UCDAVIS HEALTH

Logistic Regression: Application to Clinical Classification

CLINICAL AND TRANSLATIONAL SCIENCE CENTER

Sandra Taylor, Ph.D.

Principal Biostatistician

CTSC Biostatistics Program Manager

The UC Davis CTSC receives support from the NIH National Center for Advancing Translational Sciences (award TR001860).

Sponsored by Biostatistics Programs affiliated with the following NIH Centers

Clinical and Translational Sciences Center

MIND Intellectual and Developmental Disability Center

Comprehensive Cancer Center

Environmental Health Sciences Center

Objectives

- Be able to use logistic regression for classification
- Understand the link between logistic regression and ROC curves, AUC, sensitivity and specificity
- Appreciate trade-offs associated with selecting classification cut-offs.

Classification

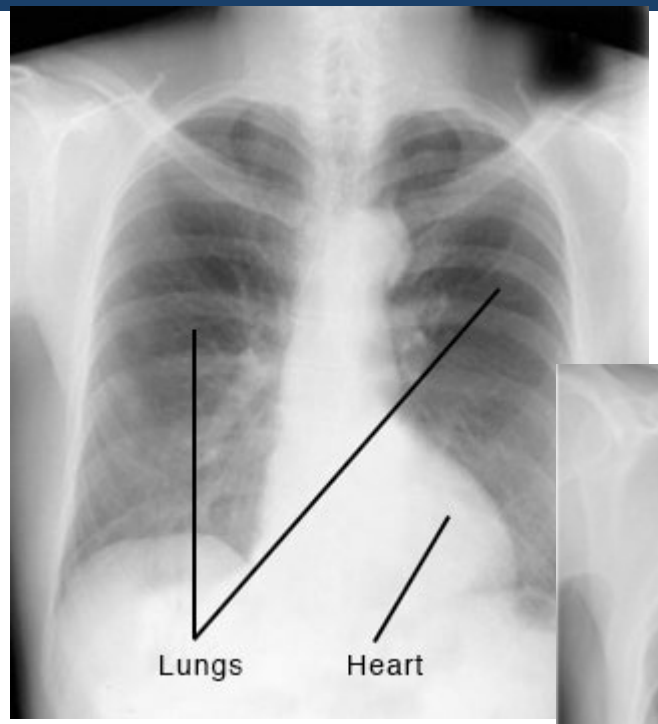
Objective: Based on observed data, develop a rule that assigns patients to a group of clinical interest.

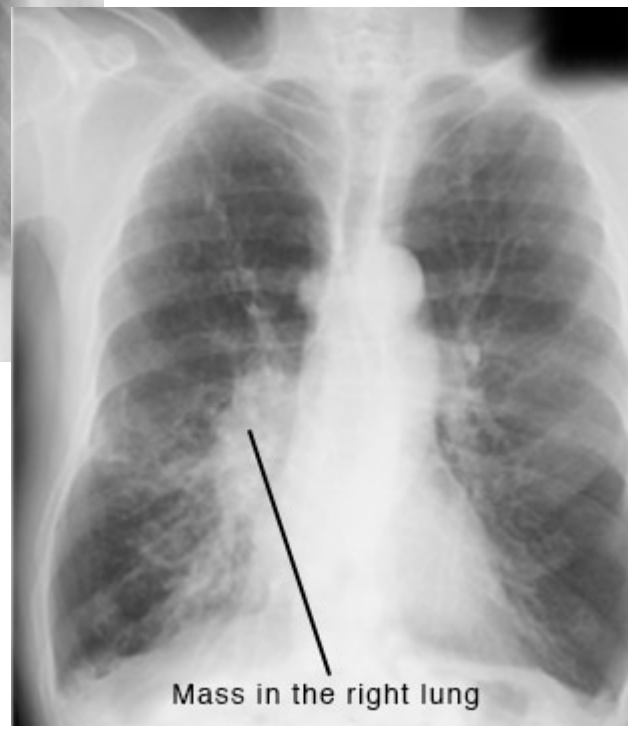
<u>Examples:</u>

Cancer vs. No Cancer

Responder vs. Non-responder

Adverse event vs. No adverse event





Note: Classification often entails converting a quantitative value to a qualitative value which results in some loss of information

Motivating Example: CA-125 for Ovarian Cancer Diagnosis

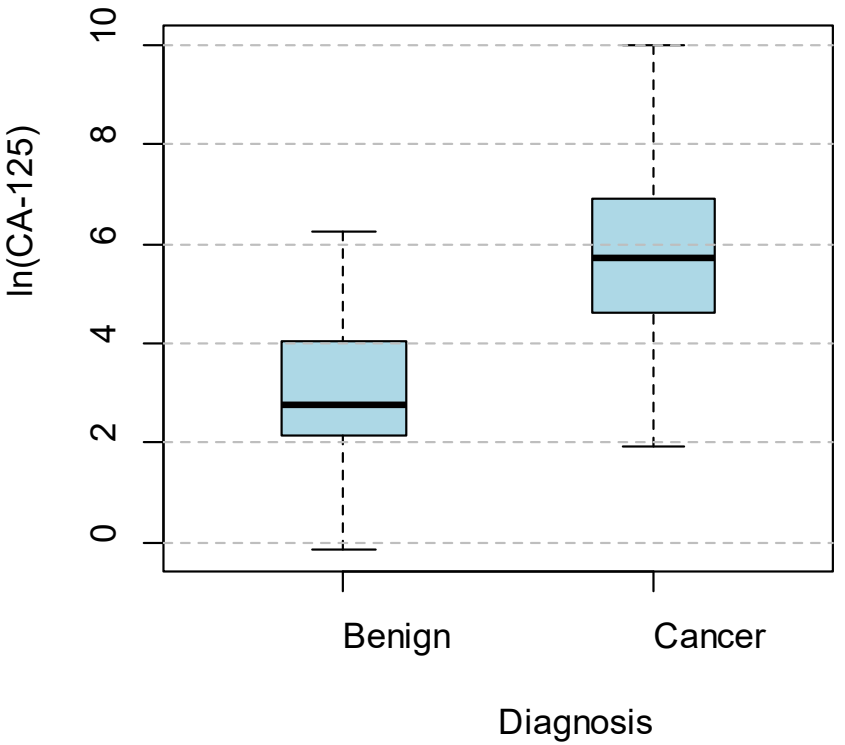
CA-125 is a glycoprotein with potential as a biomarker for ovarian cancer.

Data of CA-125 levels in serum of women diagnosed with stage III/IV ovarian cancer and women with benign ovarian masses.

CA-125 higher in women with cancer

- Benign: 3.1 units/mL (natural log transformed)
- Cancer: 5.7 units/mL (natural log transformed)

Highly significant difference: t-test p-value < 0.001



Note overlap in distributions
No rule will be perfect.

CA-125 Classification Rule

Consider a threshold value of 4 and classify

Benign: log(CA125) < 4

Cancer: log(CA125) ≥ 4

How good do we do?

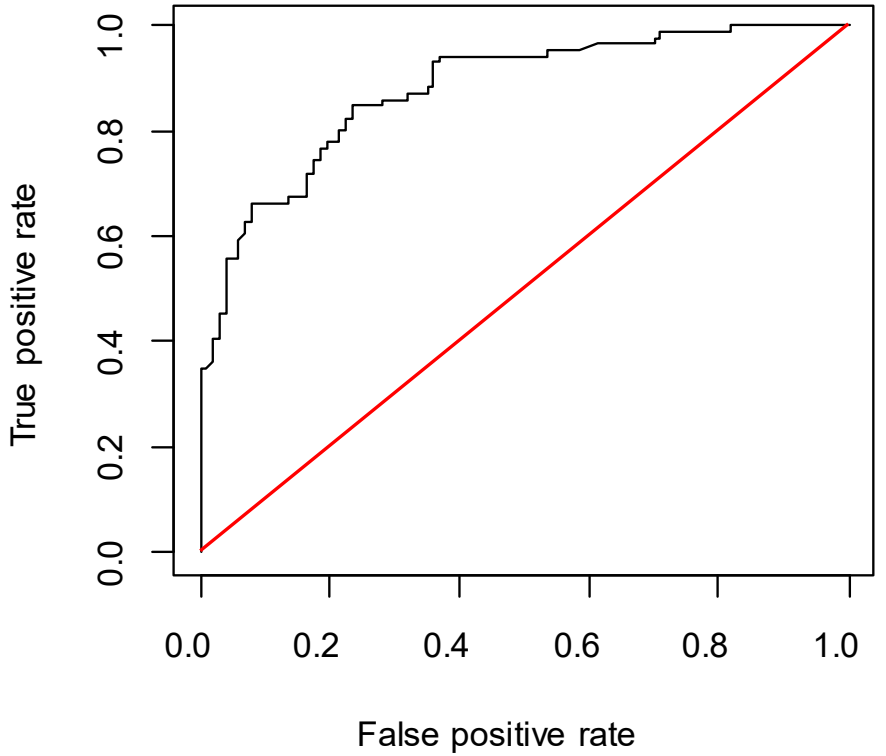
	Predicted Benign	Predicted Cancer	Total
True Benign	75	28	103
True Cancer	13	73	86
Total	88	101	189

Classifier performance

AUC (Area under the curve)		0.88
Sensitivity TP/(TP+FN)	73/(73+13)	0.85
Specificity TN/(TN+FP)	75/(75+28)	0.73

	Predicted Benign	Predicted Cancer	Total
True Benign	75	28	103
True Cancer	13	73	86
Total	88	101	189

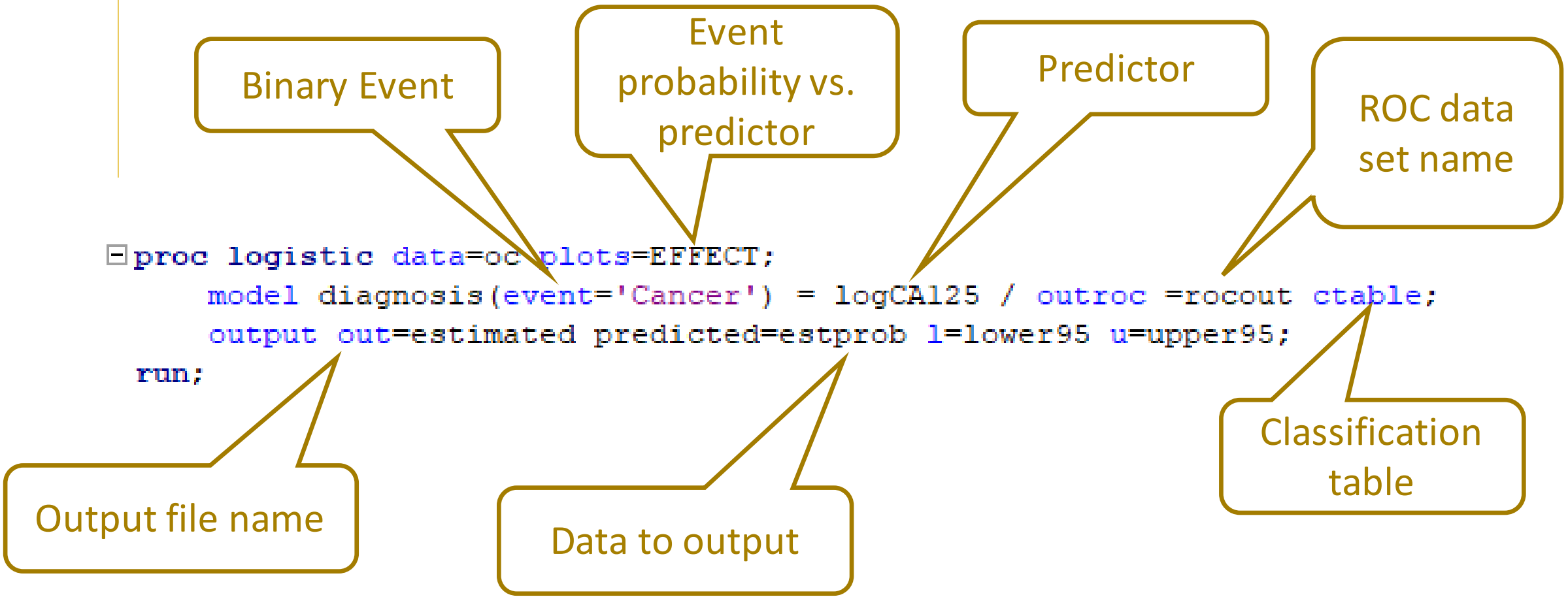
Receiver-Operating Characteristic (ROC) Curve



Develop classifier using logistic regression

- Fit logistic regression, modeling log odds of ovarian cancer (Y/N) vs. CA-125 levels
- Estimate the relationship between probability of cancer and CA-125
- Construct Receiver-Operating Characteristic curve
- Calculate AUC values
- Select probability threshold for classification
- Construct confusion matrix
- Calculate sensitivity and specificity

Fit logistic regression using Proc Logistic



SAS Logistic Regression Output

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.8999	0.6894	50.5128	<.0001
logCA125	1	1.0854	0.1513	51.4524	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
logCA125	2.961	2.201	3.983

- Significant positive relationship between the log odds of cancer and CA-125 levels
  - Estimate = 1.08
- Odds ratio = 2.96
  - For every 1 point increase in log transformed values of CA-125 the odds of cancer increases by nearly 3

Odds ratios aren't helpful for classification.

Need to convert output to estimate the probability of cancer for a given CA-125 level.

Converting from log odds to event probability

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.8999	0.6894	50.5128	<.0001
logCA125	1	1.0854	0.1513	51.4524	<.0001

Logistic Regression Model

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

p = probability of cancer
α = intercept = -4.90
β = CA-125 effect = 1.08
x = log(CA-125 value)

Re-arrange to estimate probability of cancer

$$\left(\frac{p}{1-p}\right) = e^{\alpha+\beta x} \longrightarrow p = (1-p)e^{\alpha+\beta x} = e^{\alpha+\beta x} - pe^{\alpha+\beta x} \longrightarrow p + pe^{\alpha+\beta x} = e^{\alpha+\beta x}$$

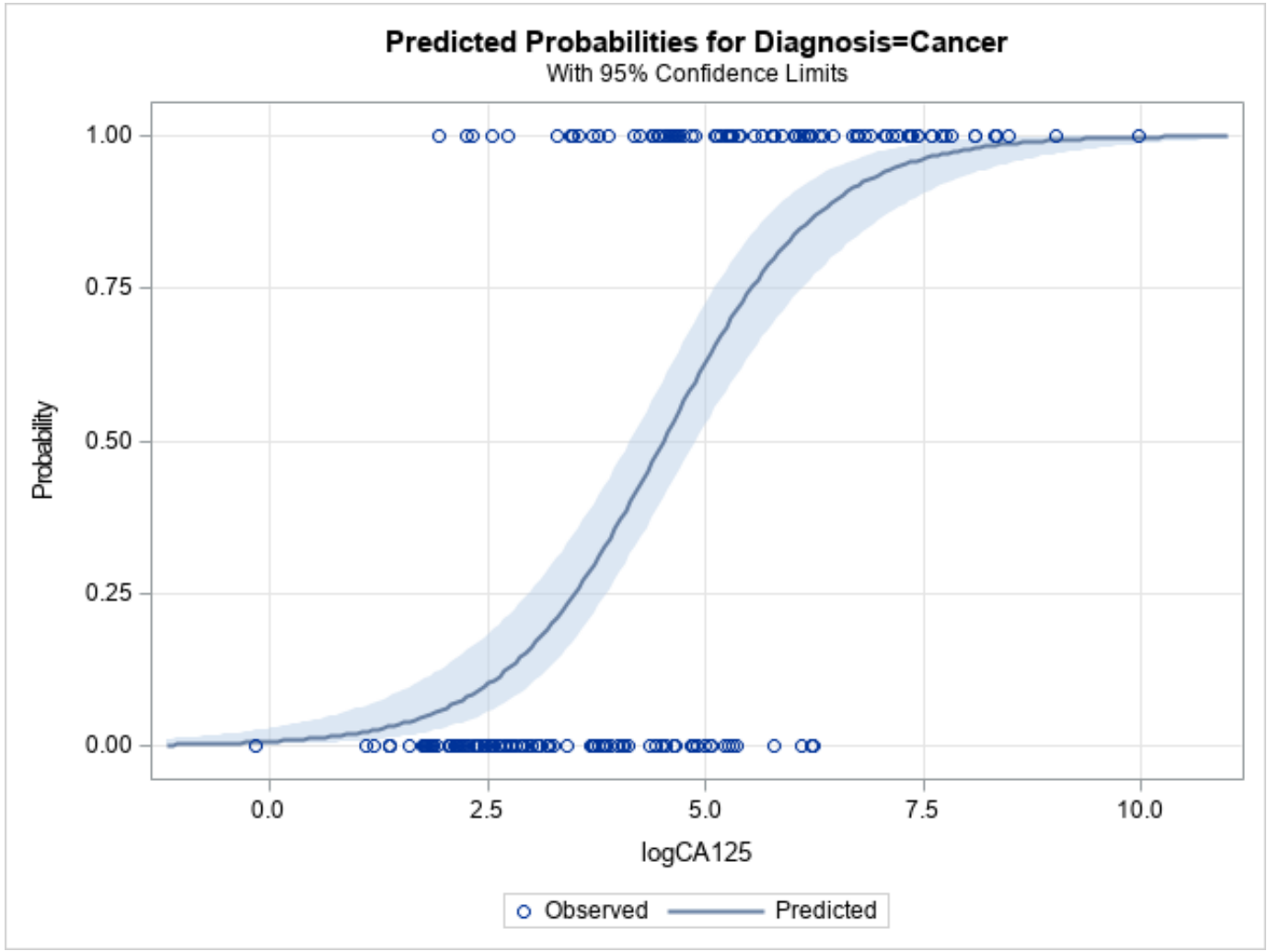
$$\longrightarrow p(1+e^{\alpha+\beta x}) = e^{\alpha+\beta x} \longrightarrow \boldsymbol{p = \frac{e^{\alpha+\beta x}}{(1+e^{\alpha+\beta x})}}$$

Probability of cancer for specified value of CA-125

Probability of Cancer vs. CA-125 levels

Probability of cancer increases with CA-125 levels

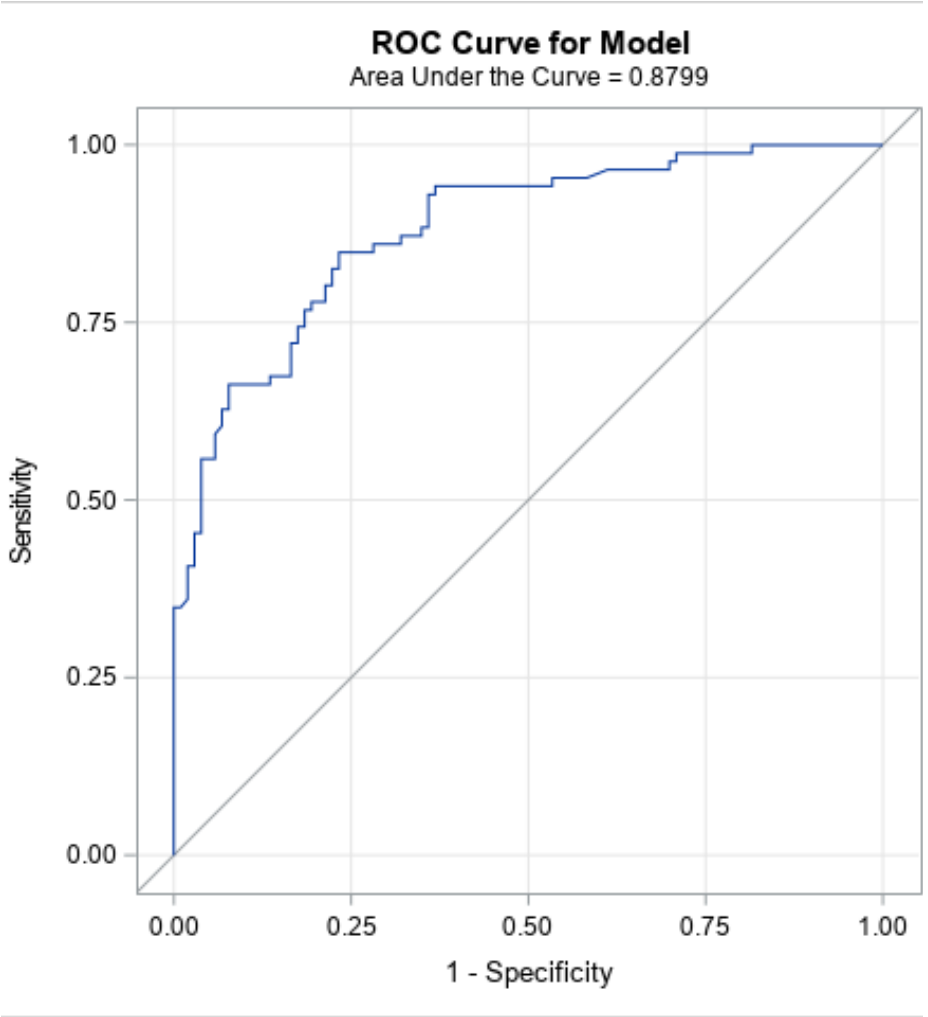
Considerable overlap in distribution of CA-125 by cancer status



How good is our model?

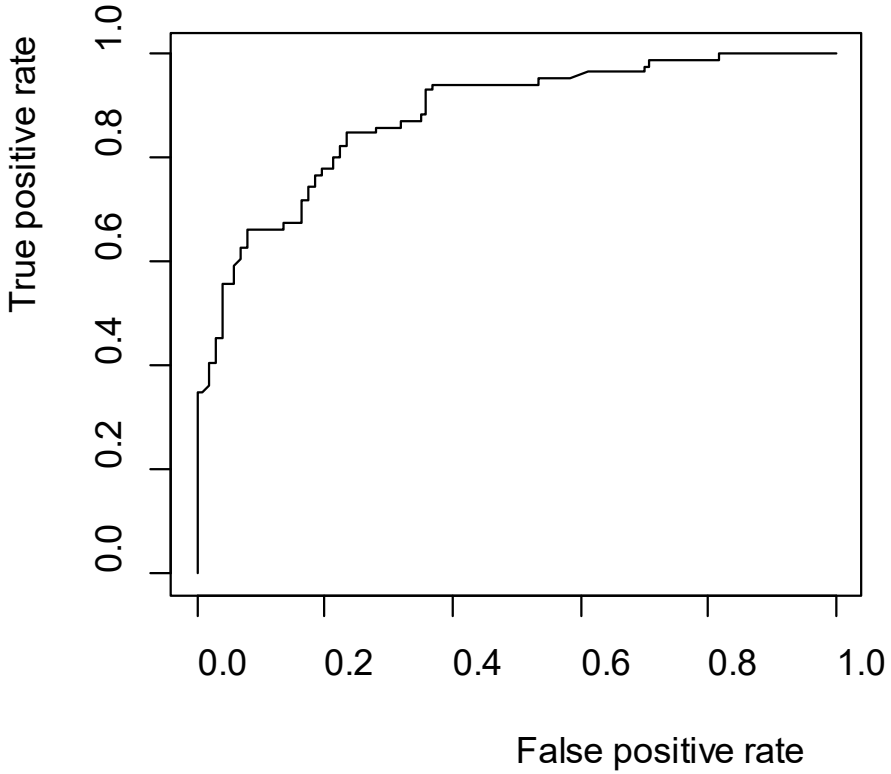
As we did using observed values of CA-125, we can construct an ROC curve using the probabilities of cancer estimated with fitted logistic regression.

How does this one compare to the previous one?

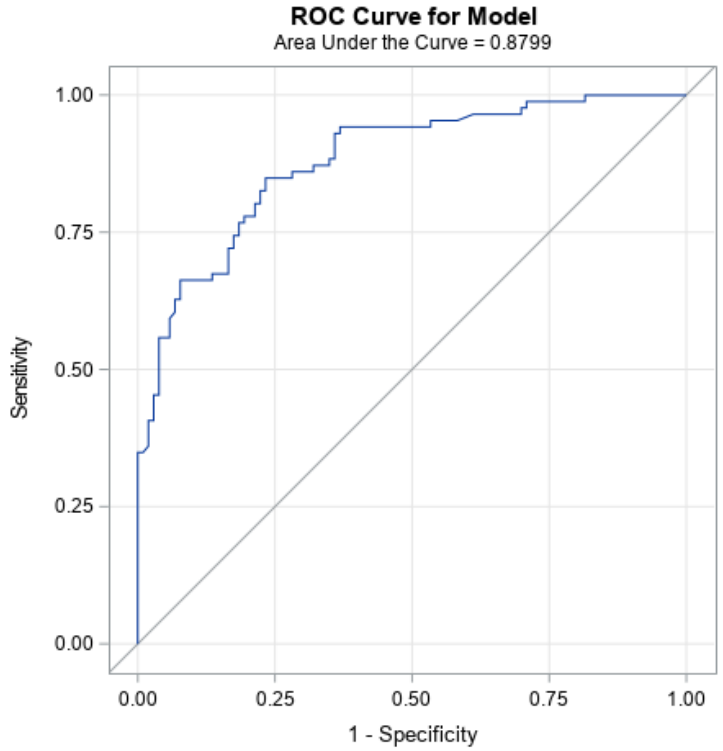


How good is our model, continued?

Observed Data



Logistic Regression Model



They are identical. With only one predictor, the logistic regression model simply maps the observed values onto a probability scale.

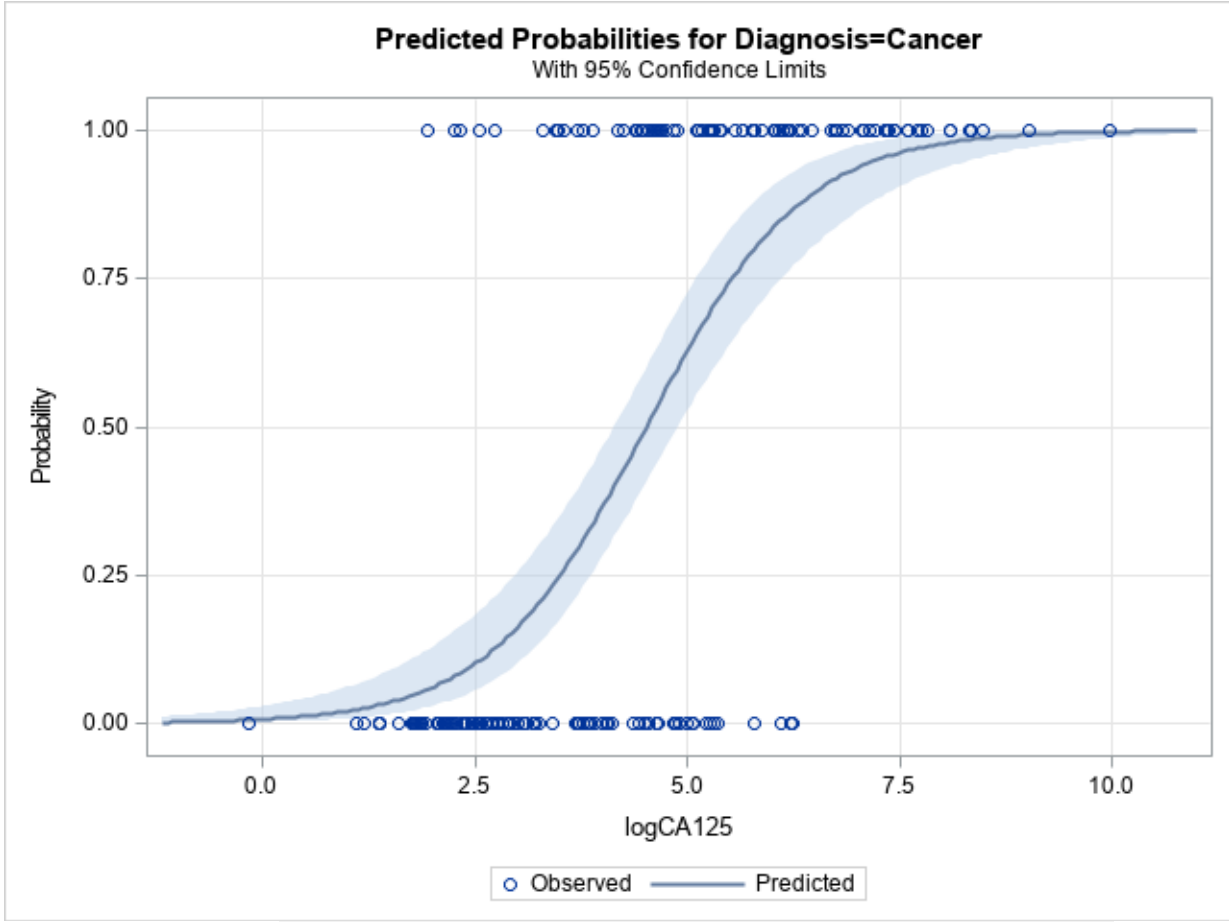
Classification with logistic regression

Results of a logistic regression model can be expressed as the probability of the condition (e.g., cancer)

- This approach retains the most information and is encouraged.

Often though, a binary classification result is desired.

- Can use in concert with predicted probabilities to provide context.

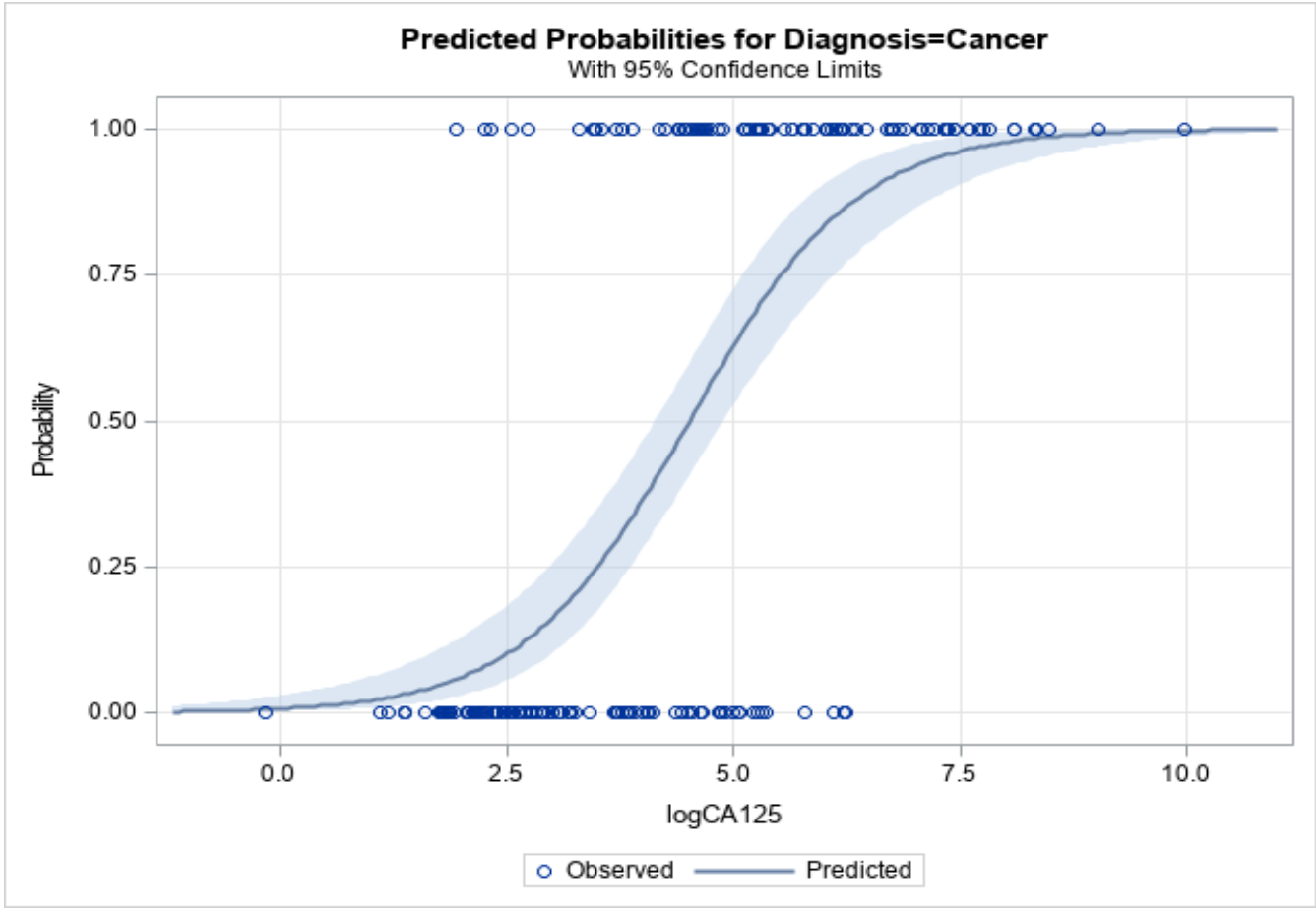


How to choose a classification threshold?

In the context of logistic regression, classification threshold is a probability value above which a patient will be classified as having the condition and below which the patient will be classified as not having the condition or vice versa depending on the relationship.

For this example, what do you think?

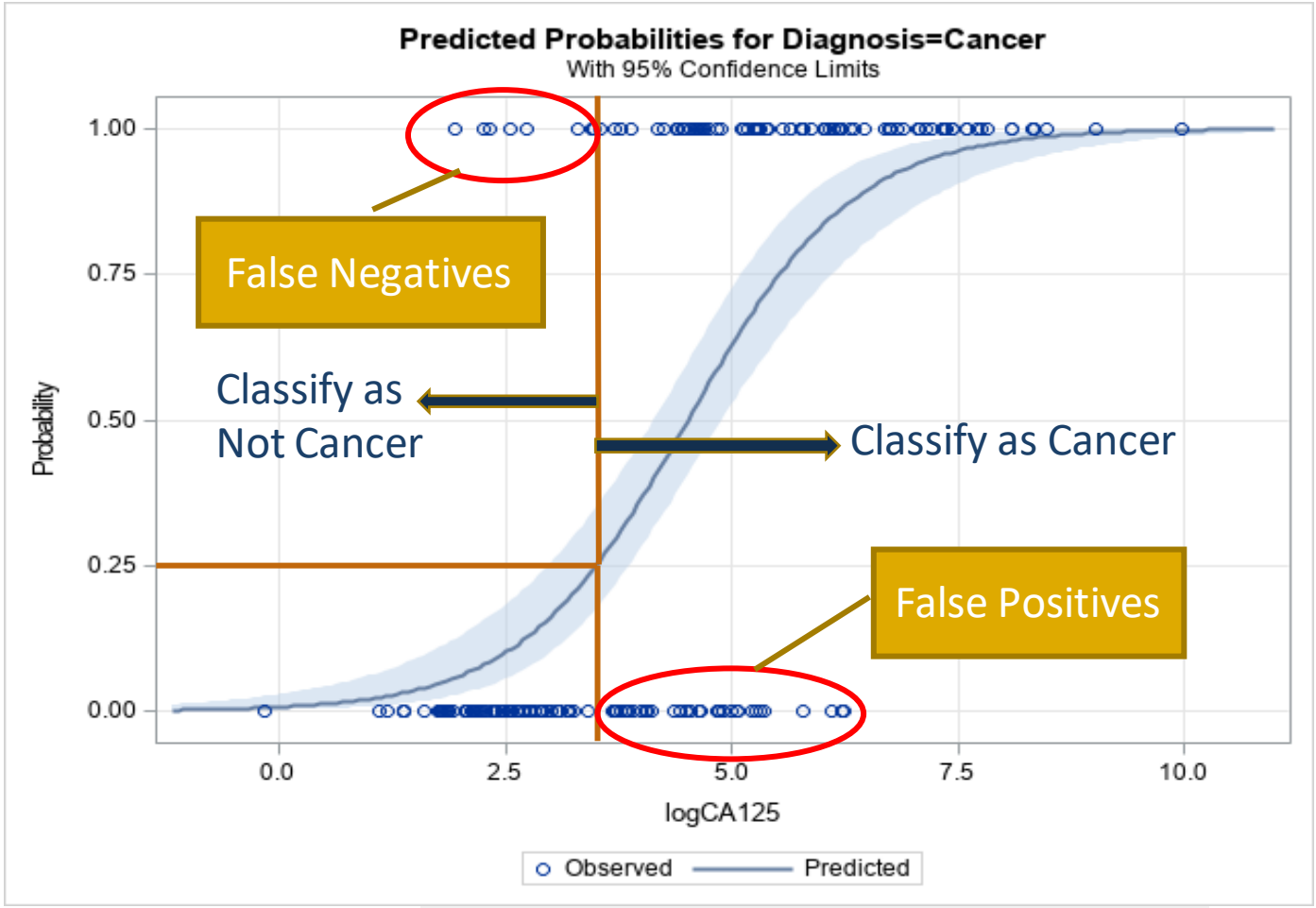
What probability would you suggest for classifying cancer vs. benign?



Considerations for specifying a classification threshold

Suppose we selected 25% probability as our threshold.

- 9 false negatives
- 37 false positives

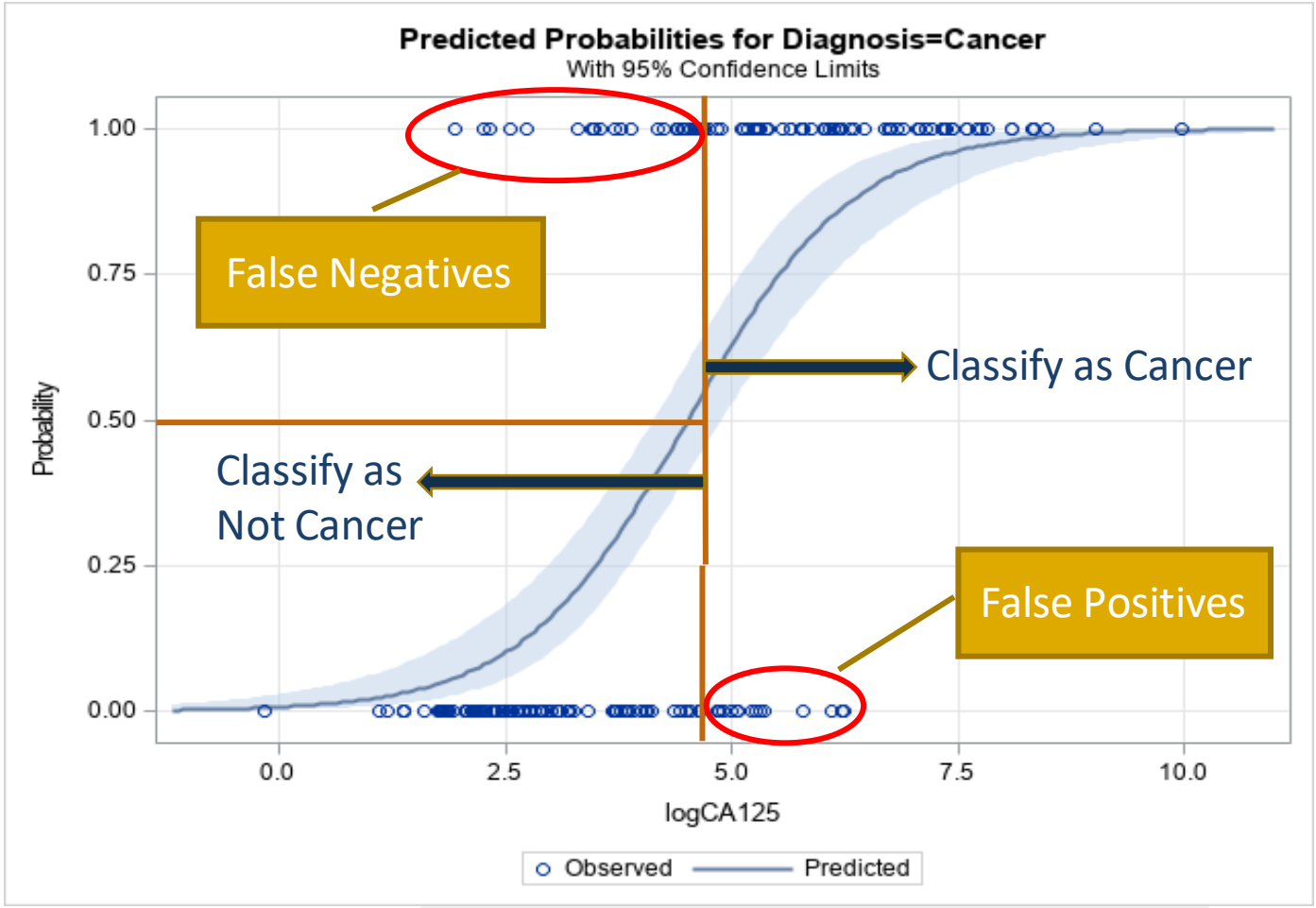


Considerations for specifying a classification threshold

Suppose we selected 50% probability as our threshold.

- 19 false negatives
- 22 false positives

Fewer false positives but many more false negatives which could be fatal.

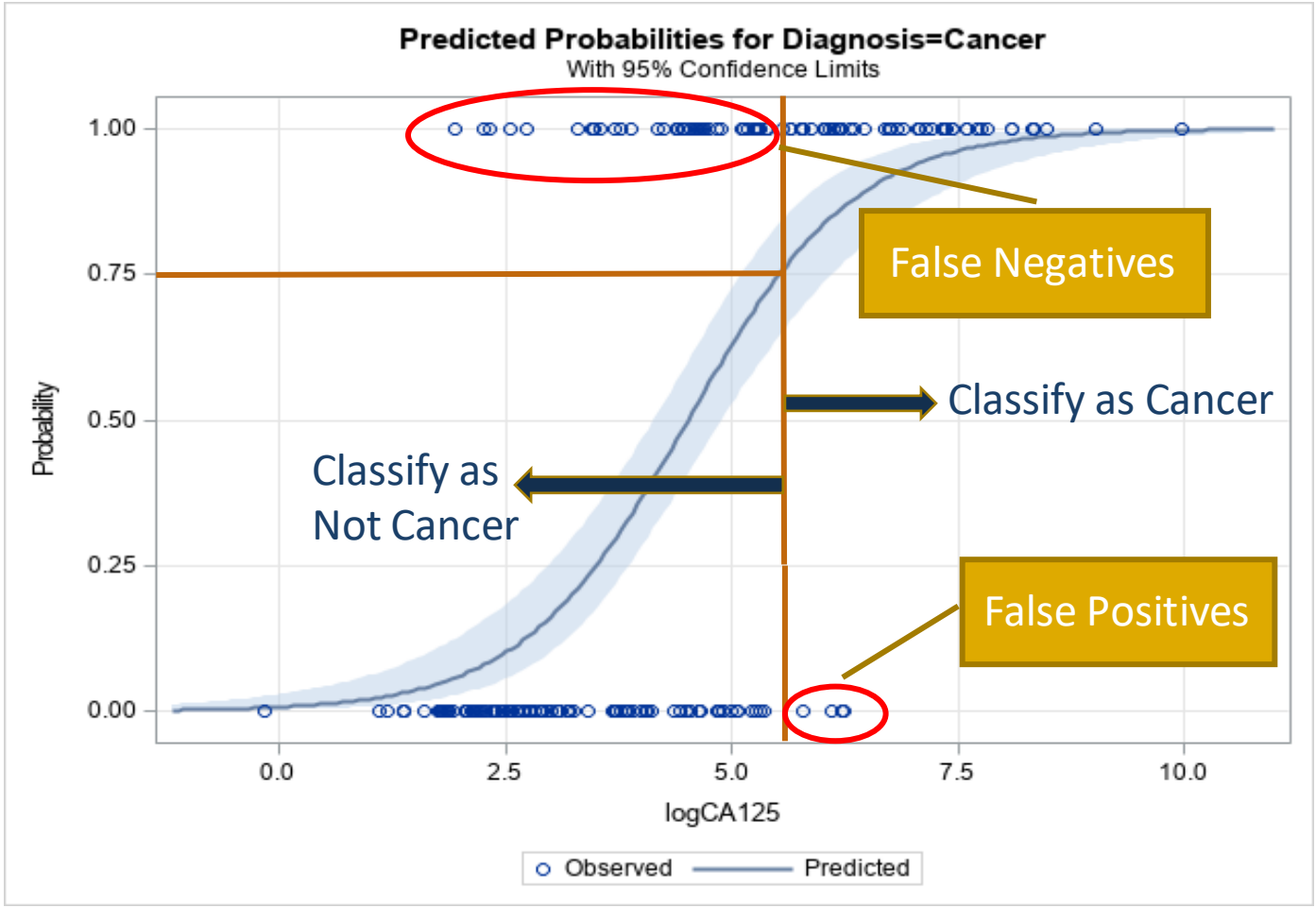


Considerations for specifying a classification threshold

Suppose we selected 75% probability as our threshold.

- 40 false negatives
- 4 false positives

Almost no false positives but many more false negatives.



Trade-off Between Sensitivity and Specificity

Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	86	0	103	0	45.5	100.0	0.0	54.5	.
0.020	86	1	102	0	46.0	100.0	1.0	54.3	0.0
0.040	86	7	96	0	49.2	100.0	6.8	52.7	0.0
0.060	85	19	84	1	55.0	98.8	18.4	49.7	5.0
0.080	83	30	73	3	59.8	96.5	29.1	46.8	9.1
0.100	82	37	66	4	63.0	95.3	35.9	44.6	9.8
0.120	81	46	57	5	67.2	94.2	44.7	41.3	9.8
0.140	81	52	51	5	70.4	94.2	50.5	38.6	8.8
0.160	81	57	46	5	73.0	94.2	55.3	36.2	8.1
0.180	81	61	42	5	75.1	94.2	59.2	34.1	7.6
0.200	81	64	39	5	76.7	94.2	62.1	32.5	7.2
0.220	80	65	38	6	76.7	93.0	63.1	32.2	8.5
0.240	77	66	37	9	75.7	89.5	64.1	32.5	12.0
0.260	76	66	37	10	75.1	88.4	64.1	32.7	13.2
0.280	76	66	37	10	75.1	88.4	64.1	32.7	13.2
0.300	75	69	34	11	76.2	87.2	67.0	31.2	13.8
0.320	74	71	32	12	76.7	86.0	68.9	30.2	14.5
0.340	73	74	29	13	77.8	84.9	71.8	28.4	14.9
0.360	73	75	28	13	78.3	84.9	72.8	27.7	14.8
0.380	73	77	26	13	79.4	84.9	74.8	26.3	14.4
0.400	73	78	25	13	79.9	84.9	75.7	25.5	14.3
0.420	72	79	24	14	79.9	83.7	76.7	25.0	15.1
0.440	71	79	24	15	79.4	82.6	76.7	25.3	16.0
0.460	70	79	24	16	78.8	81.4	76.7	25.5	16.8
0.480	68	80	23	18	78.3	79.1	77.7	25.3	18.4
0.500	67	81	22	19	78.3	77.9	78.6	24.7	19.0
0.520	65	84	19	21	78.8	75.6	81.6	22.6	20.0
0.540	61	85	18	25	77.2	70.9	82.5	22.8	22.7
0.560	59	86	17	27	76.7	68.6	83.5	22.4	23.9
0.580	59	86	17	27	76.7	68.6	83.5	22.4	23.9
0.600	57	88	15	29	76.7	66.3	85.4	20.8	24.8
0.620	57	91	12	29	78.3	66.3	88.3	17.4	24.2
0.640	57	92	11	29	78.8	66.3	89.3	16.2	24.0
0.660	55	95	8	31	79.4	64.0	92.2	12.7	24.6
0.680	54	95	8	32	78.8	62.8	92.2	12.9	25.2
0.700	49	96	7	37	76.7	57.0	93.2	12.5	27.8
0.720	47	98	5	39	76.7	54.7	95.1	9.6	28.5
0.740	46	99	4	40	76.7	53.5	96.1	8.0	28.8
0.760	45	99	4	41	76.2	52.3	96.1	8.2	29.3
0.780	43	99	4	43	75.1	50.0	96.1	8.5	30.3
0.800	39	99	4	47	73.0	45.3	96.1	9.3	32.2
0.820	38	100	3	48	73.0	44.2	97.1	7.3	32.4
0.840	36	100	3	50	72.0	41.9	97.1	7.7	33.3
0.860	31	100	3	55	69.3	36.0	97.1	8.8	35.5
0.880	28	103	0	58	69.3	32.6	100.0	0.0	36.0
0.900	27	103	0	59	68.8	31.4	100.0	0.0	36.4
0.920	23	103	0	63	66.7	26.7	100.0	0.0	38.0
0.940	20	103	0	66	65.1	23.3	100.0	0.0	39.1
0.960	12	103	0	74	60.8	14.0	100.0	0.0	41.8
0.980	5	103	0	81	57.1	5.8	100.0	0.0	44.0

- "ctable" option in model statement yields this table
- Number of correct and incorrect classifications for each probability level
- Sensitivity and specificity for each probability level
- False POS = FP/(FP+TP)
- False NEG = FN/(FN+TN)

  These are 1-PPV and 1-NPV for the prevalence in the data set.

ROC table output provides similar information

VIEWTABLE: Work.Rocout (Receiver Operating Characteristics)

	Probability Level	No. of Correctly Predicted Events	No. of Correctly Predicted Nonevents	No. of Nonevents Predicted as Events	No. of Events Predicted as Nonevents	Sensitivity	1 - Specificity
1	0.9973405612	1	103	0	85	0.011627907	0
2	0.9926047612	2	103	0	84	0.023255814	0
3	0.9867056343	3	103	0	83	0.0348837209	0
4	0.9845261303	4	103	0	82	0.0465116279	0
5	0.9842980604	5	103	0	81	0.0581395349	0
6	0.9797870453	6	103	0	80	0.0697674419	0
7	0.979720279	7	103	0	79	0.0813953488	0
8	0.9729536122	8	103	0	78	0.0930232558	0
9	0.9712772322	9	103	0	77	0.1046511628	0
10	0.9698687886	10	103	0	76	0.1162790698	0
11	0.9661419337	11	103	0	75	0.1279069767	0
12	0.9661064287	12	103	0	74	0.1395348837	0
13	0.9598306697	13	103	0	73	0.1511627907	0
14	0.9594581406	14	103	0	72	0.1627906977	0
15	0.9569359178	15	103	0	71	0.1744186047	0
16	0.9563949929	16	103	0	70	0.1860465116	0
17	0.9552821903	17	103	0	69	0.1976744186	0
18	0.9494563156	18	103	0	68	0.2093023256	0

- “outroc” option in model statement yields this table
- Every point for ROC curve

What is the "optimal" a classification threshold?

It depends on relative "cost" of false positives and false negatives.

What are the "costs" of a false negative?

1. Missed cancer diagnosis
2. Missed sepsis diagnosis
3. Failing to identify patient no-show

What are the "costs" of a false positive?

1. Incorrect cancer diagnosis
   - Unnecessary procedures, patient anxiety
2. False sepsis alert
   - Alert fatigue, unnecessary tests
3. Incorrect prediction of patient no-show

No one optimal answer.

Some options for threshold identification

1. Maximize Youden' Index

   Youden's Index = Sensitivity + Specificity – 1

2. Closest to [0,1] point of ROC curve.

   Minimize ER $\quad \mathrm{ER}(c) = \left(\sqrt{(1 - \mathrm{Se}(c))^2 + (1 - \mathrm{Sp}(c))^2} \right)$

3. Maximize Concordance Probability

   CP = Sensitivity*Specificity

4. Maximize sensitivity at lowest acceptable specificity

Threshold identification for CA-125

Youden's Index, Distance from [0,1], and Concordance Method all identify 41.3% as "optimal" cut-off

Sensitivity = 84.9% Specificity = 76.7%

Suppose we are more concerned about sensitivity but want specificity to be at least 70%.

For these criteria, the optimal cut-off is 33.0%.

Sensitivity = 86.0% Specificity = 70.0%

	Predicted Benign	Predicted Cancer
True Benign	79	24
True Cancer	13	73

	Predicted Benign	Predicted Cancer
True Benign	73	30
True Cancer	12	74

Minimum specificity approach picks up 1 more cancer case but at the expense of 6 more false positives.

Summary

- Fit logistic regression model to relate probability of cancer to CA-125 levels
- Quantified model's overall performance for classification
- Identified some alternative classification thresholds and considered the trade-offs associated with these thresholds

Questions?

Multiple Logistic Regresion

What if you want to consider more than 1 predictor?

- With only one predictor, fitting a logistic regression isn't necessary to identify a cut-off.
  - However, logistic regression could still be helpful by mapping observed values to probabilities of the outcome about which we have some intuition.
- With more than one predictor, a model is necessary in order to consider the compositive effects of the predictors on the risk of the outcome
- Multiple logistic regression integrates the effect of multiple predictors on the probability of the outcome

Fit multiple logistic regression using Proc Logistic

- Suppose we want to include age in our classification model
- Fit a logistic regression using Proc Logistic modeling cancer outcome versus CA-125 and age
- Estimate age-adjusted probability of cancer based on CA-125

```
proc logistic data=oc plots=EFFECT;
     model diagnosis(event='Cancer') = logCA125 age / outroc=rocout ctable;
     output out=estimated predicted=estprob l=lower95 u=upper95;
run;
```

Age added as predictor

Characterize performance in same way as before

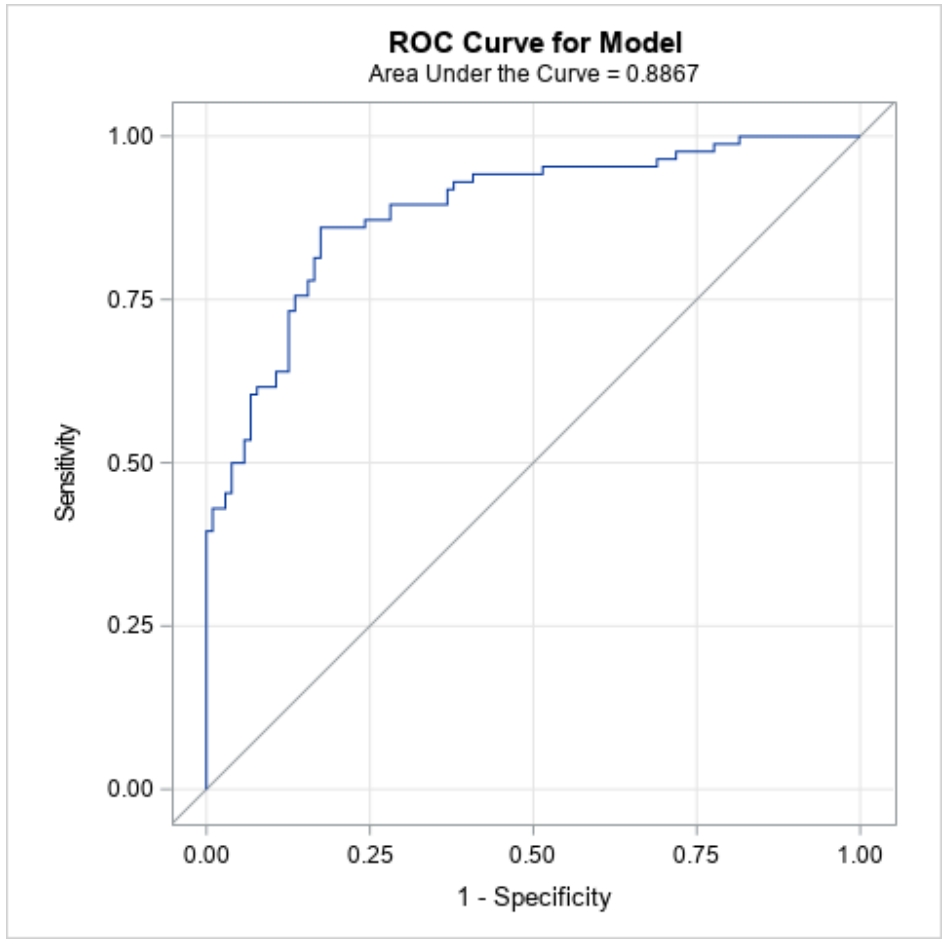
Inclusion of age slightly strengthened relationship with CA-125

Because of age-matching, age effect isn't expected

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-7.1162	1.4100	25.4716	<.0001
logCA125	1	1.1479	0.1617	50.3822	<.0001
Age	1	0.0323	0.0169	3.6535	0.0560

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
logCA125	3.152	2.295	4.327
Age	1.033	0.999	1.068

Very small increase in AUC



Threshold identification for age-adjusted CA-125 model

Youden's Index, Distance from [0,1], and Concordance Method

Age-Adjusted Model

All methods identify 45.0% as "optimal"

Sensitivity = 86.0% Specificity = 82.5%

	Predicted Benign	Predicted Cancer
True Benign	85	18
True Cancer	12	74

Not Age-Adjusted Model

Sensitivity = 84.9% Specificity = 76.7%

	Predicted Benign	Predicted Cancer
True Benign	79	24
True Cancer	13	73

Age-adjusted model reduced the false positives by 6 and false negatives by 1.

Multiple Logistic Regression Classification

- With one predictor, a probability from logistic regression can be translated back to a CA-125 value
  - One-to-one correspondence between CA-125 and cancer probability
- With multiple predictors, no longer have this
- With multiple logistic regression, many predictors are taken into account to estimate probability of cancer
  - For classification, would need to calculate this probability

To summarize

- Logistic regression can be used to estimate the probability of a binary outcome based on one or more predictors
- Classification thresholds can be selected using these probabilities
  - Several methods are available for choosing a threshold
- Selecting a classification threshold entails balancing the relative costs of false positives and false negatives.
  - Costs are context dependent
- A statistically significant difference between cases and controls does not guarantee acceptable discriminatory performance for clinical use

Cautions on developing classification models: Need for training and test sets

If your objective is to develop a clinical classification model, it is imperative to have completely separate training and test sets.

- Conduct ALL model development steps using ONLY the training set
- Build a model using a training set and evaluate performance on test sets
  - Models perform better on the data used to build them than on independent data
- Models should be validated on a third independent data set reflective of world conditions (e.g., event prevalence, data availability and quality, etc.)

Cautions on developing classification models: Predictor selection

- Carefully consider predictors to include in model development
- Smaller of the number of events and non-events drives maximum number of predictors that can be reliably estimated
- Rough guideline is 10-20 events per predictor
- Avoid including highly correlated predictors
- Penalized approaches (LASSO, Ridge, Elastic Net) can be valuable variable selection methods in logistic regression context

Help is available

- CTSC and Cancer Center Biostatistics Office Hours
  - Every Tuesday from 12 – 2:00 currently via WebEx
  - Sign-up through the CTSC Biostatistics Website
- EHS Biostatistics Office Hours
  - Upon request
- Request Biostatistics Consultations
  - CTSC
  - MIND IDDRC
  - Cancer Center Shared Resource
  - EHS Center

References

- Bewick et al. 2004. Statistics review 13: Receiver operating characteristic curves. Critical Care 8(6): 508-512.
- Bewick et al. 2005. Statistics review 14: Logistic regression. Critical Care 9(1): 112-118.
- Anderson et al. 2003. Understanding logistic regression analysis in clinical reports: An introduction. Ann Thorac Surg 2003(75): 753-757.
- Bagley et al. 2001. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. J. Clinical Epidemiology 54:979-985
- Leisman et al. 2021 Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals. Critical Care Medicine 48(5): 624-633.