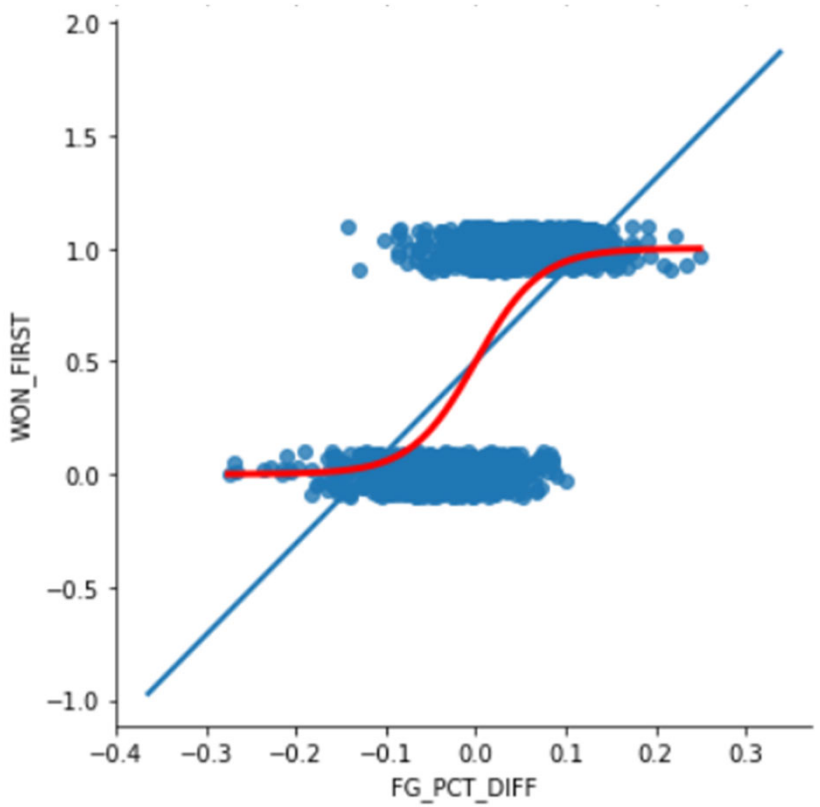
$$w_0 + w_1 * t$$

for intercept w_0 and slope w_1, then we can adjust the shape of the curve



Review

- Each value of the explanatory variable determines a guess for the category of the response variable.
- However we predict numbers between 0 and 1. Each number represents the probabilities of the response variable equaling 1 conditional on the value of the explanatory variable

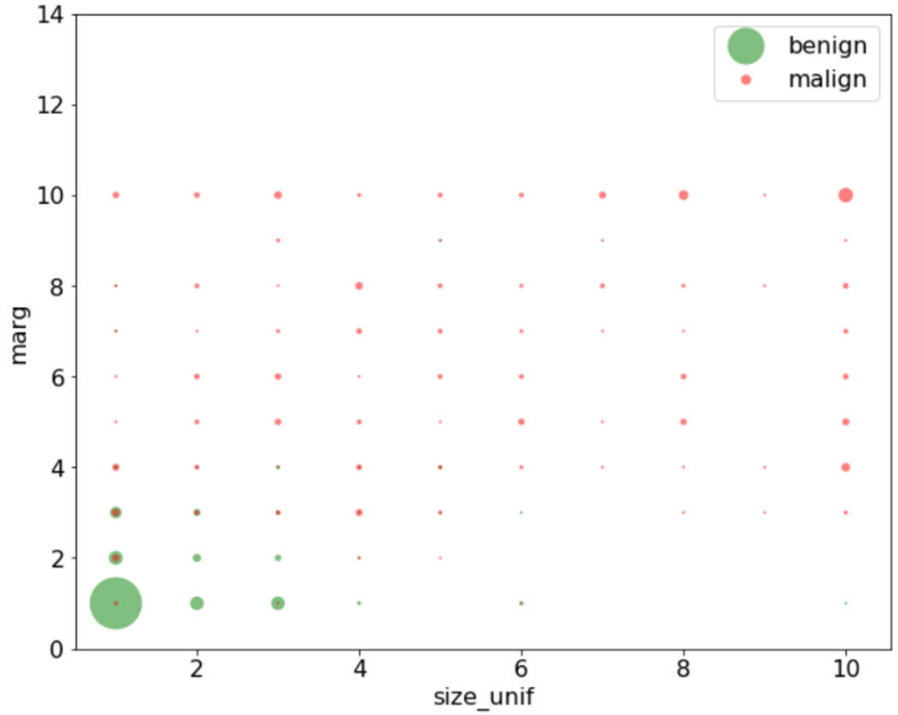


Agenda

- Logistic Regression
  - Thresholds
- Classification
  - Accuracy
  - Precision
  - Recall

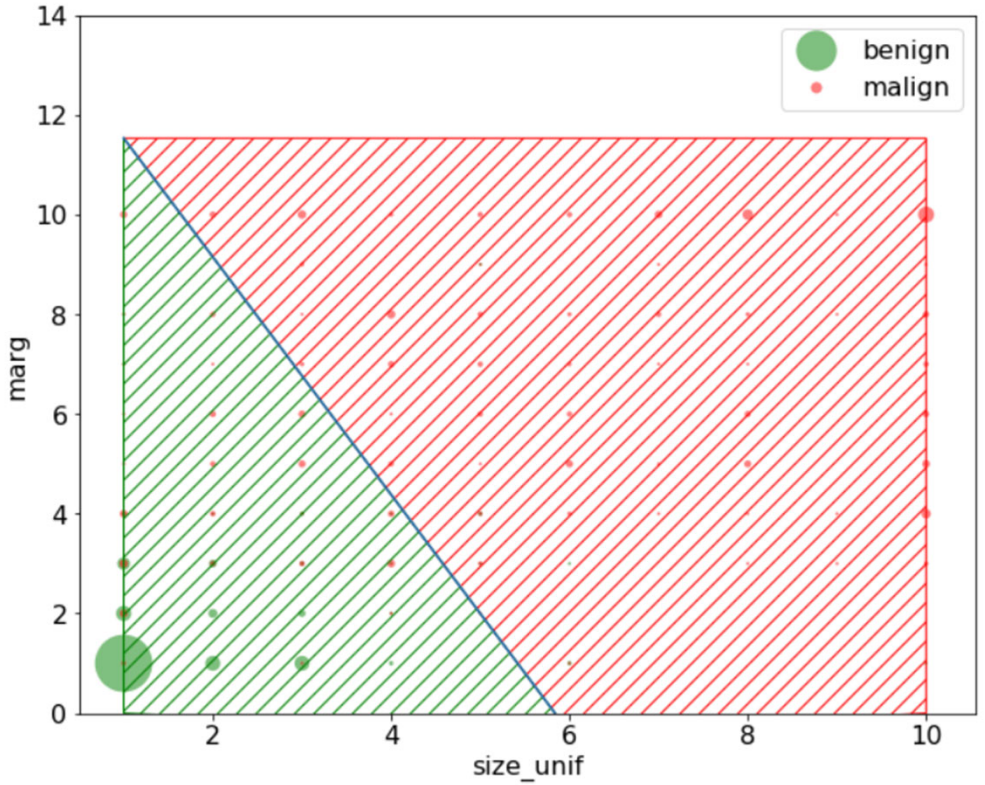
Histograms

- Remember that we use scatter-plots to visualize two quantitative variables. We represent the values of the variables with the horizontal coordinate and vertical coordinate
- Additionally we can adjust the size and color of the points.
  - The color could represent the two categories
  - The size could represent the frequency of points in a region
- Using size and color in scatter-plots allows us to represent 3-dimensional histograms in 2-dimensions



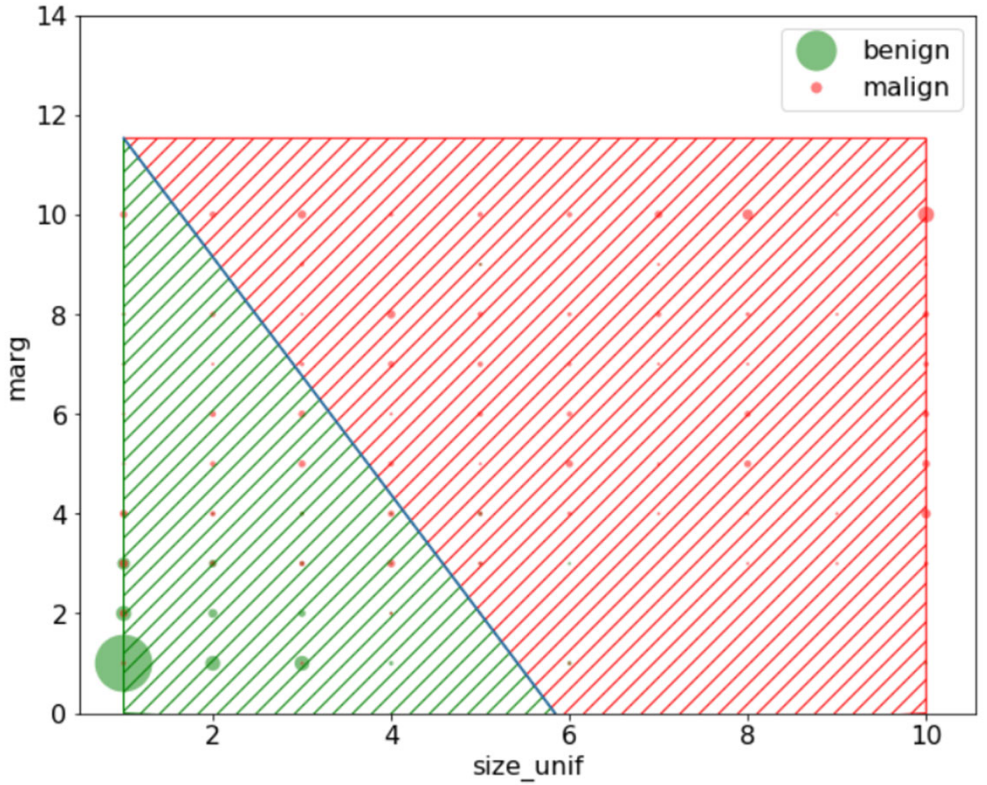
Thresholds

- We predict numbers between 0 and 1 in logistic regression. However we need either 0 or 1 for classification into categories.
- We can round the numbers up to 1 or down to 0 based on a threshold.
- Usually we take the threshold to be 0.5. However we can decrease or increase the threshold depending on the context



Decision Boundary

- Based on the explanatory variables, we classify the response variable into category 1 or category 0.
- These classifications determine two regions separated by the decision boundary
- The points on the decision boundary correspond to predicted probabilities equal to the threshold



Metrics

- We have different approaches to evaluating the classifications. We compare the observations and predictions with various metrics.
- The accuracy measures the number of correct classification. The error measures the number of incorrect classifications

$$\text{accuracy} = \frac{\#\text{ of points classified correctly}}{\#\text{ points total}}$$

$$\text{error} = 1 - \text{accuracy} = \frac{\#\text{ of points classified incorrectly}}{\#\text{ points total}}$$

Metrics

- We have different approaches to evaluating the classifications. We compare the observations and predictions with various metrics.
- The accuracy measures the number of correct classification. The error measures the number of incorrect classifications

$$\text{accuracy} = \frac{\#\text{ of points classified correctly}}{\#\text{ points total}}$$

$$\text{error} = 1 - \text{accuracy} = \frac{\#\text{ of points classified incorrectly}}{\#\text{ points total}}$$

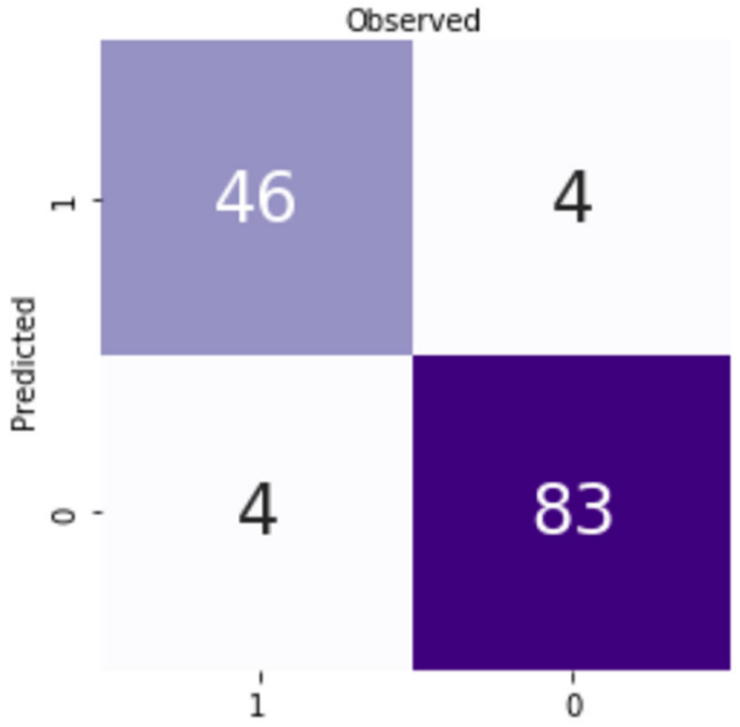
Confusion Matrix

- The observation take the value 1 or 0. The predictions take the value 1 or 0. So we have four possibilities
  - True Positive
  - False Positive
  - False Negative
  - True Negative

		Truth	
		1	**0**
Prediction	**1**	**TP**: True Positive	**FP**: False Positive
	0	**FN**: False Negative	**TN**: True Negative

Confusion Matrix

- The observation take the value 1 or 0. The predictions take the value 1 or 0. So we have four possibilities
  - True Positive
  - False Positive
  - False Negative
  - True Negative
- We can visualize the number of each possibility for a dataset with a confusion matrix



Confusion Matrix

- We can determine metrics from different combinations of these four possibilities.

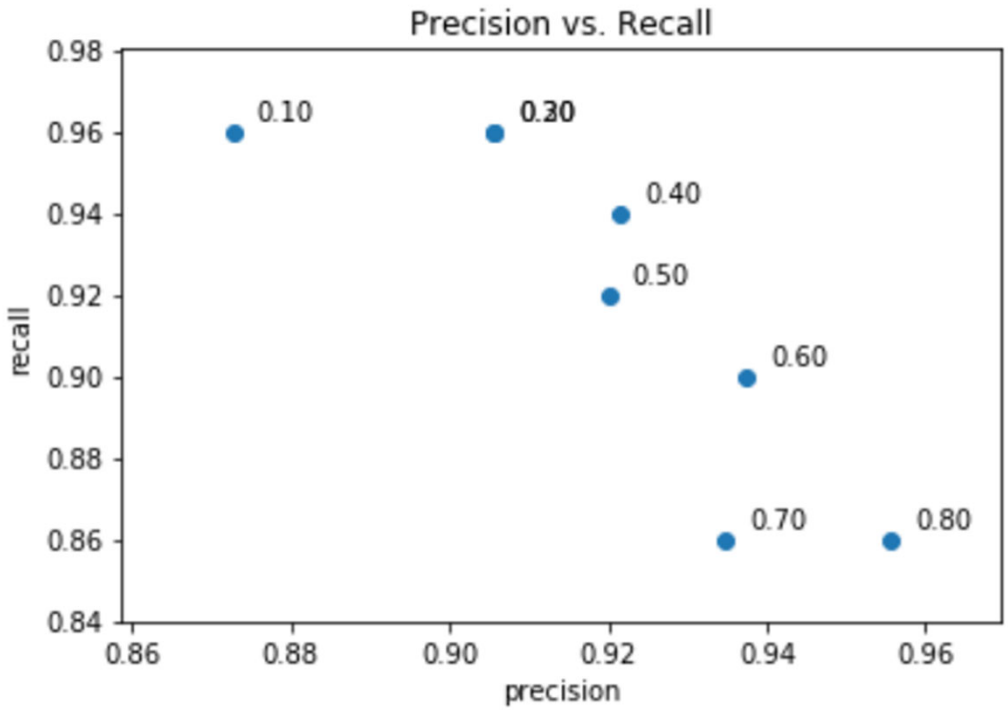
$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{n}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

Precision Recall Curve

- Accuracy might not capture the differences between observations and prediction with an imbalance between categories
  - Precision penalizes false positives
  - Recall penalizes false negative
- We can visualize the trade-off between recall and precision through a precision-recall curve



Summary

- Logistic Regression
  - Thresholds
- Classification
  - Accuracy
  - Precision
  - Recall

Goals

- Convert a probability to a category with a threshold to use logistic regression for classification
- Use the metrics accuracy, recall and precision to evaluate classifications