Finding Cutoff For Ovarian Cancer Using ROC

Yuning Hu

2023-11-13

Glance of data:

Data is from the `OC_Marker.csv` available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9394434/.

```
## [1] "Age"       "Menopause" "CA19-9"    "CA72-4"    "AFP"       "CA125"
## [7] "HE4"       "CEA"       "TYPE"
```

CA19-9: most commonly used for ovarian cancer.

CA72-4: most commonly used for gastric cancer.

AFP: most commonly used for liver, ovary or testicle cancer.

CA125: related to ovarian cancer.

HE4: recently discovered, most commonly used for epithelial ovarian cancer.

CEA: related to cancer.

TYPE: indicate whether the patient have ovarian cancer.

```
##       Age         Menopause    CA19-9            CA72-4           AFP
##  Min.   :15.00    0:230     Min.   :   0.60   Min.   :  0.20   Min.   :   0.61
##  1st Qu.:33.00    1:119     1st Qu.:   8.34   1st Qu.: 10.17   1st Qu.:   1.67
##  Median :45.00              Median :  15.11   Median : 10.17   Median :   2.42
##  Mean   :45.05              Mean   :  45.87   Mean   : 10.17   Mean   :  11.39
##  3rd Qu.:57.00              3rd Qu.:  34.22   3rd Qu.: 10.17   3rd Qu.:   3.77
##  Max.   :83.00              Max.   :1000.00   Max.   :158.50   Max.   :1210.00
##     CA125            HE4             CEA           TYPE
##  Min.   :   3.75   Min.   :  16.71   Min.   :  0.200   0:171
##  1st Qu.:  20.32   1st Qu.:  42.58   1st Qu.:  0.850   1:178
##  Median :  49.28   Median :  54.94   Median :  1.400
##  Mean   : 345.78   Mean   : 183.95   Mean   :  3.309
##  3rd Qu.: 255.82   3rd Qu.: 166.20   3rd Qu.:  2.300
##  Max.   :5000.00   Max.   :3537.60   Max.   :138.800
```
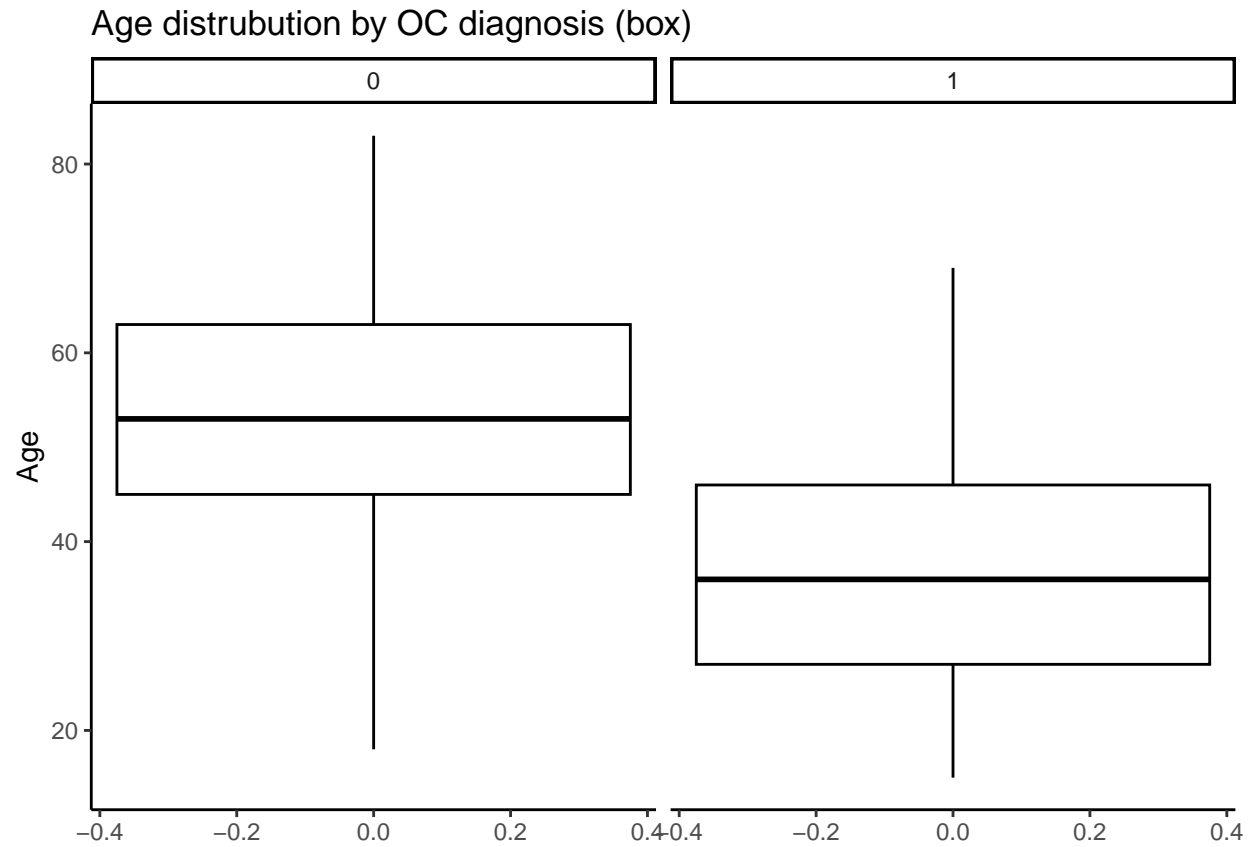
Number of patients who have / not have ovarian cancer.

```
## [1] 178
```
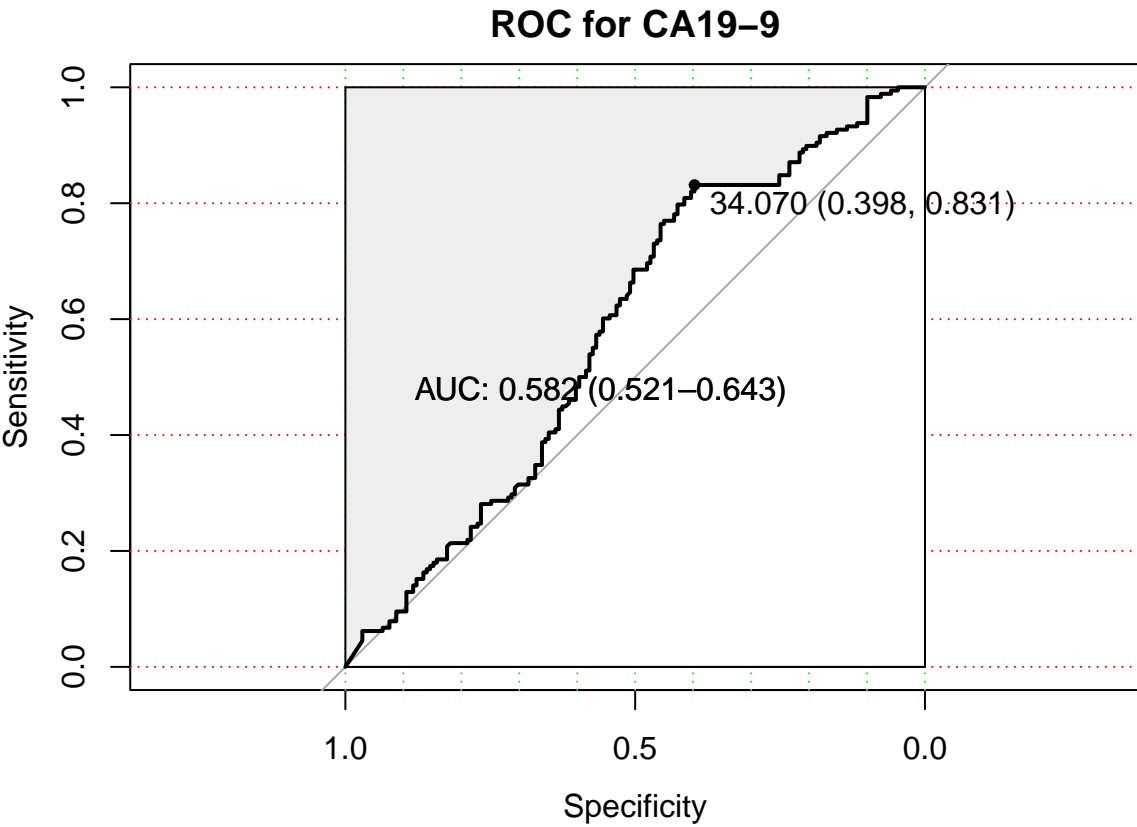
```
## [1] 171
```

Age distribution by diagnosis in this dataset:

Age distrubution by OC diagnosis (hist)
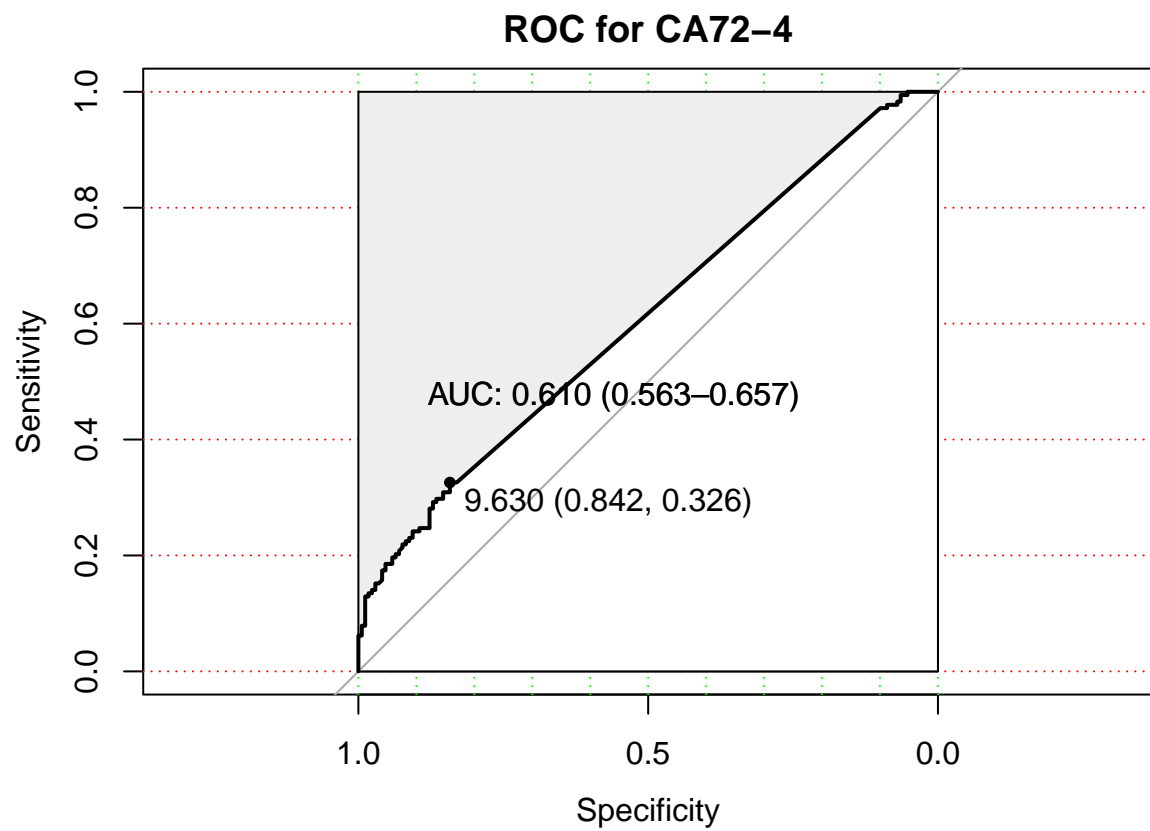
Age distrubution by OC diagnosis (box)
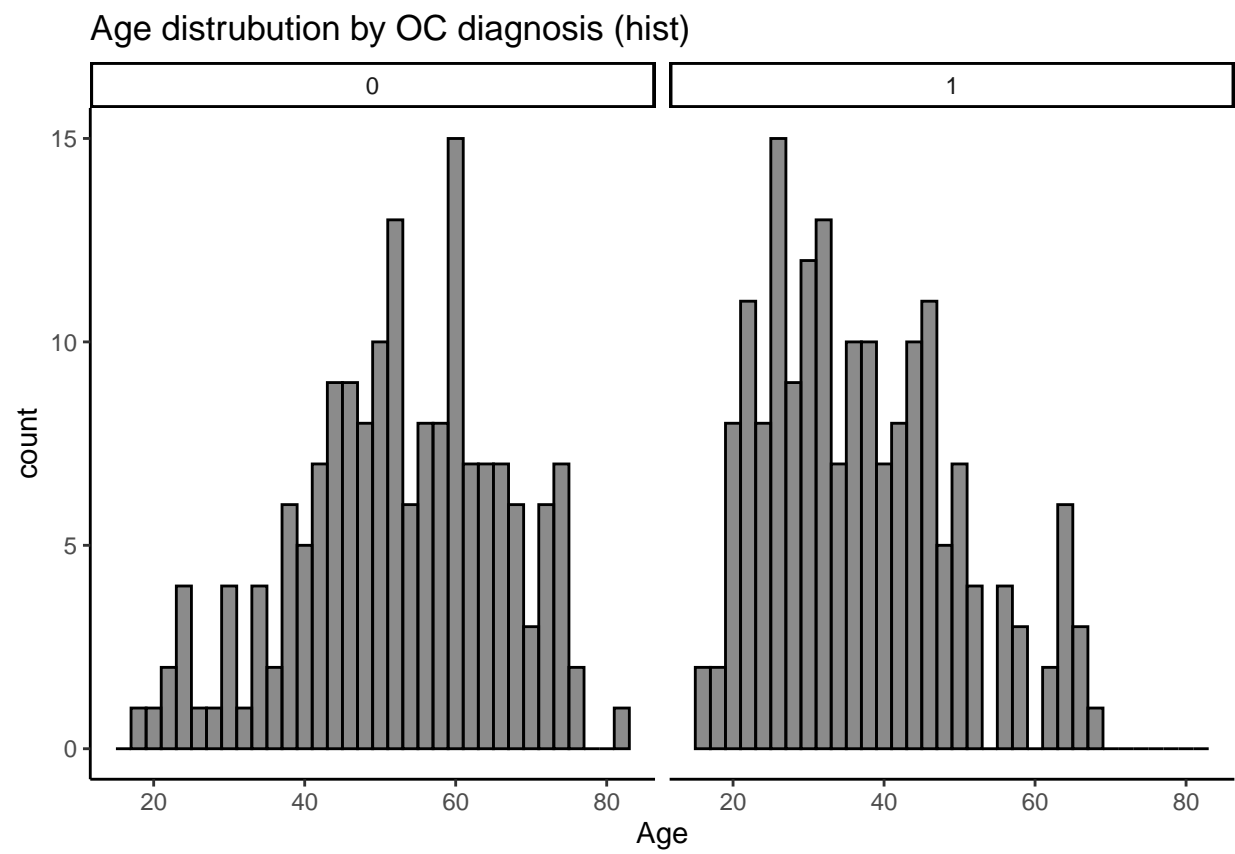
We cannot see a clear relationship between age and ovarian cancer. Younger people (who are not likely to encounter a menopause) may also have ovarian cancer.
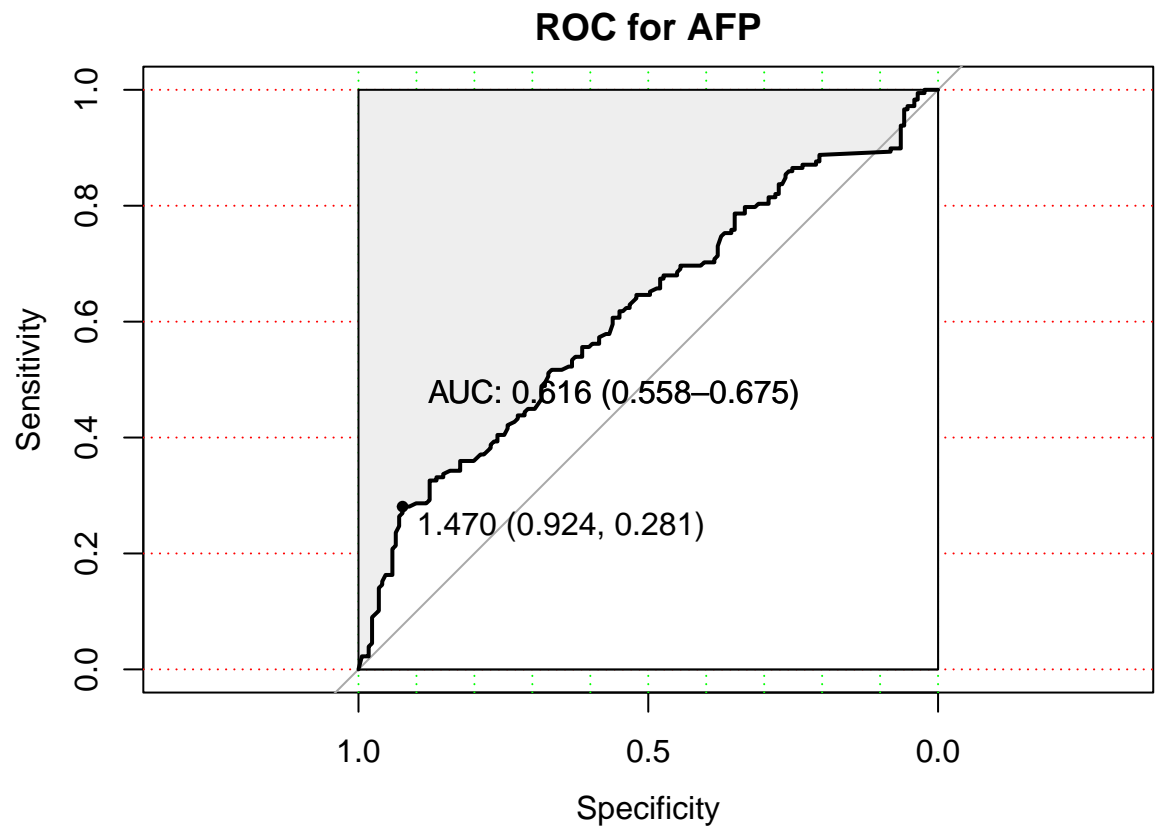
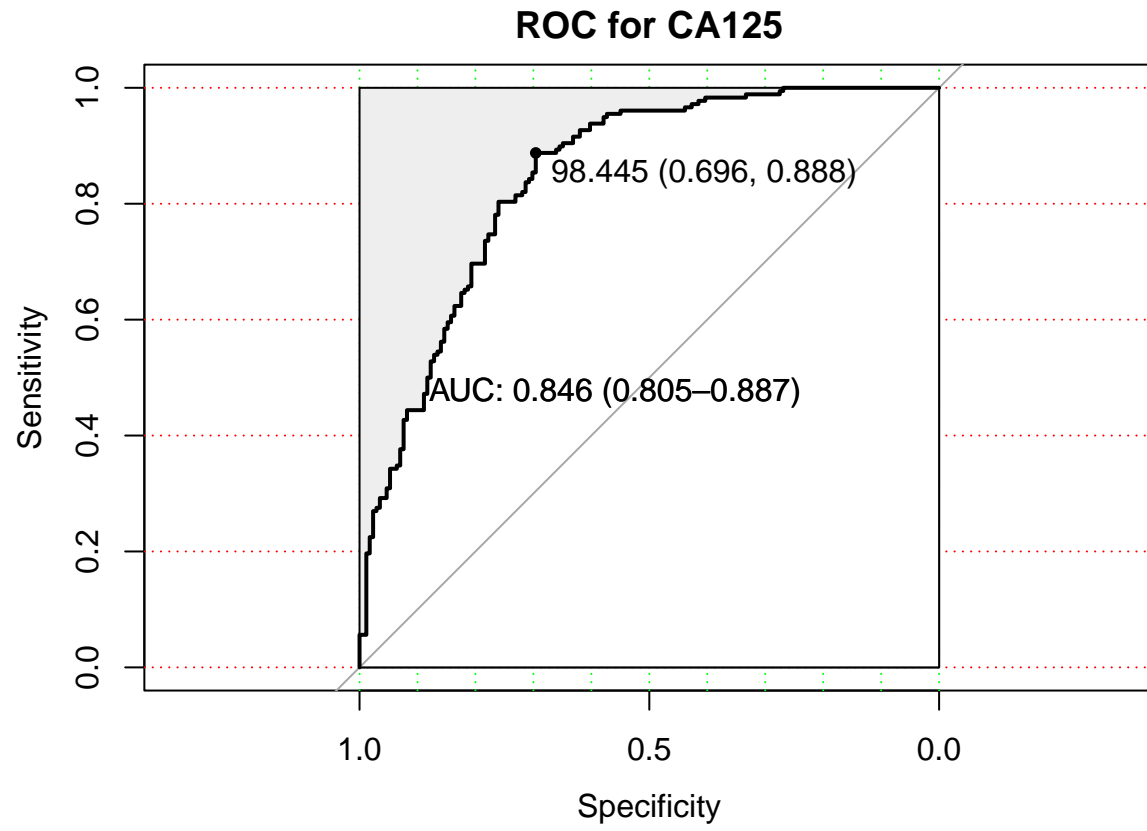**Visualize the predictors' sensitivity and specificity by ROC:**

ROC for CA19−9

ROC curve for CA19-9:
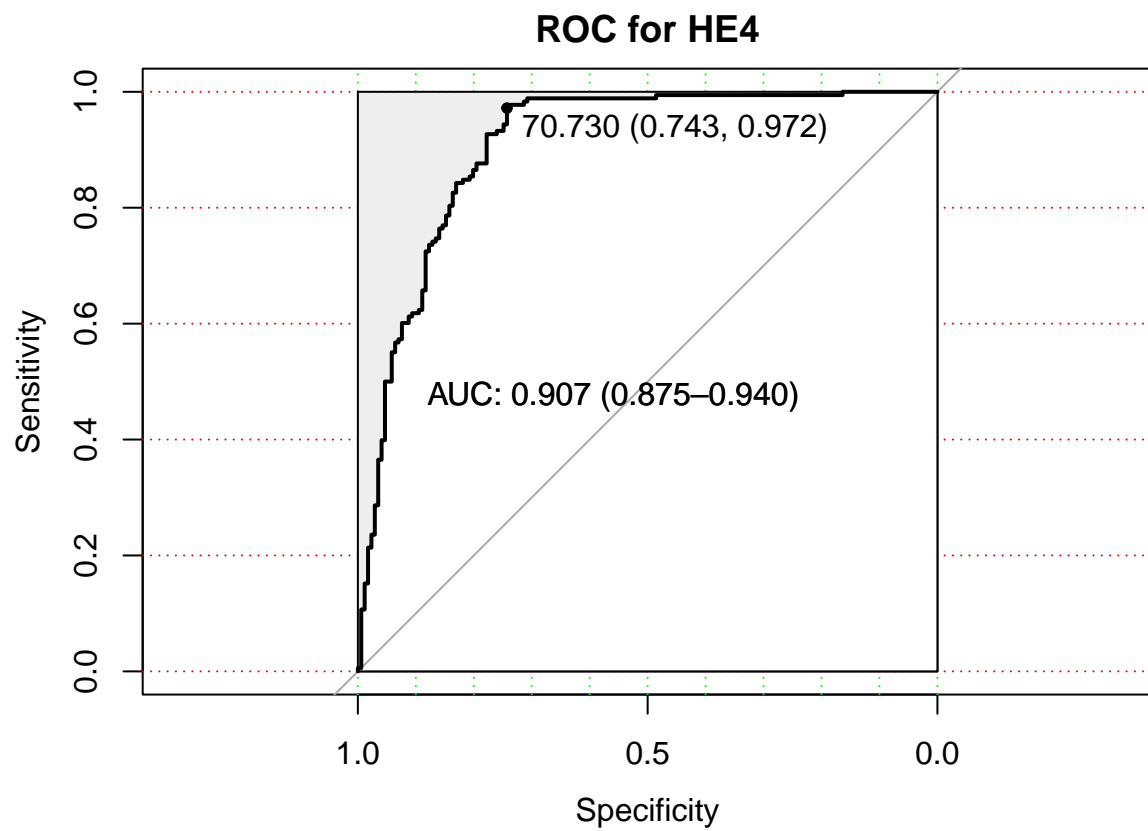
ROC for CA72−4

AUC: 0.610 (0.563–0.657)

9.630 (0.842, 0.326)

ROC curve for CA72-4:

ROC for AFP

AUC: 0.616 (0.558–0.675)

1.470 (0.924, 0.281)

Sensitivity

Specificity

ROC curve for AFP:

ROC for CA125

ROC curve for CA125:

ROC for HE4

70.730 (0.743, 0.972)

AUC: 0.907 (0.875–0.940)

Sensitivity

Specificity

ROC curve for HE4:

ROC for CEA

2.625 (0.357, 0.921)

AUC: 0.606 (0.545–0.666)
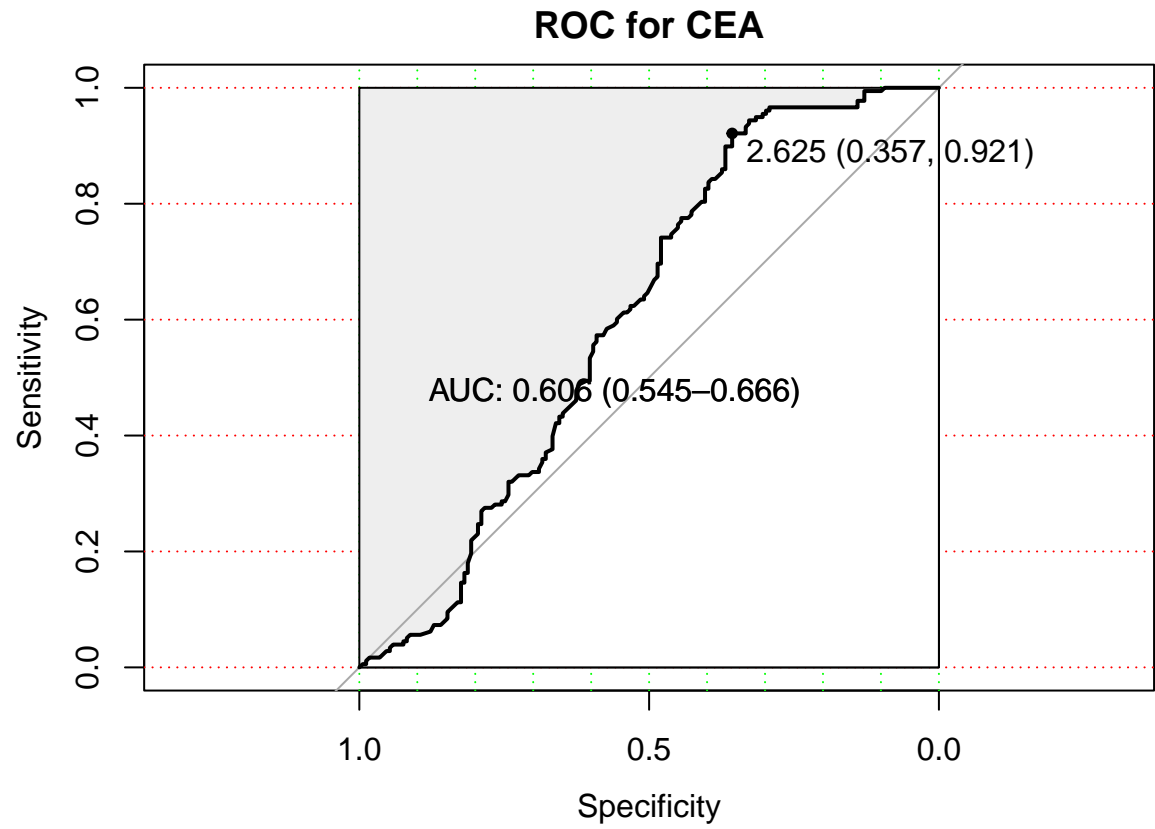
ROC curve for CEA:

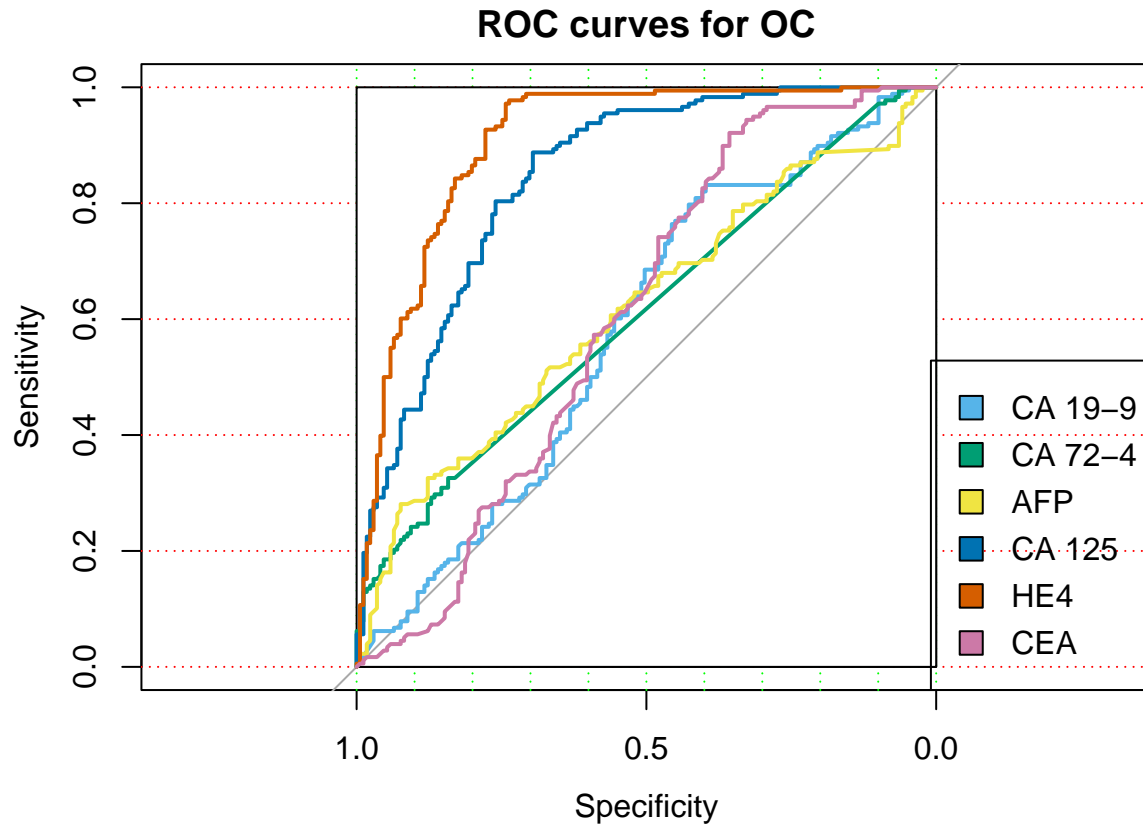**Comparing ROC curves with ROC of HE4:**

ROC curves for OC

We can see that HE4 has the best AUC, implying it is on average the best predictor of ovarian cancer.

Comparing the AUCs with the delong method:

(checking if HE4 is obviously better than others)

```
##
##  DeLong's test for two correlated ROC curves
##
## data:  roche4 and roc19_9
## Z = 9.4932, p-value < 2.2e-16
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  0.2580743 0.3923627
## sample estimates:
## AUC of roc1 AUC of roc2
##   0.9072212   0.5820028


##
##  DeLong's test for two correlated ROC curves
##
## data:  roche4 and roc72_4
## Z = 10.587, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  0.2422431 0.3523097
## sample estimates:
## AUC of roc1 AUC of roc2
##   0.9072212   0.6099448


##
##  DeLong's test for two correlated ROC curves
##
## data:  roche4 and rocafp
## Z = 8.3445, p-value < 2.2e-16
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  0.2225505 0.3591895
## sample estimates:
## AUC of roc1 AUC of roc2
##   0.9072212   0.6163513


##
##  DeLong's test for two correlated ROC curves
##
## data:  roche4 and roc125
## Z = 3.0874, p-value = 0.002019
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  0.02230291 0.09984703
## sample estimates:
## AUC of roc1 AUC of roc2
##   0.9072212   0.8461463


##
##  DeLong's test for two correlated ROC curves
##
## data:  roche4 and roccea
## Z = 8.8491, p-value < 2.2e-16
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  0.2348604 0.3684972
## sample estimates:
## AUC of roc1 AUC of roc2
##   0.9072212   0.6055424
```

We can see that HE4 is significantly better than all other ROC curves based on their AUC. CA125 is the second best. Since HE4 is best but expensive, we would also look at the second best, CA125 for its accessibility.

Finding the cutoff:

Traditional approach - penalize sensitivity and specificity equally.

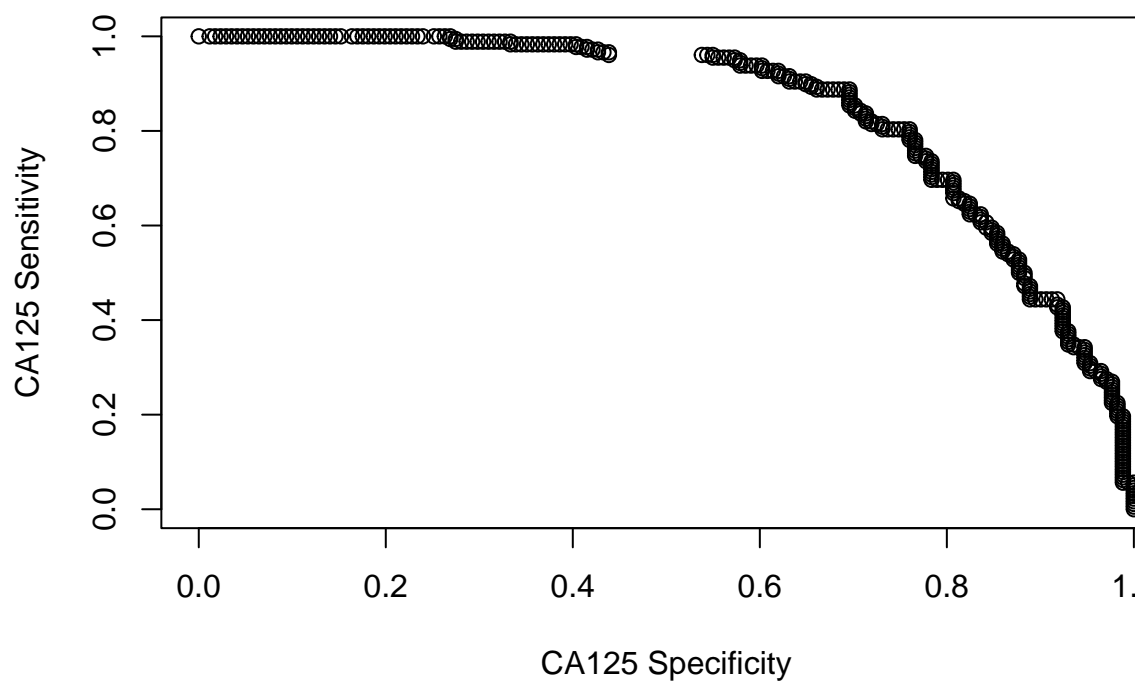Adjusted approach - penalize sensitivity with 1.5 pentalty of specificity.
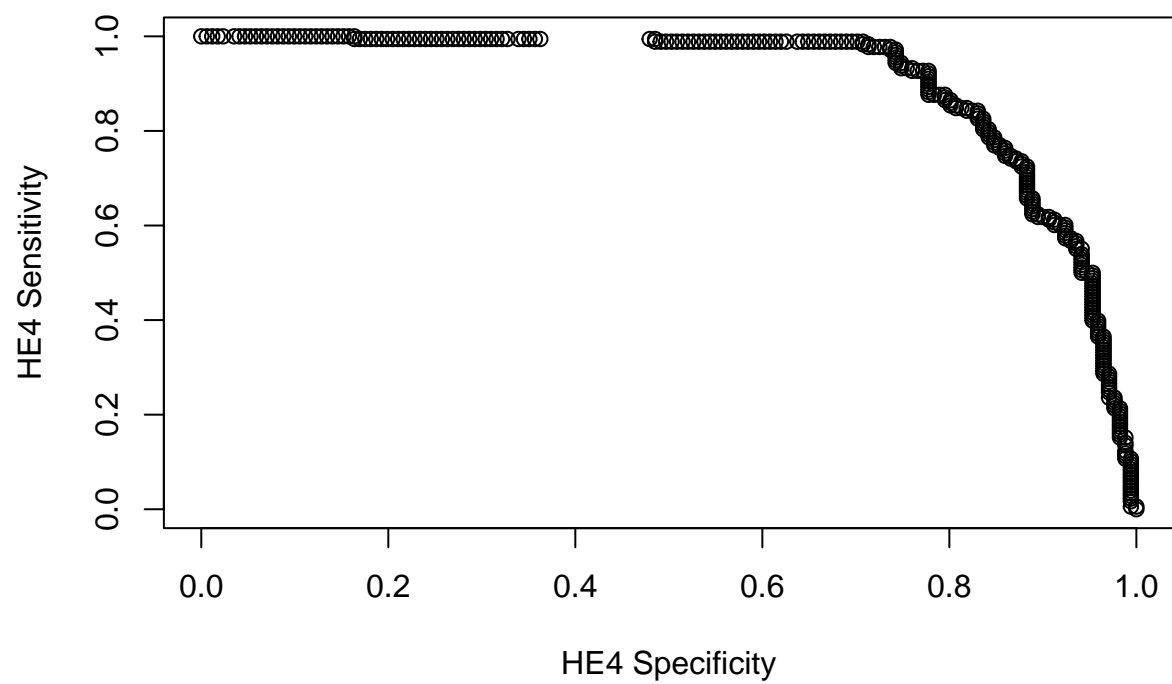
Find the cutoff for HE4

```
## [1] 70.73
```

```
## [1] 71.44
```

(break in plot due to break in data)

Find the cutoff for CA125:
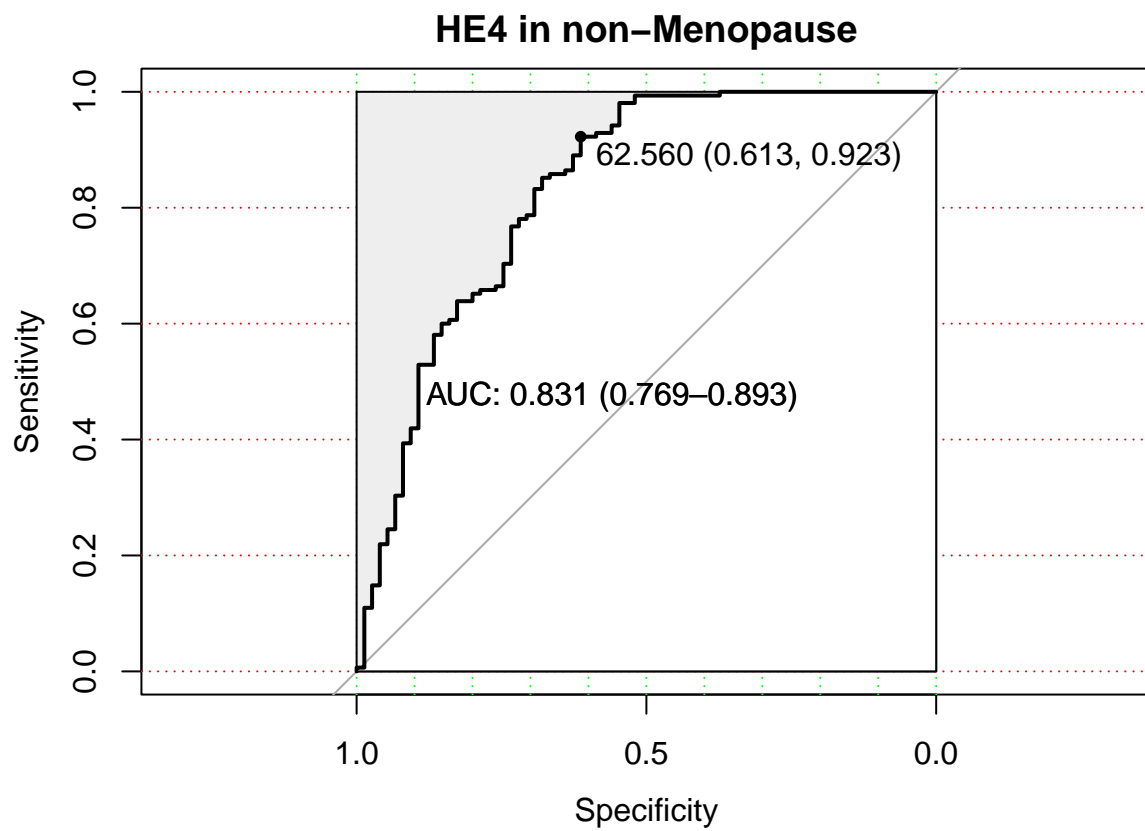
```
## [1] 98.445
```

```
## [1] 98.445
```
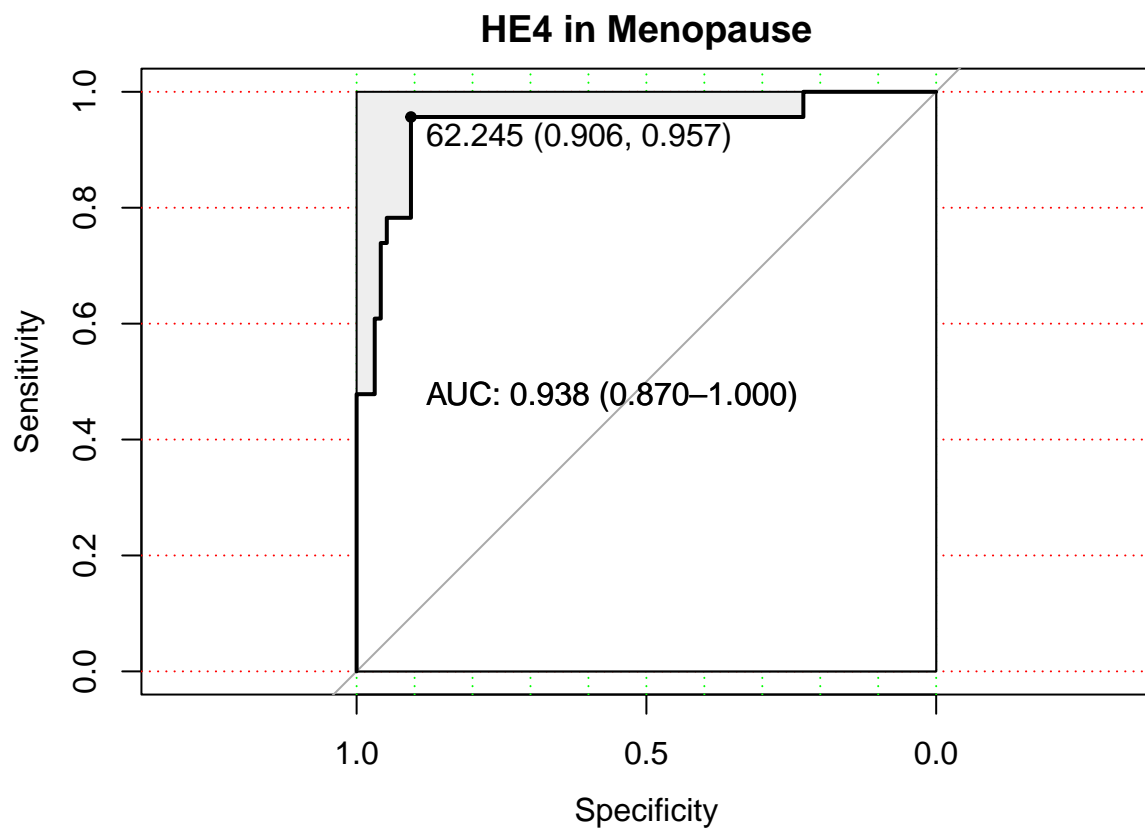
```
## [1] 0.5786517
```

```
## [1] 0.8538012
```

The cutoff for CA19-9: 73, adjusted: 73.

Further:

HE4 in Menopause and non-Menopause populations:

HE4 in non–Menopause

62.560 (0.613, 0.923)

AUC: 0.831 (0.769–0.893)

HE4 in Menopause

AUC: 0.938 (0.870–1.000)

62.245 (0.906, 0.957)

CA125 in Menopause and non-Menopause populations.

CA125 in non−Menopause

98.445 (0.613, 0.877)

AUC: 0.777 (0.708−0.846)

CA125 in Menopause

22.250 (0.938, 0.913)

AUC: 0.956 (0.910–1.000)

Sensitivity

Specificity