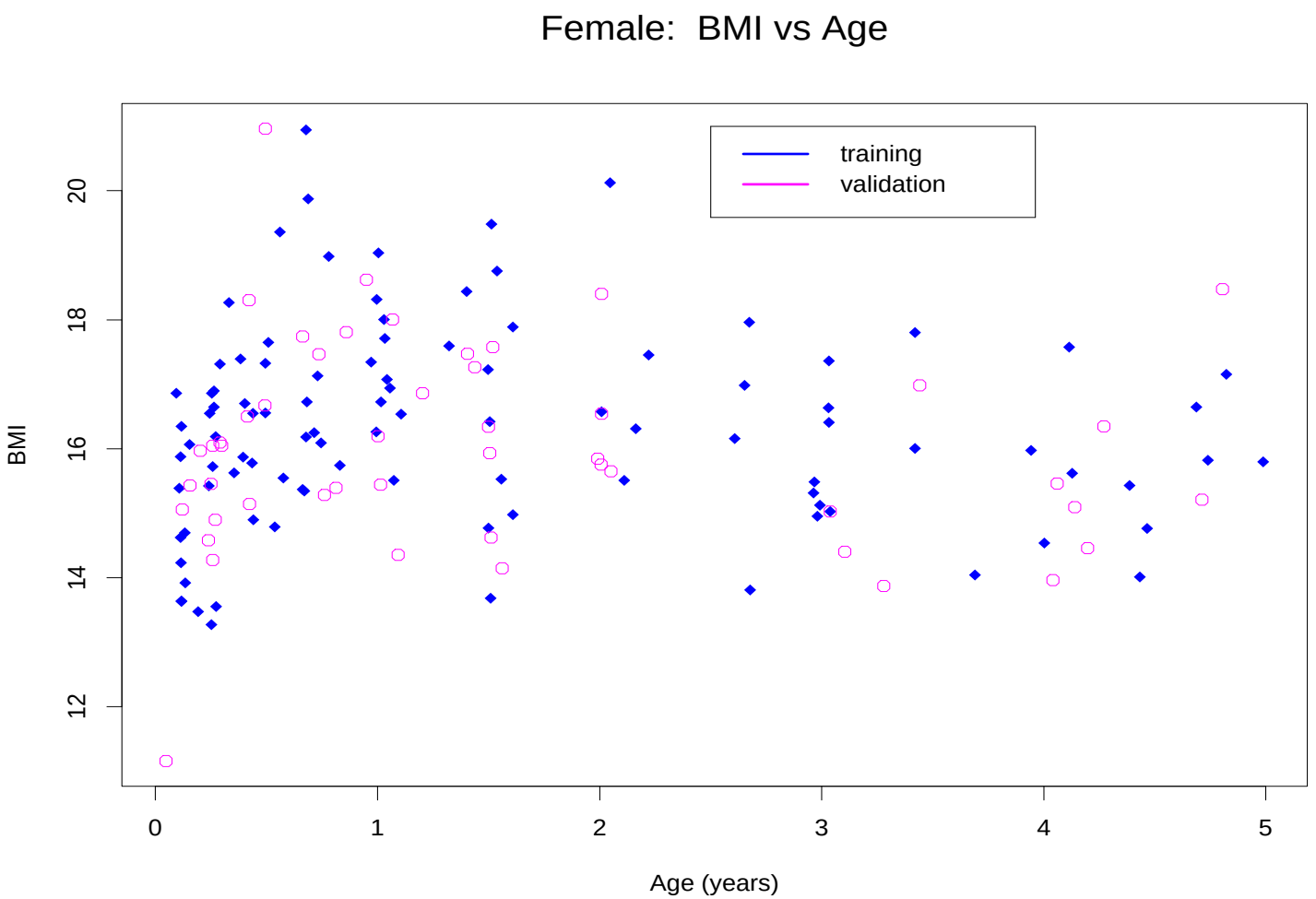
Model Choice & Checking

- Predictive Model Development & Assessment.
  - Defining error of prediction
  - Bias versus variance
  - AIC and BIC as criteria
  - Internal vs. External validation
- Accuracy Ideas for Survival Data.
  - ROC curves?
  - C index?
  - Extension of R^2?

Example: Continuous Response

- Diagnostic classification often relies on the comparison of an observed measurement with "normal" or reference values.
  - ▷ percent-predicted for FEV1
  - ▷ height & weight growth charts for children
- In such situations we desire a good prediction for an outcome, and typical ranges (e.g. 5th and 95th percentiles).
- As of (1999) there were no published reference percentiles for BMI (body mass index) for children between 0 and 36 months of age.
- The following data are a sample of data used by Heagerty & Pepe (1999) to create reference percentiles.
- The following data come from Group Health Cooperative based on subjects born between 1965 and 1971.

Example: BMI Data



Prediction Error

- One Statistical View:
  - **Objective**: find a curve that is a function of age, $f(\text{age})$, that can be used as a prediction function for the mean BMI at each age, $E(Y \mid \text{age})$ where Y denotes BMI.
  - Specifically, suppose that we will use splines to estimate the function. Let p denote the number of splines that we are using for the function. **Q**: what value of p should we pick?
  - **Define**: a curve will be judged in terms of its ability to create predictions that are "close" to real observations. We will choose the curve that is best in terms of minimizing our estimate of its error of prediction.
- Error In order to make progress we will need to choose some method for estimating the "error" of prediction.

Prediction Error

- Measuring the Error
  - For continuous measurements we often choose to measure the distance from the prediction to the observation, and call this the error:

$$\texttt{error}_i = \left[Y_i - \widehat{f}(\text{age}_i)\right]^2$$

- Sources of Error
  - **variance** in observation Y_i.
  - **variance** in prediction function estimate $\widehat{f}(\text{age}_i)$.
  - **bias** in prediction function choice $f(\text{age}_i)$.

Prediction Error

- **Bias** :
  - $f(\text{age}_i)$ = our chosen curve if estimated with a huge sample. Shape is determined by what model we allow (e.g. linear, quadratic, splines).
  - $\mu(\text{age}_i)$ = true mean of Y as a function of age.
  - Unless our model for $f(x)$ is correct these two curves will be different.

$$\texttt{bias}_i = |\mu(\text{age}_i) - f(\text{age}_i)|$$

Decomposing Prediction Error

- Expected Prediction Error :
  - We can calculate the expected value of the prediction error:

$$
\begin{aligned}
E\left\{\texttt{error}_i\right\} &= E\left\{\left[Y_i - \widehat{f}(\texttt{age}_i)\right]^2\right\} \\
&= E\left\{\left[\, Y_i - \mu(\texttt{age}_i) \;\; + \right.\right. \\
&\qquad \mu(\texttt{age}_i) - f(\texttt{age}_i) \;\; + \\
&\qquad \left.\left. f(\texttt{age}_i) - \widehat{f}(\texttt{age}_i)\right]^2\right\}
\end{aligned}
$$

Decomposing Prediction Error

- Expected Prediction Error :
  - If Y_i is "new" data, not used to create the estimated prediction function, $\widehat{f}(x)$ then:

$$E\{\texttt{error}_i\} = E\left\{\left[Y_i - \widehat{f}(\texttt{age}_i)\right]^2\right\}$$

$$= \textbf{variance}(Y_i) +$$
$$\textbf{bias}^2[f(\texttt{age}_i)] +$$
$$\textbf{variance}[\widehat{f}(\texttt{age}_i)]$$

Balancing Bias and Variance

- Objective: our task is to create a "good" estimated prediction function, $\widehat{f}(\text{age}_i)$, based on the available data (so-called "training data").
- Issue:
  - ▷ Make function flexible $\Rightarrow$ decrease **bias**.
  - ▷ Make function flexible $\Rightarrow$ increase **variance** since it's harder to estimate more parameters associated with a more flexible model.
- Goal: find a function that minimizes the expected prediction error, which will balance **bias** and **variance** of our prediction.

Balancing Bias and Variance

- We have made progress: we have defined the error measurement that we want to use. We see that more flexible models won't necessarily lead to better predictions.
- Next:
  - If we can estimate the mean error for each possible $f(x)$ form (e.g. number of splines used) then we simply choose the function that's best.
  - **Q**: How to estimate the error?
- Validation:
  - Apply your prediction, $\widehat{f}(\text{age}_i)$, to some "new" data (so-called "validation" data, or "test" data).
  - Somehow use the data you have to estimate the out-of-sample error.

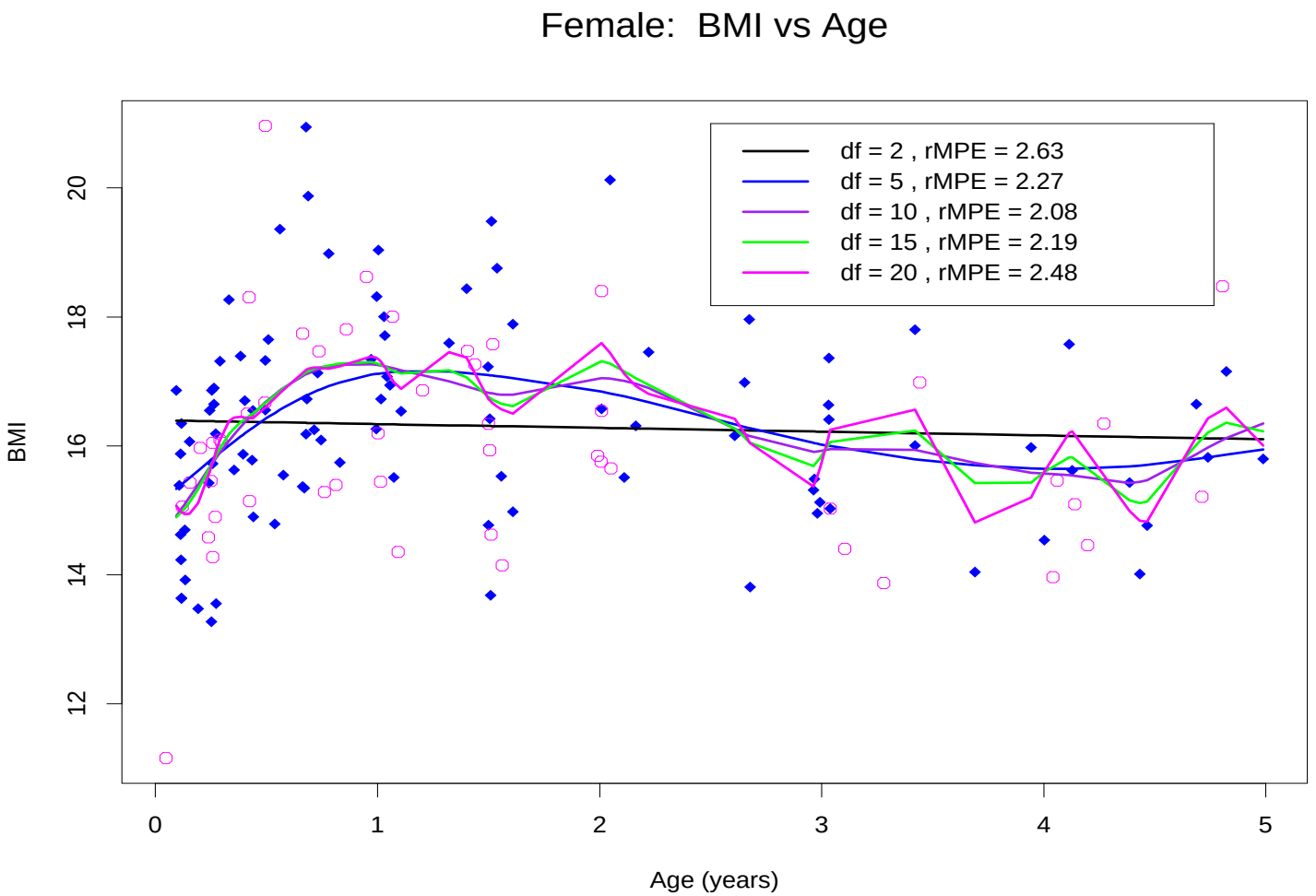
External Validation

- If you have some external test data (or have saved some data for validation) then the job is easy – just compute the possible prediction functions, apply them to the test data, and calculate their error.
- Example: For the BMI data we have 100 observations that we use to create prediction functions, and then apply the prediction function to the new data (50 observations). We measure:

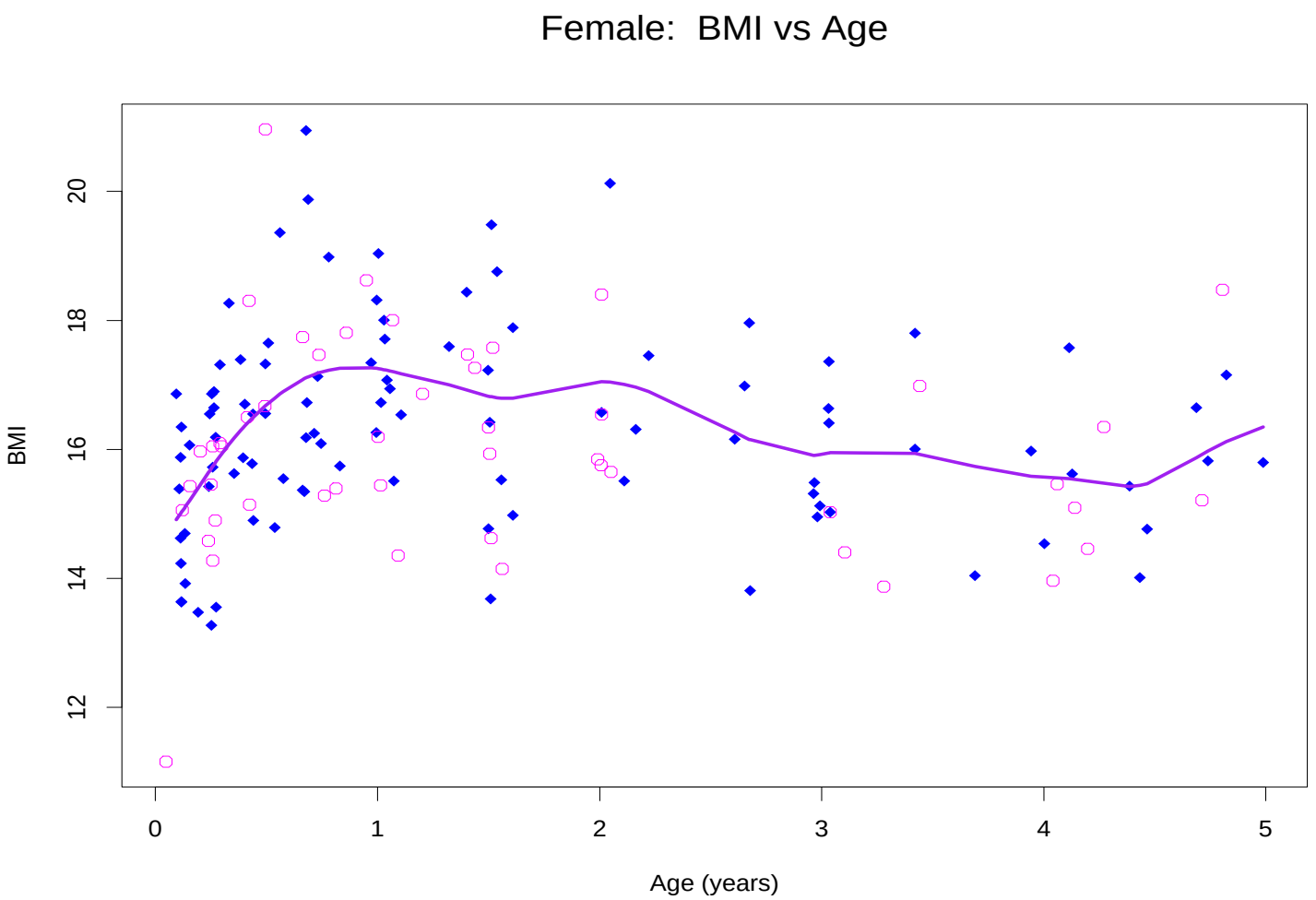
$$\texttt{mean error} = \frac{1}{N^{\texttt{new}}} \sum_j \left[Y_j^{\texttt{new}} - \widehat{f}(\texttt{age}_j) \right]^2$$

- We show the square root of this mean average as rMPE.

Example: BMI Data



Example: BMI Data (fit with df=10)



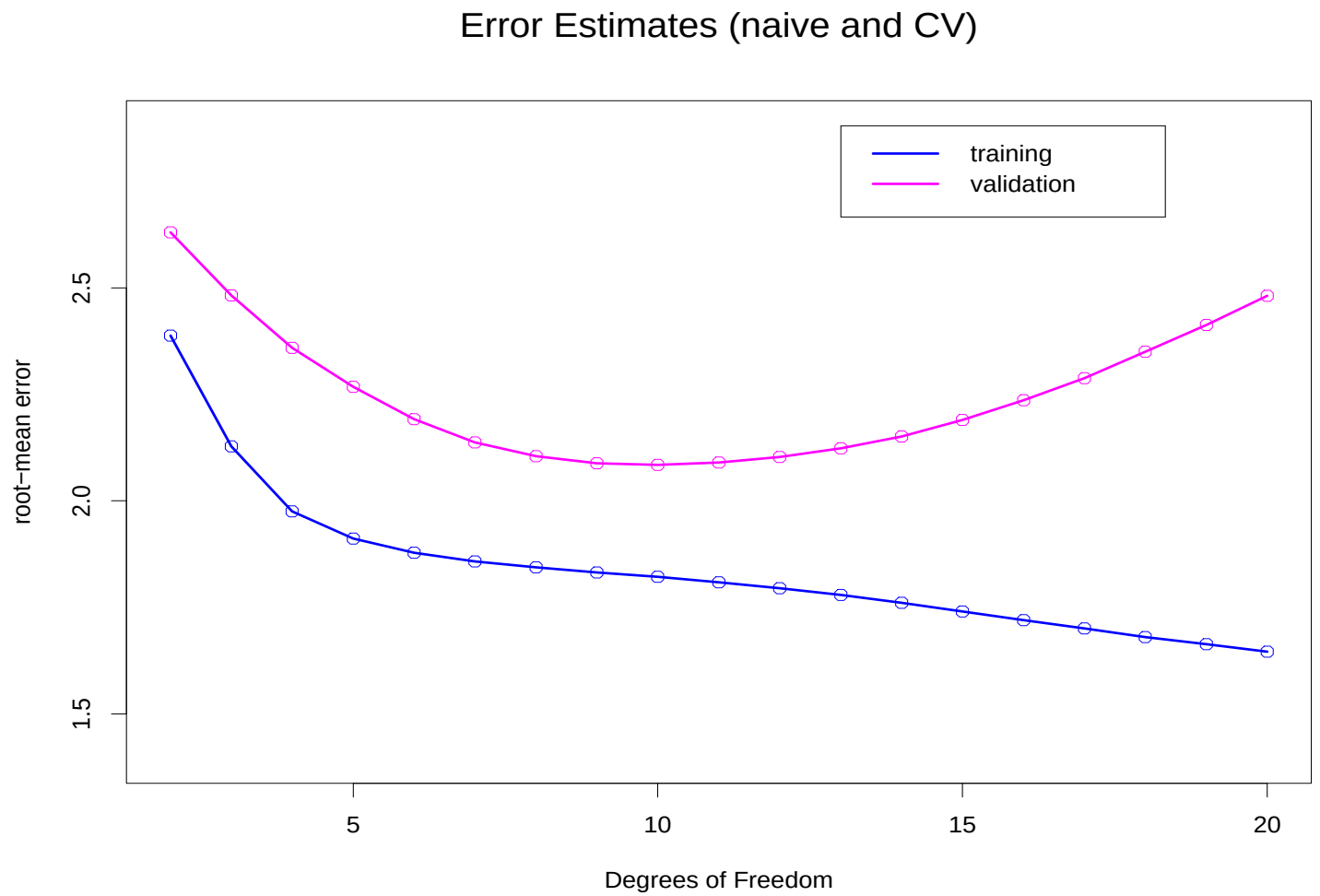
Validation

- **Q**: Why not just estimate the error of prediction based on the data you used to develop the prediction function:

$$\texttt{mean within} - \texttt{error} = \frac{1}{N^{\texttt{old}}} \sum_i \left[Y_i^{\texttt{old}} - \widehat{f}(\text{age}_i) \right]^2$$

- **A**: Because this estimate is too optimistic – you've used the data both to estimate the model, and to evaluate how well the model performs.
- Second, you will always decrease the within-sample error by adding more predictors, or more flexibility to your curve, $f(x)$.
- However, the optimism is proportional to $1/N$, so with large sample sizes the optimism may be small.

Example: BMI Data



Within-sample Estimates of Prediction Error

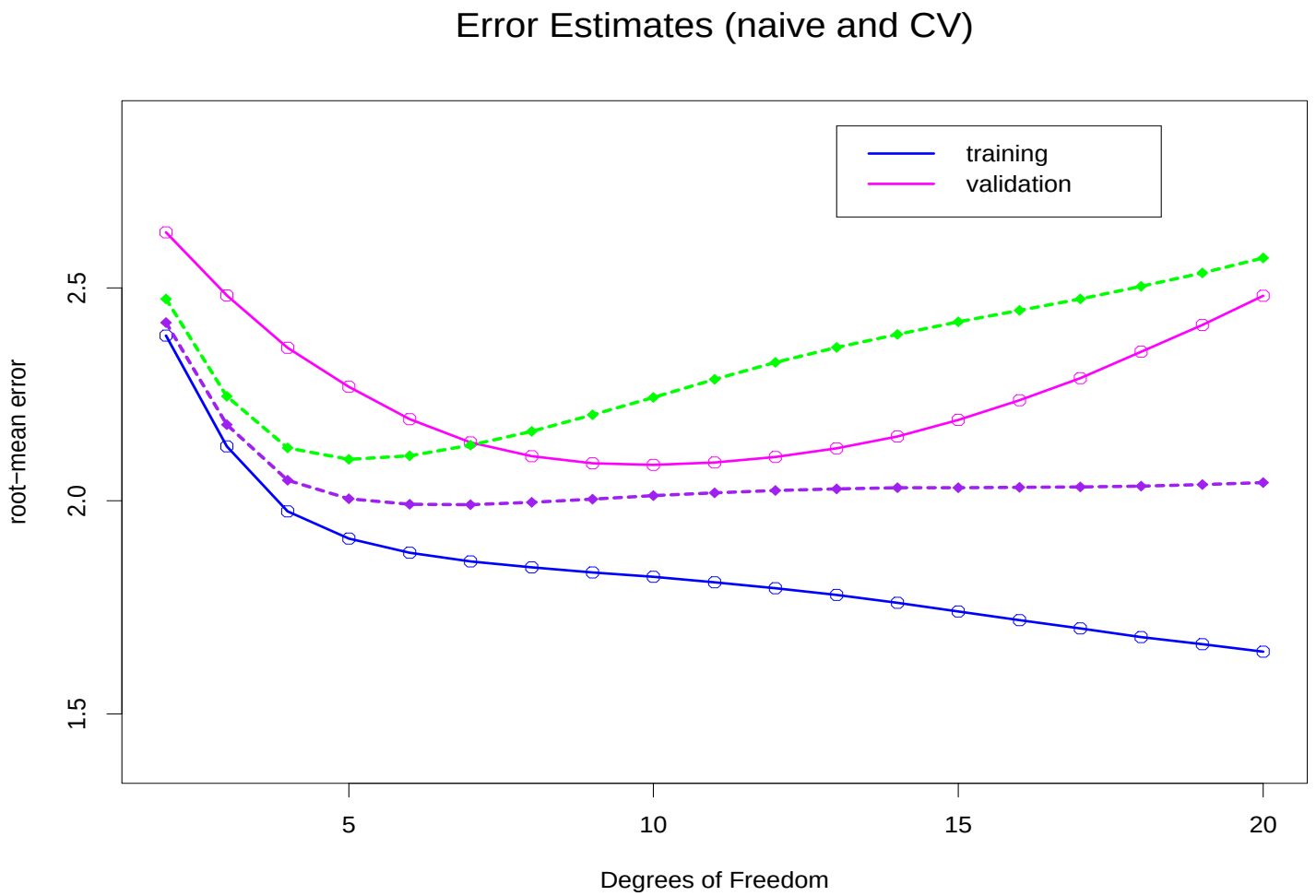
- Cross Validation: CV works by leaving points (Y_i, X_i) out one at a time, and estimating the function $f(x)$ based on the remaining $N-1$ points. This is an attempt to mimic the use of training and test samples for prediction. The CV estimate is then:

$$\texttt{CV} = \frac{1}{N}\sum_i \left[Y_i - \widehat{f}^{(-i)}(\texttt{age}_i)\right]^2$$

- Here $\widehat{f}^{(-i)}(x)$ is the estimated function based on data removing observation i.
- Estimate: There is an estimate called C_p that estimates the prediction error based on theoretical calculations. This is given as

$$C_p = \texttt{MSE}(\widehat{\texttt{f}}) + 2 \cdot p \cdot \frac{\sigma^2}{N}$$

Example: BMI Data



Example: BMI Data

- The **purple** curve on the previous plot is for C_p.
- The **green** curve on the previous plot is for `CV`.
- Cross-validation methods (CV) provide an approximately unbiased estimate of the prediction error.
- Cross-validation methods also sometimes create K "blocks" of data (for example breaking the data into K=10 blocks) and then apply for each observation the fit obtained from data in the other $K-1$ blocks of data). Block choice is another bias/variance trade-off where larger blocks lead to less variable estimates of error, but with more potential for bias.
- Stone (1977) showed that AIC and the leave-one-out cross-validation approach are asymptotically equivalent.

Summary with Continuous Response

- Goal = good prediction.
- Define criterion for judging = error scale
- Job is now easy – choose model / function that is estimated to provide the best prediction.
- Use external validation (test) data, or use the development data and cross-validation or C_p.
- **Q**: what about binary data, or survival data?

AIC: A Generalization

- Error Scale – the previous presentation with the BMI data supposed that we have adopted a scale to measure the "distance" from the prediction to the data.
- Recall – we used the likelihood (or partial likelihood) as a method for measuring how well a model "fits" data.

$$\log \mathcal{L}(\widehat{f}) = \sum_i \log \mathcal{L}(Y_i \mid \widehat{f}) = \sum_i \log P[Y_i \mid \widehat{f}]$$

AIC: A Generalization

- Here the model $\widehat{f}$ denotes the parameters for the model – e.g. the regression coefficients, β_j, or as in the previous BMI example, the estimate of the prediction curve $f(x)$.
- Since a higher value here reflects better agreement (prediction of Y_i) we can use -1 times this probability to denote error, or:

$$\texttt{error}_i = -2 \cdot \log \mathcal{L}(Y_i \mid \widehat{f})$$

AIC: A Generalization

- log Likelihood Error Scale – provided we are willing to consider distance as measured by -2 times the probability of the observed data we can proceed to try and find a model that has the smallest error.
- Viewed another way, we seek a predictive model that when we take the model and calculate the probability of some new test (validation) data, the predictive model will assign high probability to this new data.
- **AIC**: the Akaike Information Criterion is a method that allows estimation of the expected prediction error:

$$E\left[-2 \cdot \log \mathcal{L}(Y_i^{\texttt{new}} \mid \widehat{f})\right] \approx -\frac{2}{N} E[\log \mathcal{L}(\widehat{f})] + \frac{2 \cdot p}{N}$$

AIC: A Generalization

- AIC – Using this idea, we can estimate the expected error, or more commonly presented as the total error where we don't divide by N:

$$\texttt{AIC} = -2 \cdot \log \mathcal{L}(\widehat{f}) + 2 \cdot p$$

- Here $\log \mathcal{L}(\widehat{f})$ is the maximized log-likelihood using the model f which uses p parameters.
- Similar to before we can now search over models – possibly non-nested models, or models with the same number of parameters, and judge them in terms of their predictive potential.
- Notice that similar to before, we see that as we add more parameters to the model the $\log \mathcal{L}$ will increase (so -2 times it decreases) but we correct by "penalizing" for the number of parameters.

AIC: A Generalization

- Our previous criterion C_p is just AIC/N for the normal linear model.
- Summary:
  - If we define **prediction** as our goal, and
  - If we accept **log-likelihood** units as a measure of how far/close we predict,
  - Then we can use **AIC** as a guide for identifying one or more candidate predictive models.

AIC: A Generalization

- The only task is to define the possible models that would be used for prediction, and then fit them all.
- Note: how many models possible with m predictors? If we consider only additive models where each variable is either in or out as a main effect then we have 2^m possible models.
  - m = 10, 2^m = 1,024
  - m = 20, 2^m = 1,048,576
- STATA has a function `swaic` that will sequentially add or delete predictors depending on whether they improve AIC.
- Note: we are taking an average of the error across all observations, and are therefore averaging over the covariate distribution in our sample. Will that generalize?

Example: Mayo PBC Data

```
stset time, failure(status)

#delimit;
stcox
          age
          logalb
          alkphos
          ascites
          logbil
          chol
          edema
          hepmeg
          plate
          logpro
          sex
```

```
          sgot
          spiders
          stage
          treat
          trigly
          copper ;
#delimit cr

swaic, m
```

Example: Mayo PBC Data

Cox regression -- Breslow method for ties

No. of subjects = 280 No. of failures = 112
Log likelihood = -466.65255

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.0271	0.0114	2.41	0.016	1.0050	1.0497
logalb	0.1151	0.1082	-2.30	0.021	0.0182	0.7264
alkphos	0.9999	0.0000	-0.06	0.951	0.9999	1.0000
ascites	1.2756	0.4912	0.63	0.527	0.5996	2.7135
logbil	2.1988	0.3483	4.97	0.000	1.6119	2.9995
chol	1.0001	0.0004	0.24	0.809	0.9992	1.0009
edema	1.8652	0.5380	2.16	0.031	1.0598	3.2828
hepmeg	0.9199	0.2340	-0.33	0.743	0.5586	1.5146

plate	1.0007	0.0011	0.63	0.532	0.9984	1.0030
logpro	10.7850	14.17989	1.81	0.070	0.8197	141.8928
sex	0.8436	0.2664	-0.54	0.590	0.4543	1.5666
sgot	1.0028	0.0020	1.38	0.166	0.9988	1.0067
spiders	1.0808	0.2576	0.33	0.744	0.6773	1.7246
stage	1.4187	0.2428	2.04	0.041	1.0143	1.9843
treat	0.9843	0.2080	-0.07	0.941	0.6504	1.4895
trigly	0.9979	0.0013	-1.59	0.112	0.9953	1.0004
copper	1.0015	0.0012	1.23	0.217	0.9990	1.0040

Example: Mayo PBC Data

Stepwise Model Selection by AIC

stcox regression.

number of obs = 280

		Df	Chi2	P>Chi2	-2*ll	Df Res.	AIC
Null Model					1114.80	280	1114.80
Step 1:	logbil	1	115.59	0.0000	999.21	279	1001.21
Step 2:	logalb	1	23.19	0.0000	976.02	278	980.02
Step 3:	age	1	15.56	0.0001	960.47	277	966.47
Step 4:	logpro	1	8.95	0.0028	951.52	276	959.52
Step 5:	copper	1	5.30	0.0213	946.21	275	956.21
Step 6:	stage	1	4.08	0.0435	942.14	274	954.14
Step 7:	edema	1	3.08	0.0792	939.06	273	953.06
Step 8:	trigly	1	2.68	0.1016	936.38	272	952.38

```
Step 9:     sgot  1   1.77   0.1832    934.60    271     952.60
Step 10:  ascites 1   0.37   0.5412    934.23    270     954.23
Step 11:    plate 1   0.38   0.5374    933.85    269     955.85
Step 12:      sex 1   0.25   0.6175    933.60    268     957.60
Step 13:  spiders 1   0.12   0.7258    933.48    267     959.48
Step 14:   hepmeg 1   0.10   0.7467    933.37    266     961.37
Step 15:     chol 1   0.06   0.8066    933.31    265     963.31
Step 16:    treat 1   0.00   0.9437    933.31    264     965.31
Step 17:  alkphos 1   0.00   0.9510    933.31    263     967.31
-----------------------------------------------------------------
```

Example: Mayo PBC Data

No. of subjects = 280 No. of failures = 112
Log likelihood = -468.18773

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logbil	2.400	0.3104	6.77	0.000	1.8631	3.0933
logalb	0.090	0.0737	-2.94	0.003	0.0181	0.4478
age	1.025	0.0097	2.69	0.007	1.0069	1.0450
logpro	8.572	10.7273	1.72	0.086	0.7379	99.5967
copper	1.002	0.0010	2.18	0.029	1.0002	1.0042
stage	1.334	0.1907	2.02	0.044	1.0079	1.7654
edema	1.665	0.4353	1.95	0.051	0.9982	2.7803
trigly	0.998	0.0011	-1.55	0.122	0.9958	1.0004

Example: Mayo PBC Data

- This model with the minimum AIC includes the (5) "Mayo model" variables (logbil, logpro, logalb, edema, age) and (3) additional variables.
- Some of the additional variables were not allowed as part of the Mayo model subset since they were not routinely or easily available.
- Here we have used $m = 17$ candidate covariates, leading to:

  $2^{17} = 131{,}072$ possible additive models.
- Notice that AIC includes `trigly` even though the variable doesn't achieve nominal 0.05 significance.

Example: Mayo PBC Data

- This is due to the fact that AIC will include a variable (with 1 df) provided the likelihood increases enough to compensate for the penalty.
- This implies the likelihood ratio must be at least 2.0 for AIC to add.
- The critical value for a 1 df LR test is 3.86. An LR test with value 2.0 corresponds to a p-value of 0.1573.
- **Caution**: the inference associated with the model displayed does not account for the fact that model selection was done. We only display those covariates that have p-values (approx) less than 0.15. How to interpret the p-value given that we only see it if it's less than 0.15?

A Related Idea: BIC (* = extra)

- Bayesian Information Criterion (BIC) is similar to AIC but motivated from a different perspective.
- Assume we have a set of $m = 1, 2, \ldots, M$ models that we assume all have an equal probability of being the "true" model.
- After analysis of the data we can then update our probabilities, and the model with the smallest BIC is the model with the highest probability of being the correct model.
- Define:

$$\texttt{BIC} = -2 \cdot \log \mathcal{L}(\widehat{f}) + \log(N) \cdot p$$

- Rather than a penalty of $2 \cdot p$ used for AIC, this criterion uses a penalty of $\log(N) \cdot p$, and therefore will favor more parsimonious models.

Predictive Accuracy and Survival Data

- Use of AIC presumes that we are willing to adopt the log-likelihood scale to measure error.
- **Q**: How might the prediction actually be used in practice?
  - ▷ Predict a survival curve?
  - ▷ Predict a survival time?
  - ▷ Predict mortality within 1 year?
- In some (many) situations it is meaningful to try and identify those subjects that are likely to die with a certain follow-up time (1 year, 2 years, 5 years). These subjects are candidates for aggressive therapy, while the others may not require such procedures.
- In this case, we can consider **classification errors** associated with a predictive model.

Components of Accuracy

- **Calibration**
  - Bias – does observed match predicted?
  - Evaluated graphically and formally.
- **Discrimination**
  - Does prediction separate subjects with different risks?
  - Evaluated qualitatively based on K-M plots.
- For binary outcomes we have standard concepts of classification error, and associated discrimination summaries.

Binary Classification

Sensitivity "True Positive"

$$\begin{aligned} \text{BINARY TEST} &: P(T+ \mid D=1) \\ \text{CONTINUOUS MARKER} &: P(M > c \mid D=1) \end{aligned}$$

Specificity "True Negative"

$$\begin{aligned} \text{BINARY TEST} &: P(T- \mid D=0) \\ \text{CONTINUOUS MARKER} &: P(M \leq c \mid D=0) \end{aligned}$$

ROC Curve

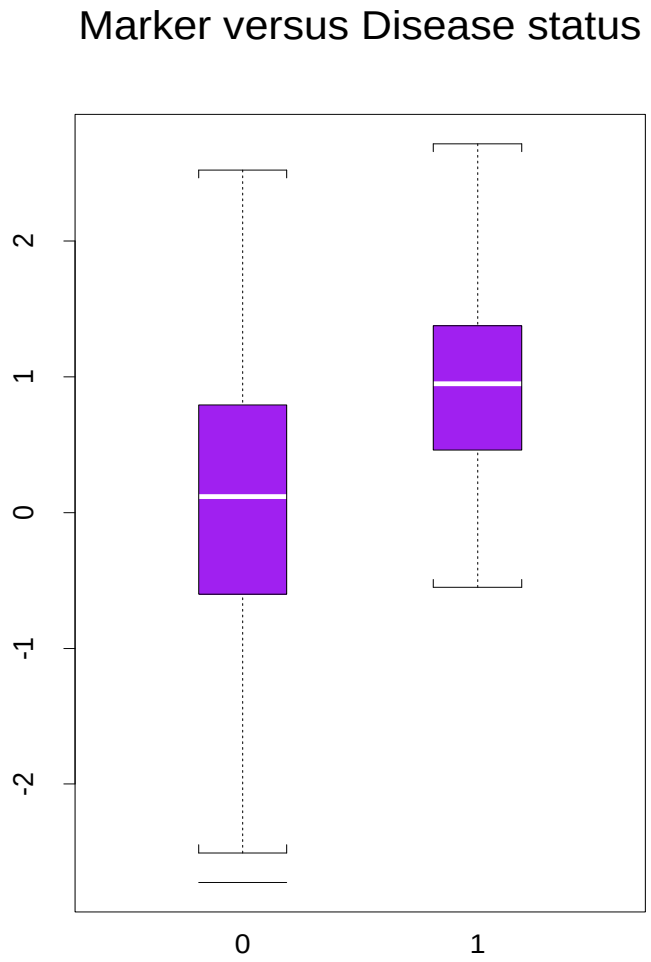
An ROC curve plots the True Positive Rate, TP(c), versus the False Positive Rate, FP(c) for all possible cutpoints, c:

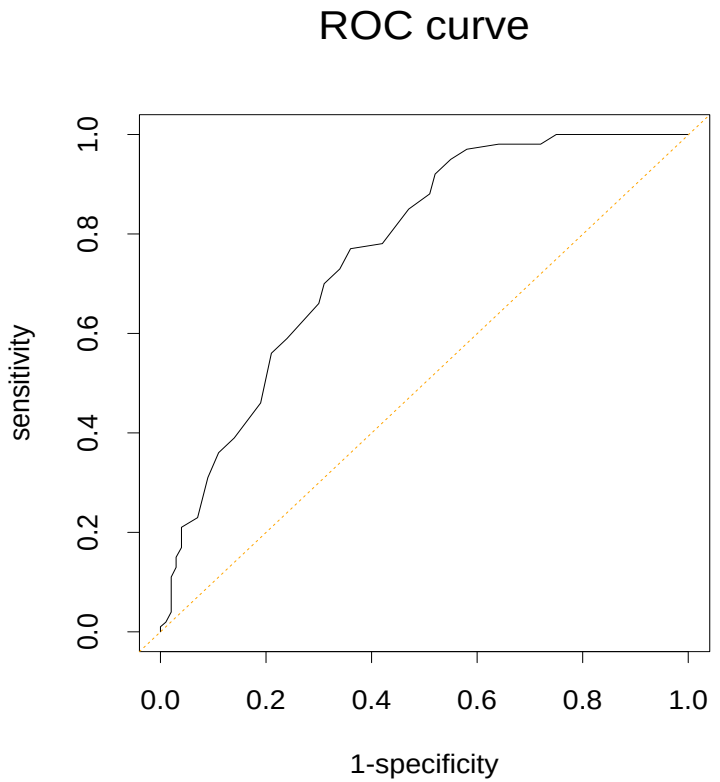
$$\text{FP}(c) \quad = \quad P(M > c \mid D = 0)$$

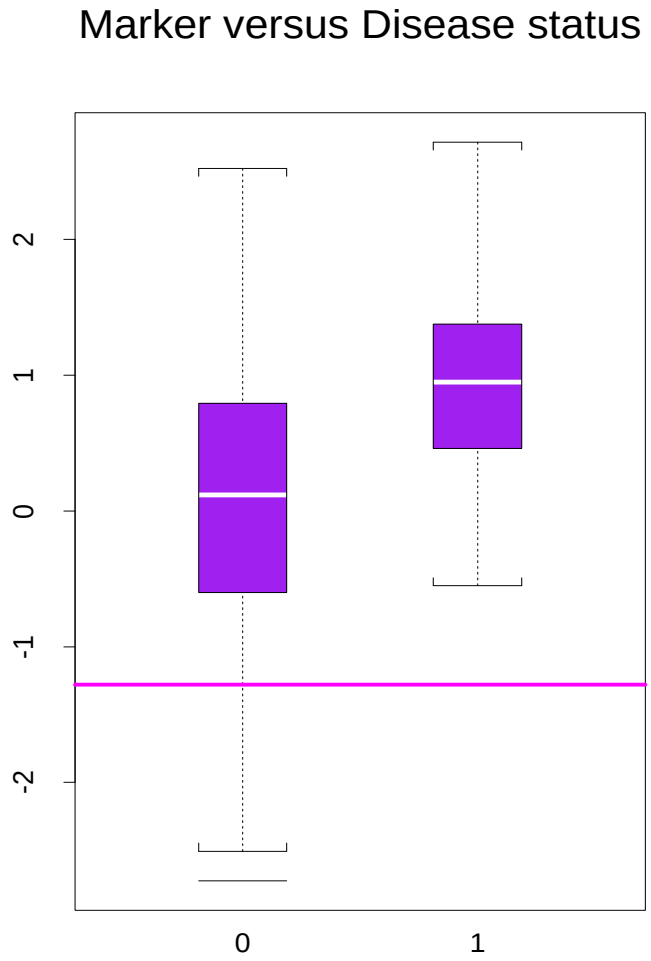
$$\text{TP}(c) \quad = \quad P(M > c \mid D = 1)$$

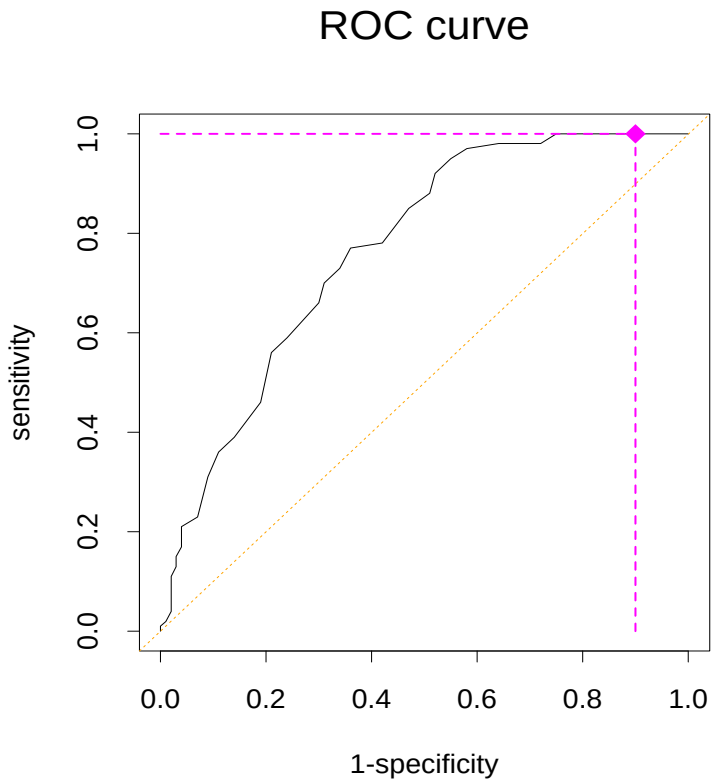
$$\text{ROC Curve} \quad : \quad [\, FP(c),\, TP(c)\,] \quad \forall c$$

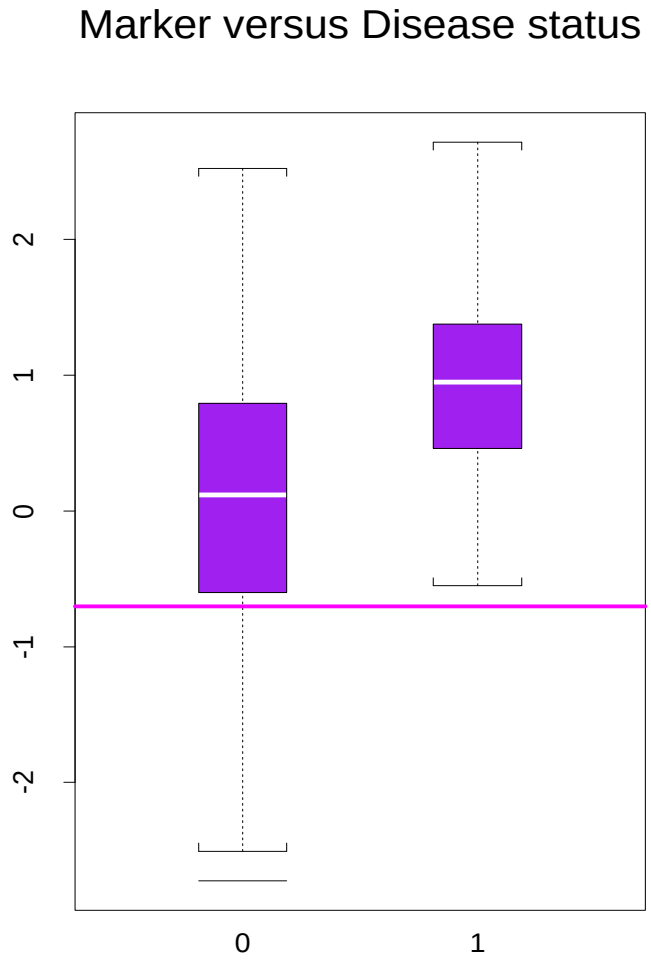
$$\text{ROC}(p) \quad : \quad [\, p = FP(c^p),\, TP(c^p)\,] \quad \text{for } \; p \in [0, 1]$$

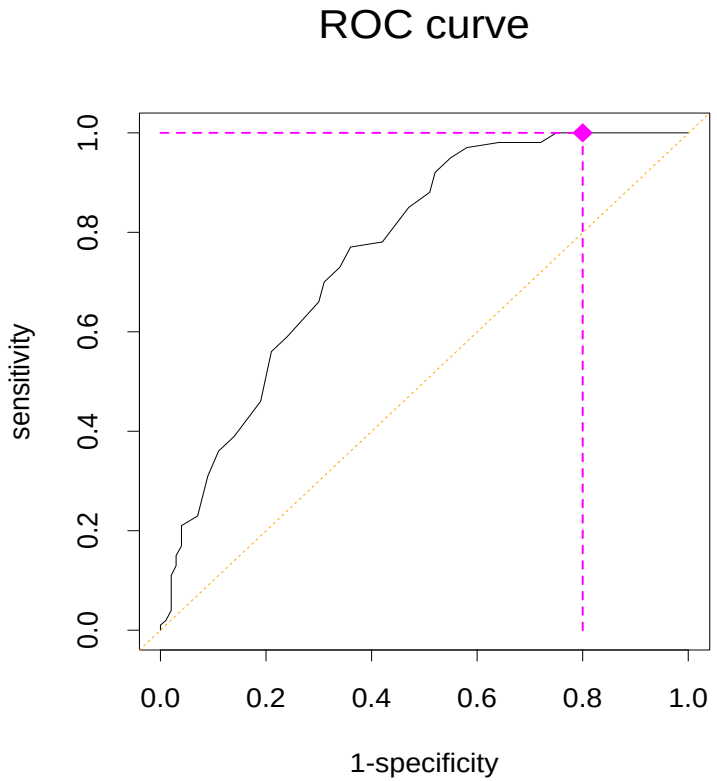

Marker versus Disease status
2
1
0
-1
-2
0
1

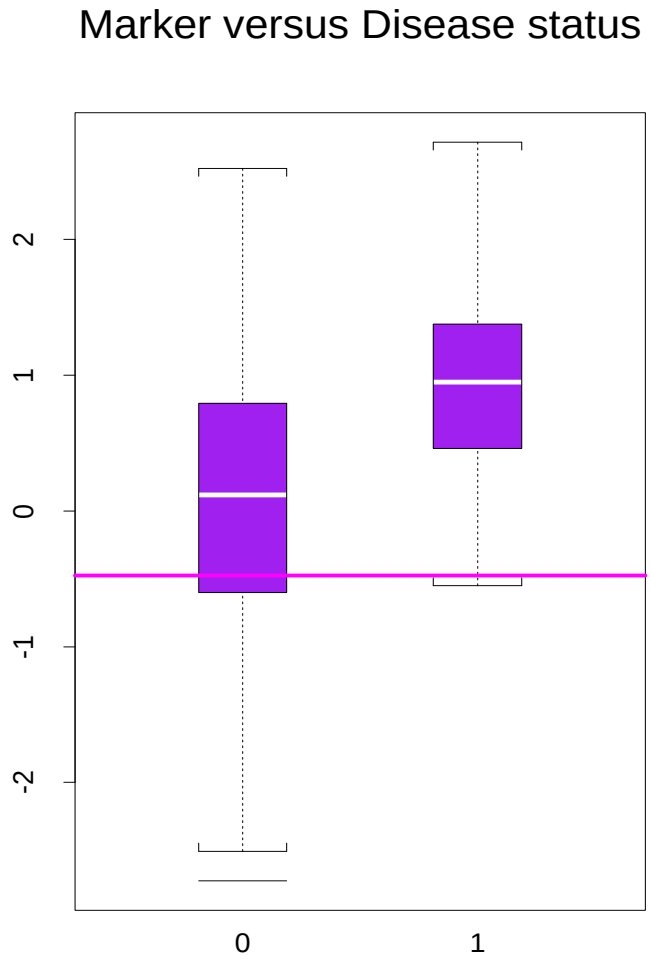

ROC curve
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity

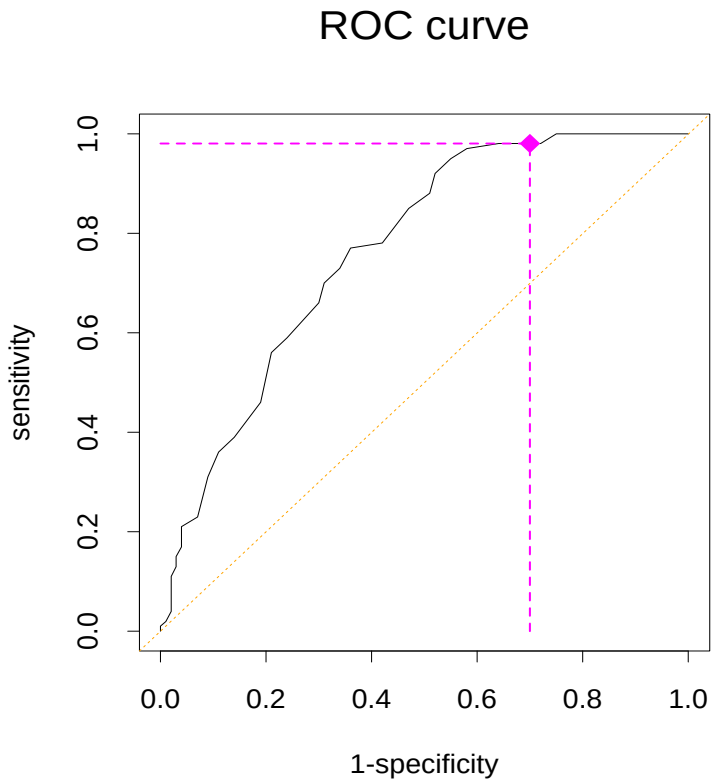

Marker versus Disease status
2
1
0
-1
-2
0
1

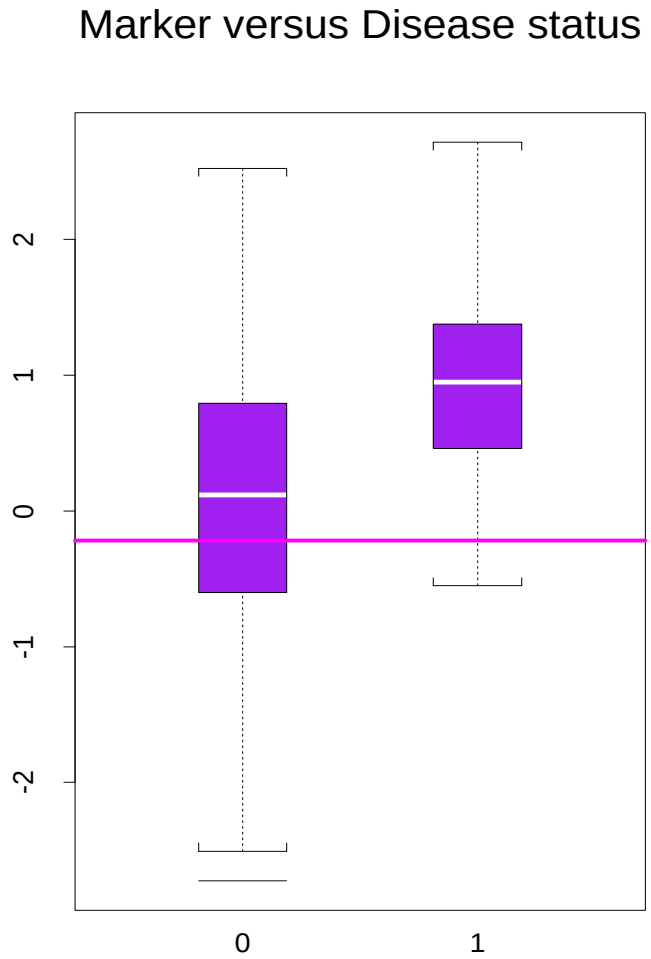

ROC curve
sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity

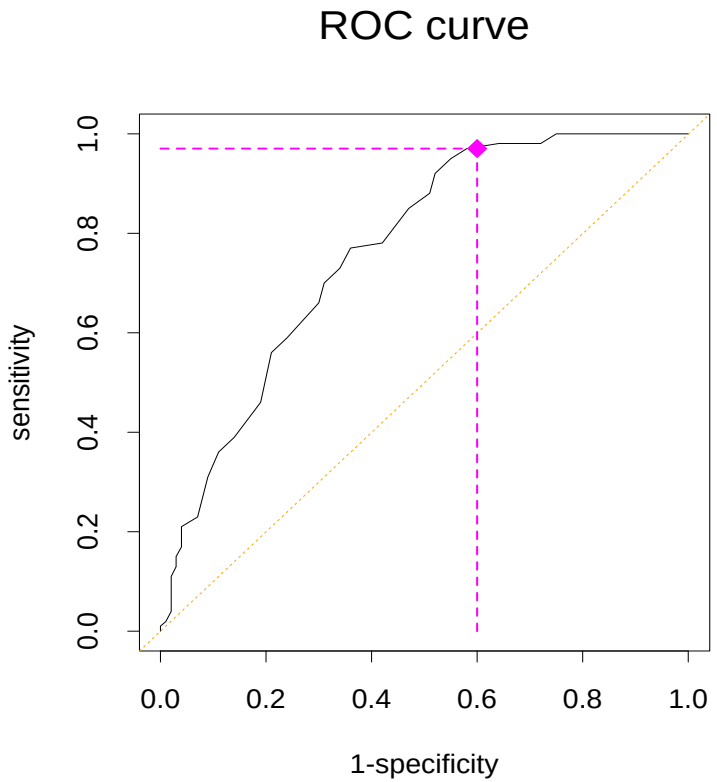

Marker versus Disease status
2
1
0
-1
-2
0
1

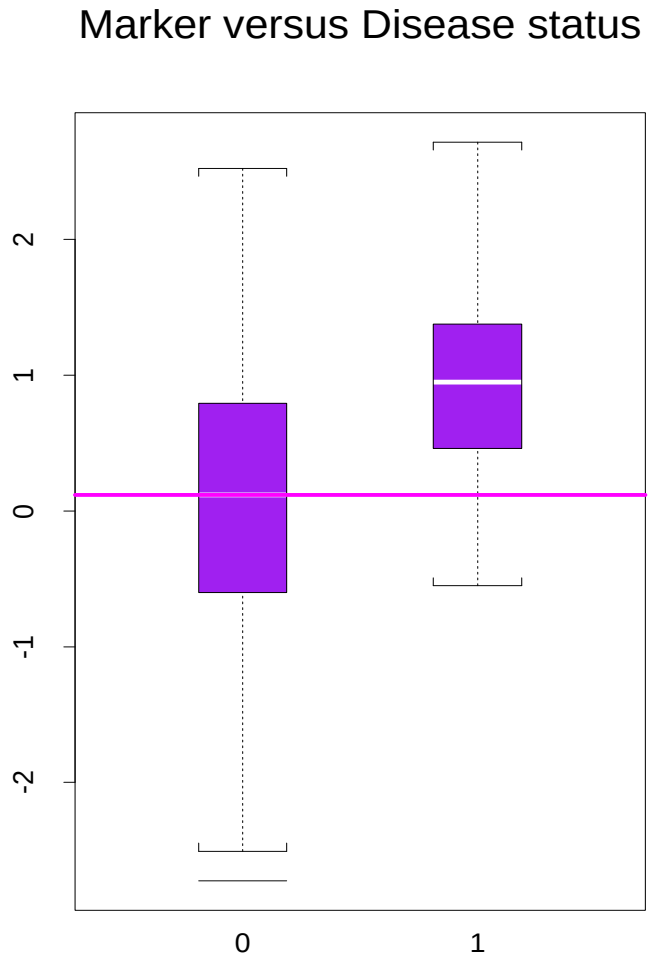

ROC curve
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity

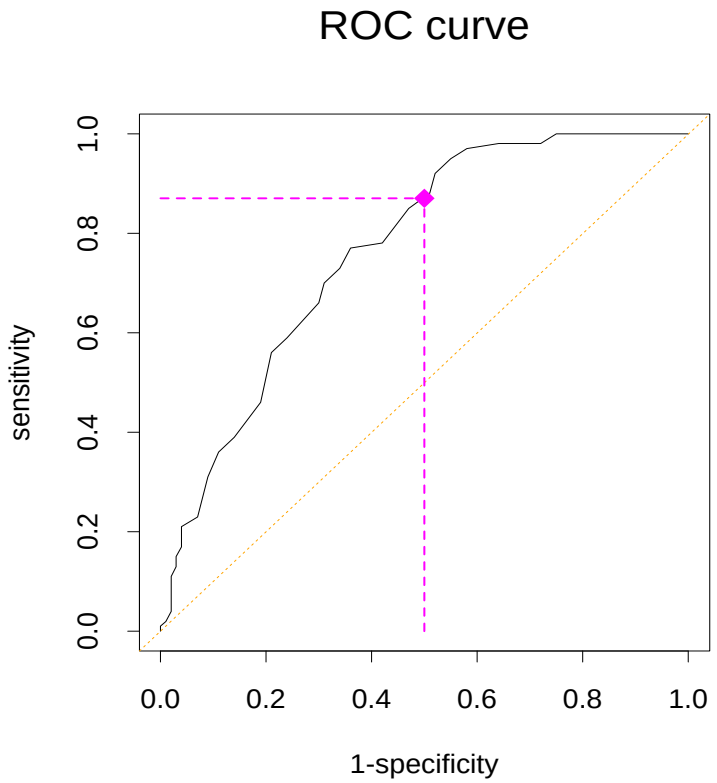

Marker versus Disease status
2
1
0
-1
-2
0
1

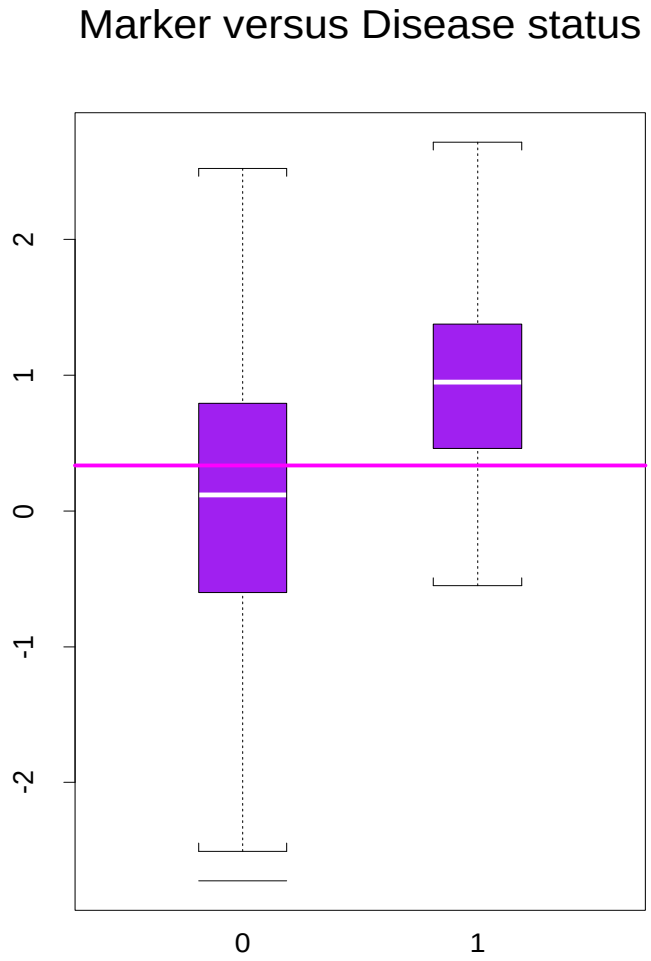

ROC curve
sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity

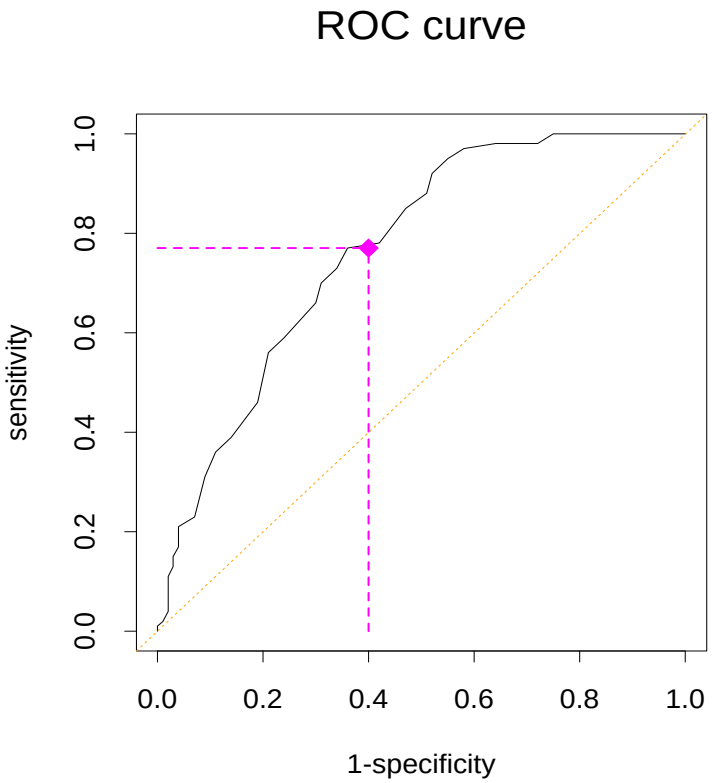

Marker versus Disease status
2
1
0
-1
-2
0
1

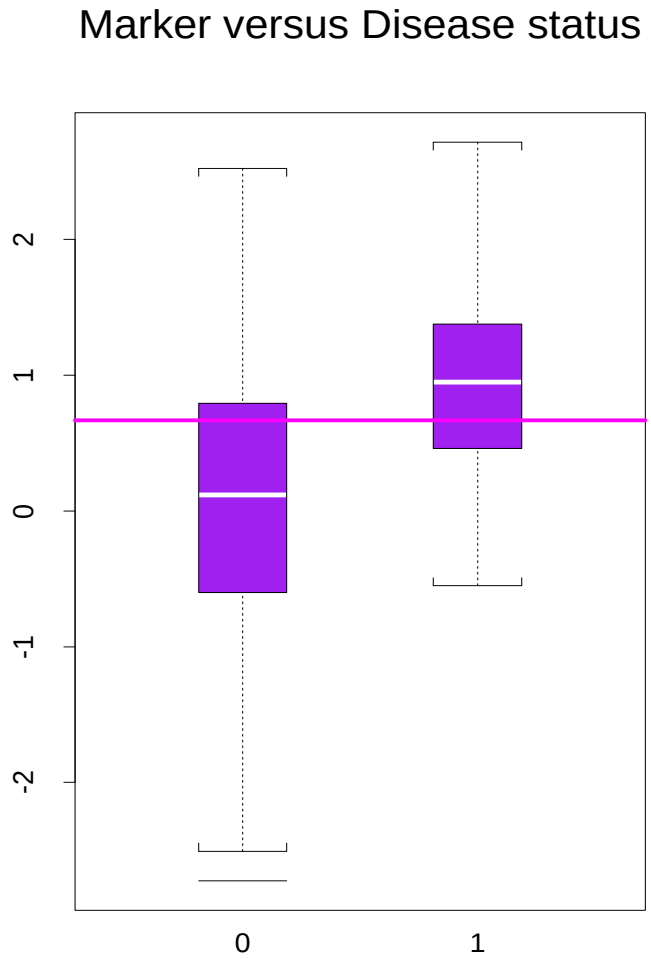

ROC curve
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity

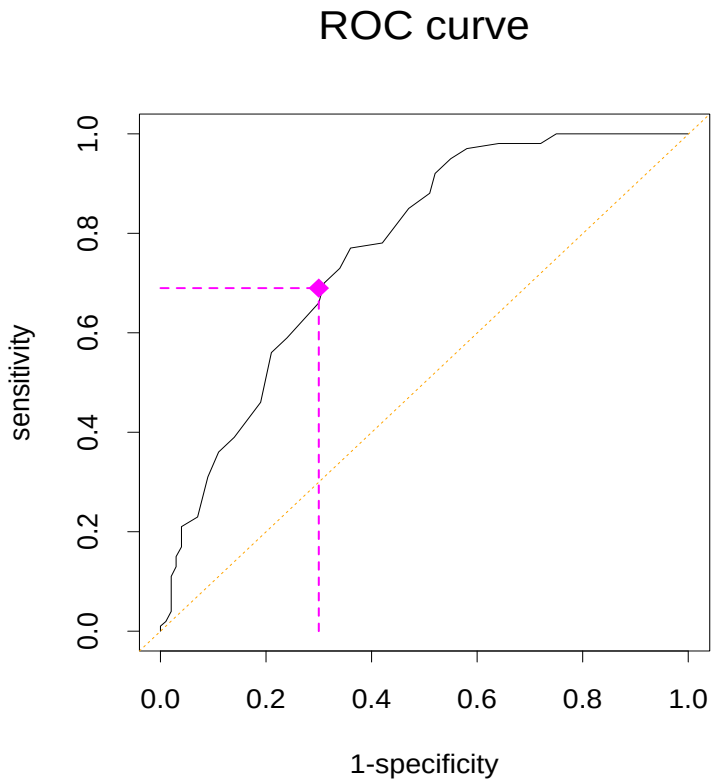

Marker versus Disease status
2
1
0
-1
-2
0
1

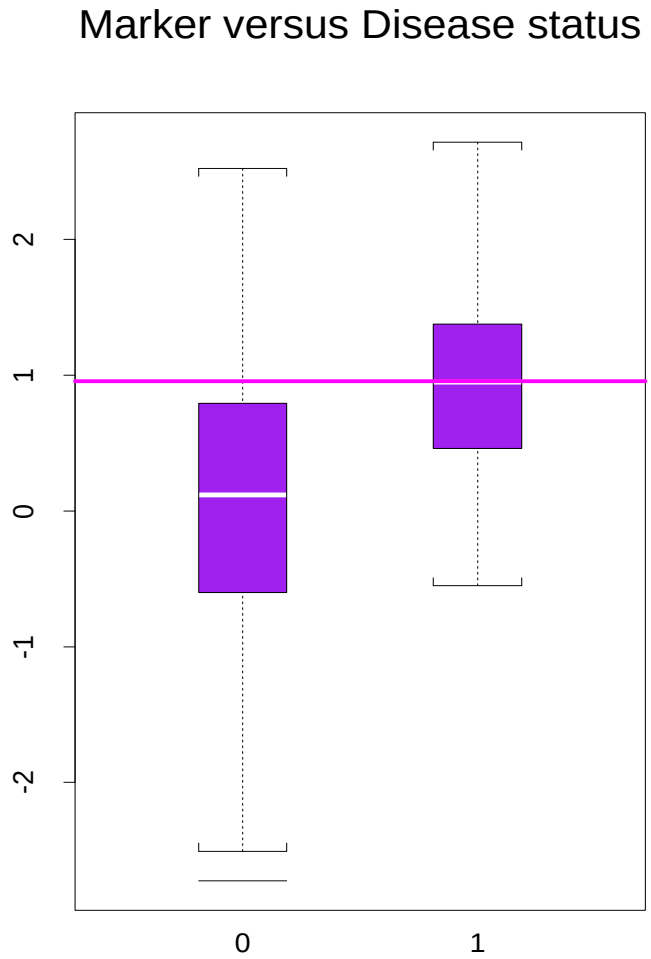

ROC curve
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity

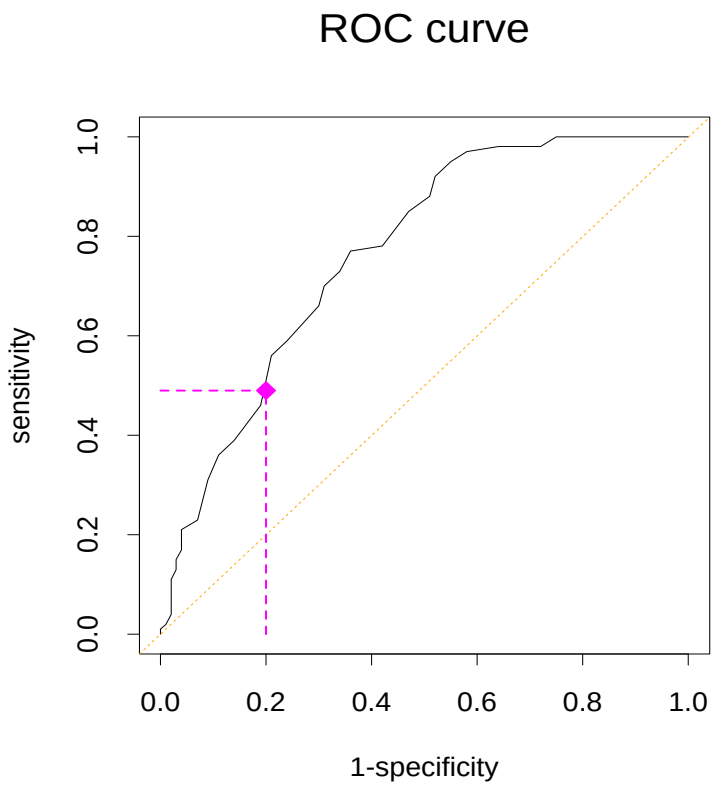

Marker versus Disease status
2
1
0
-1
-2
0
1

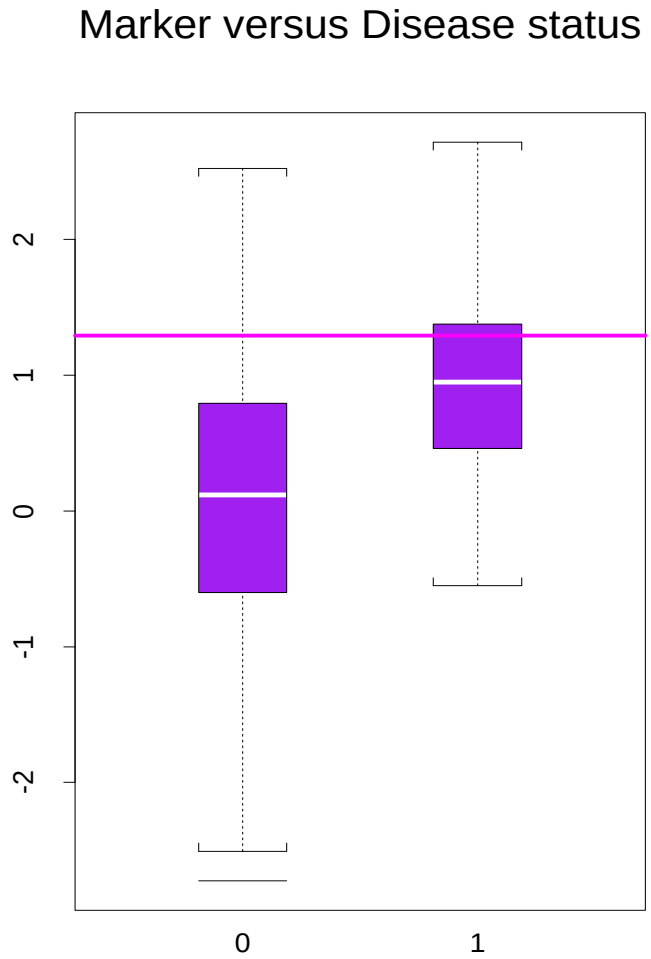

ROC curve
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity

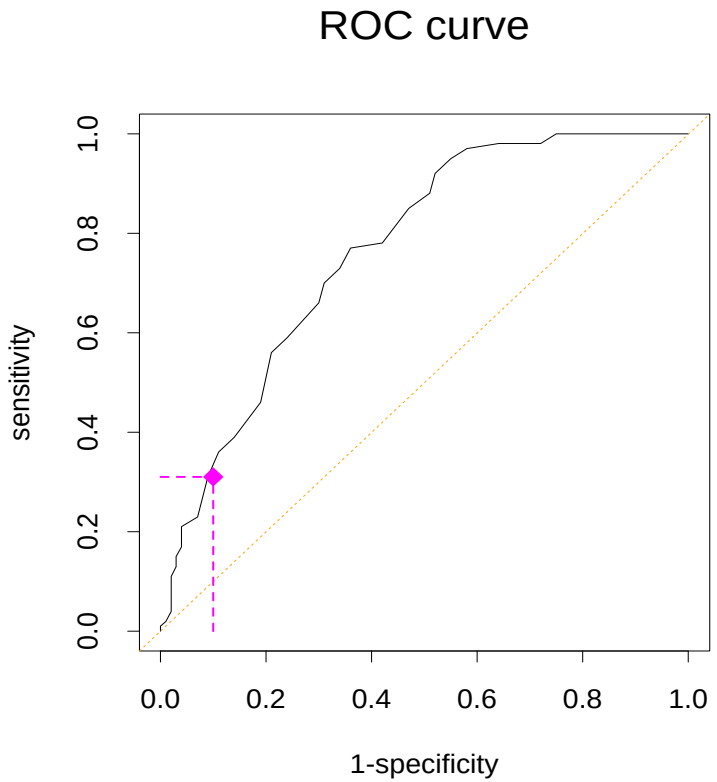

Marker versus Disease status
2
1
0
-1
-2
0
1


ROC curve
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity


Marker versus Disease status
2
1
0
-1
-2
0
1


ROC curve
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity


Marker versus Disease status
2
1
0
-1
-2
0
1


ROC curve
sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
1-specificity

ROC Curves

1. "ROC plots provide a pure index of accuracy by demonstrating the limits of a test's ability to discriminate between alternative states of health over a complete spectrum of operating conditions"
   Zweig and Campbell (1993)
2. Compare different markers.
3. Compare sensitivity when controlling specificity.
4. AUC interpretation:
   "For a randomly chosen case and control, the area under the ROC curve is the probability that the marker for the case is greater than the marker for the control."
5. AUC is a marker-outcome concordance summary (c-index).

Sensitivity and Specificity for Survival

Let T denote the survival time, and let $D(t)$ denote the counting process for the uncensored outcome:

$$D(t) = 1(T \leq t)$$

Possible definitions:

$$\text{CASE}(t) \quad : \quad \left\{ \text{Cumulative} \quad D(t) = 1 \right.$$

$$\text{CONTROL}(t) \quad : \quad \left\{ \text{Dynamic} \quad D(t) = 0 \right.$$

Example: 2-year mortality and Mayo PBC

- Consider the 5-variable "Mayo model", and the 8-variable model identified using AIC.
- Consider classification of subjects according to their "model score" defined as $\widehat{\beta}_1 \cdot X_1 + \widehat{\beta}_2 \cdot X_2 \ldots$.
- Consider **Cases** to be subjects who die within 2-years.
- Consider **Controls** to be subjects who live beyond 2-years.
- **Q**: How well do the two models discriminate between the 2-year cases and controls?

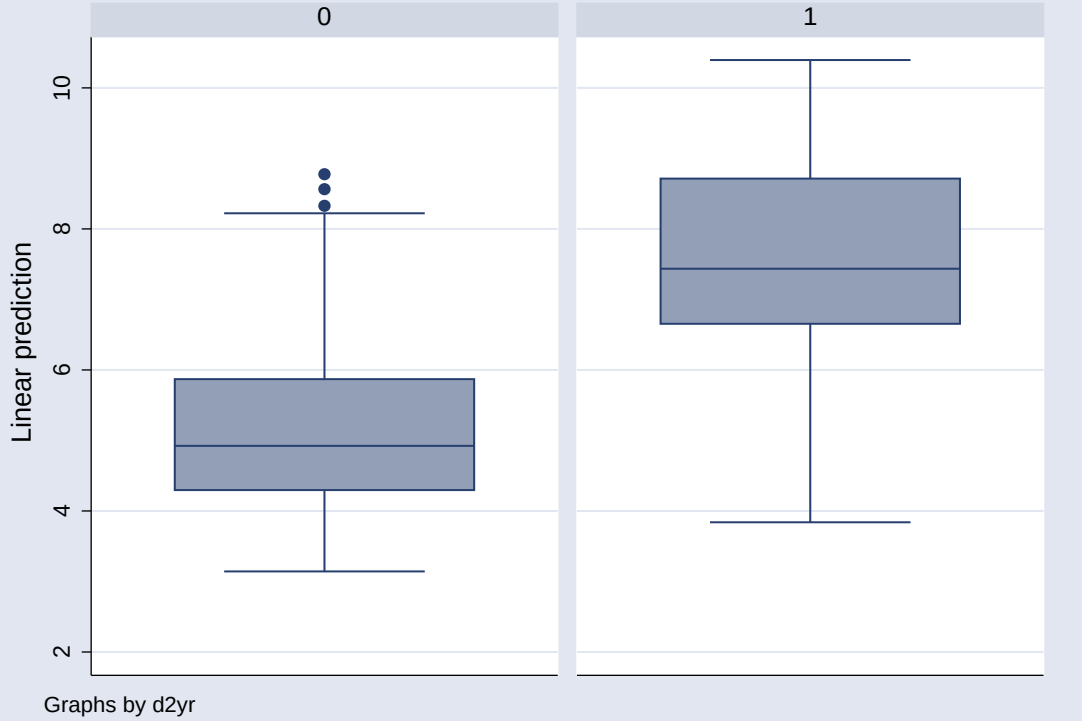
Example: 2-year mortality and Mayo PBC

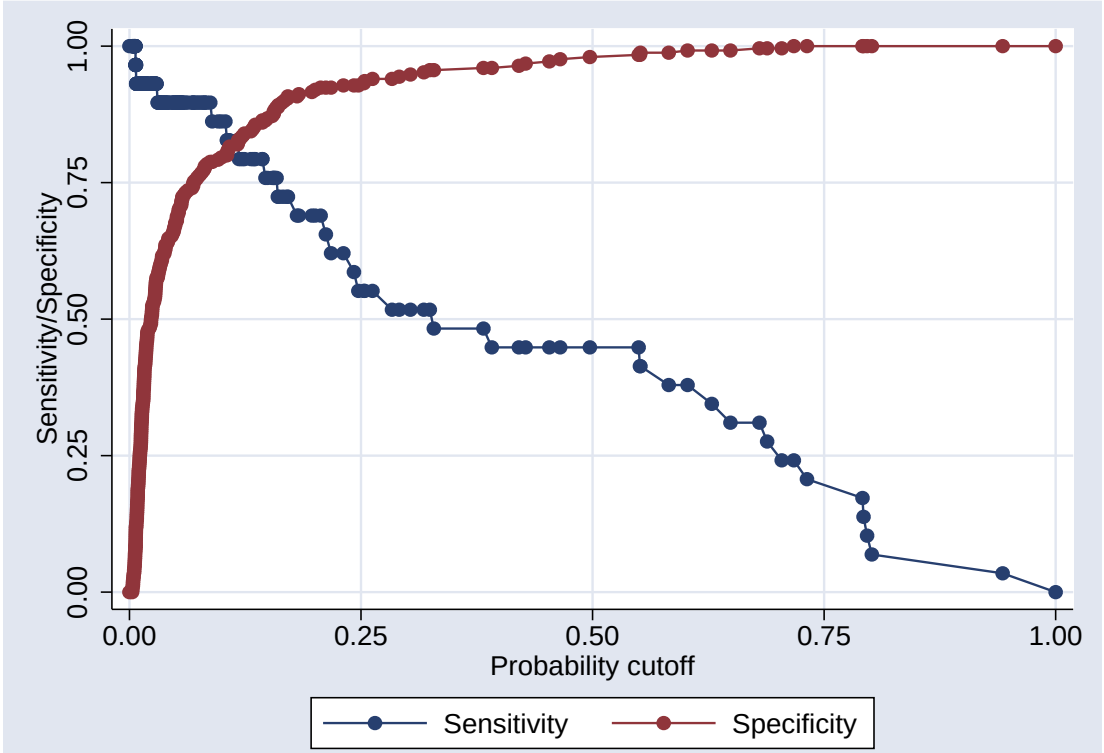
```
**********************************
***   model 1                  ***
**********************************
stcox logbil logalb logpro age edema
predict score1, xb

**********************************
***   model 2                  ***
**********************************
stcox logbil logalb logpro age edema copper stage trigly
predict score2, xb

*****
***** consider 2-year survival
*****
```

```
gen d2yr = (time <= 365*2)

*** check for censoring before 2-years
tab d2yr status
recode d2yr 1=. if status==0

*****
*****  ROC curves
*****
graph box score1, by(d2yr)
graph box score2, by(d2yr)

*** score 1 ***
logit d2yr score1
lsens, gensens( sens1 ) genspec( spec1 )
lroc

*** score 2 ***
```

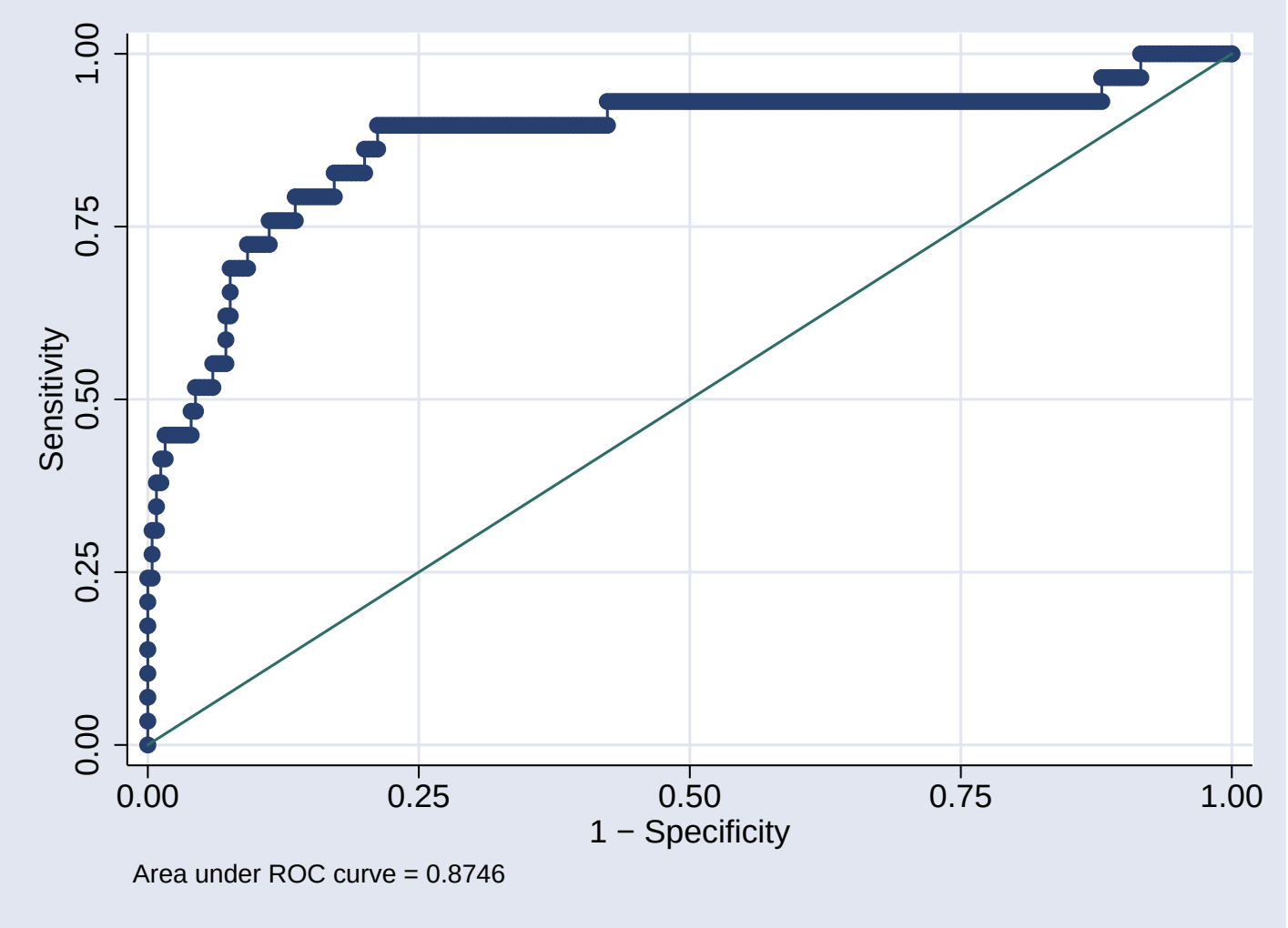
```
logit d2yr score2
lsens, gensens( sens2 ) genspec( spec2 )
lroc

*** both ROC curves on one plot ***
gen FP1 = 1-spec1
gen TP1 = sens1
gen FP2 = 1-spec2
gen TP2 = sens2

sort FP1
graph twoway (connected TP1 FP1) (scatter TP2 FP2)
```

Example: 2-year mortality and Mayo PBC

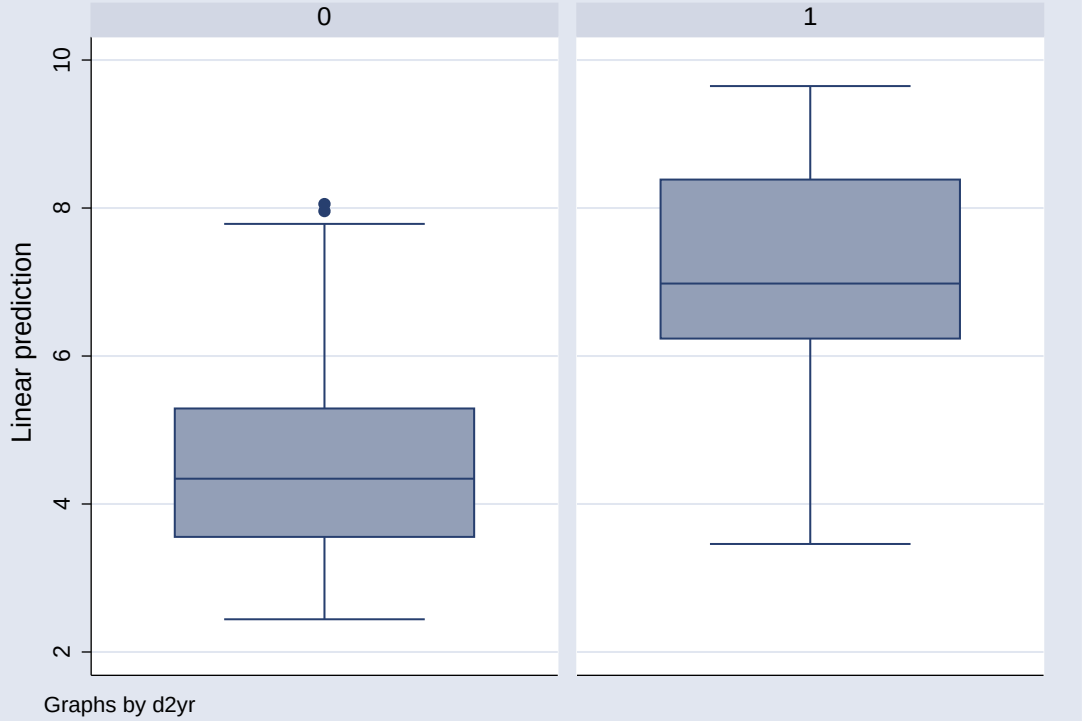


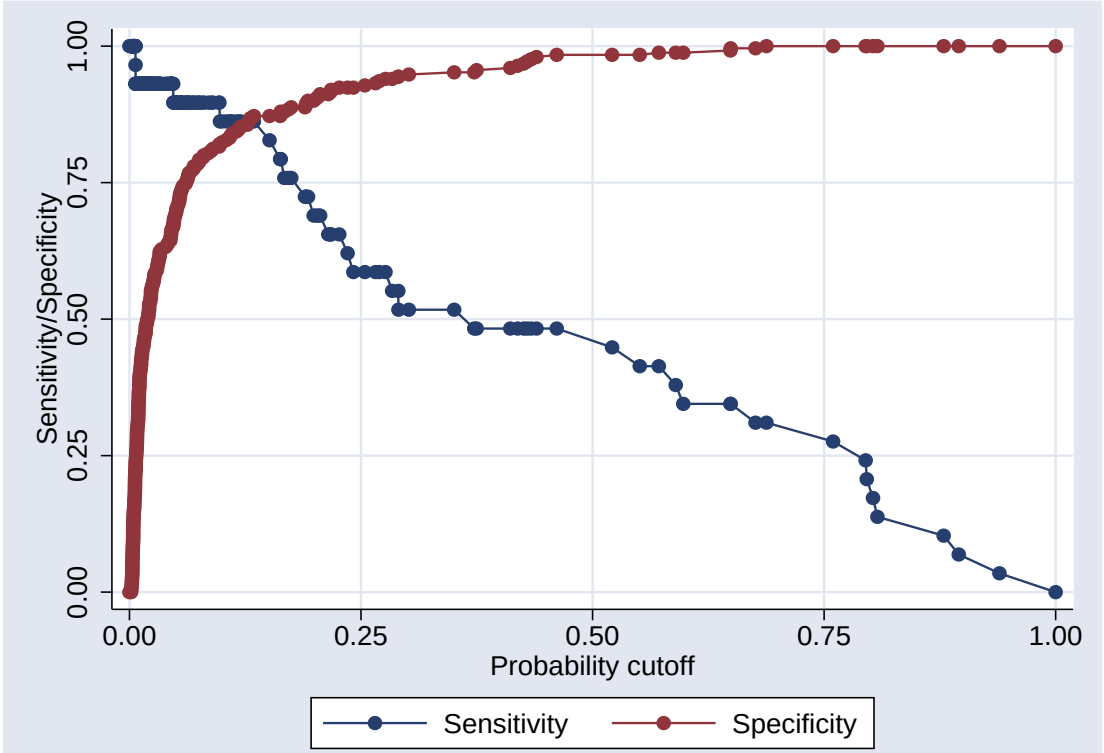


Example: 2-year mortality and Mayo PBC

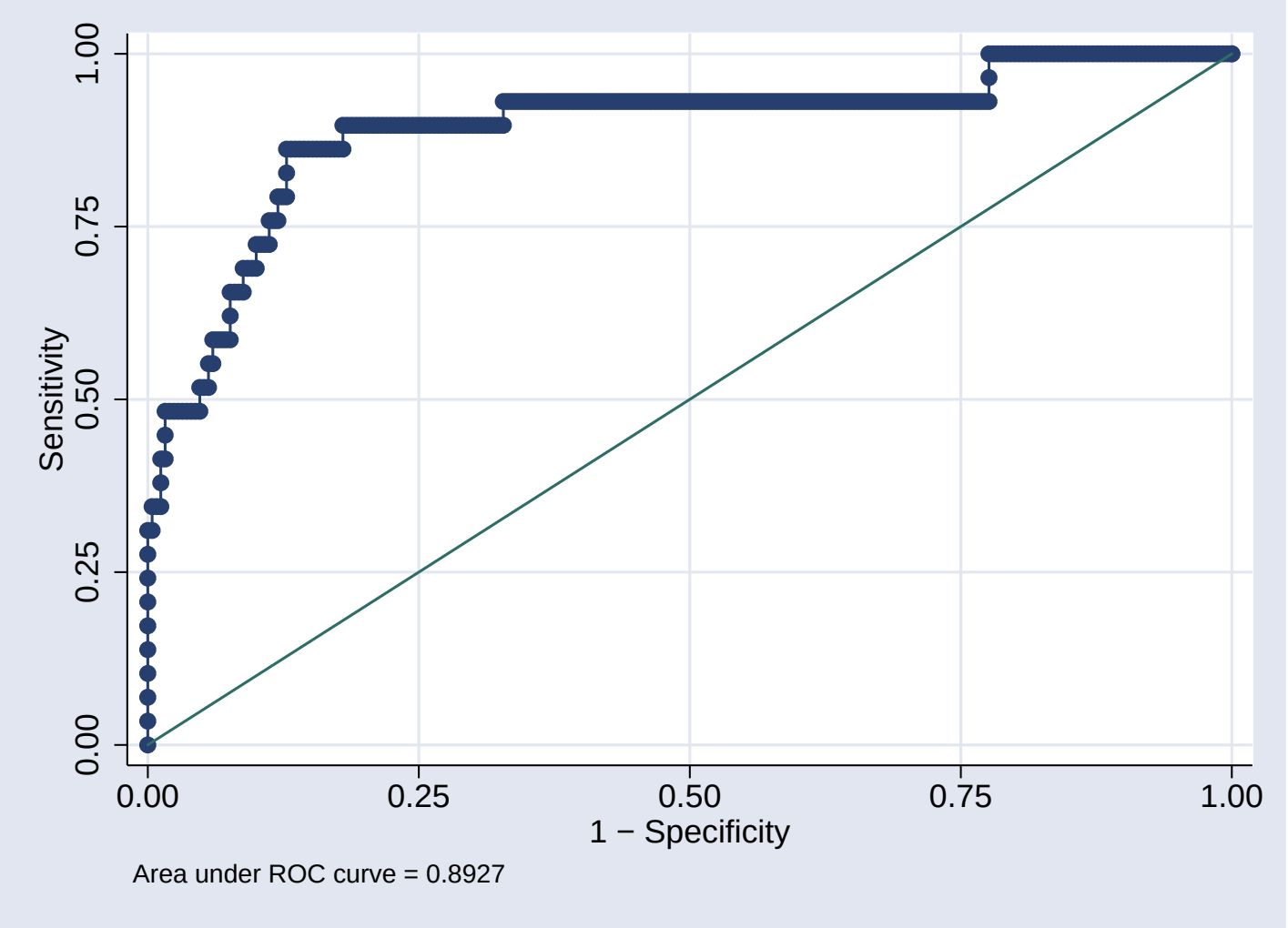


Example: 2-year mortality and Mayo PBC

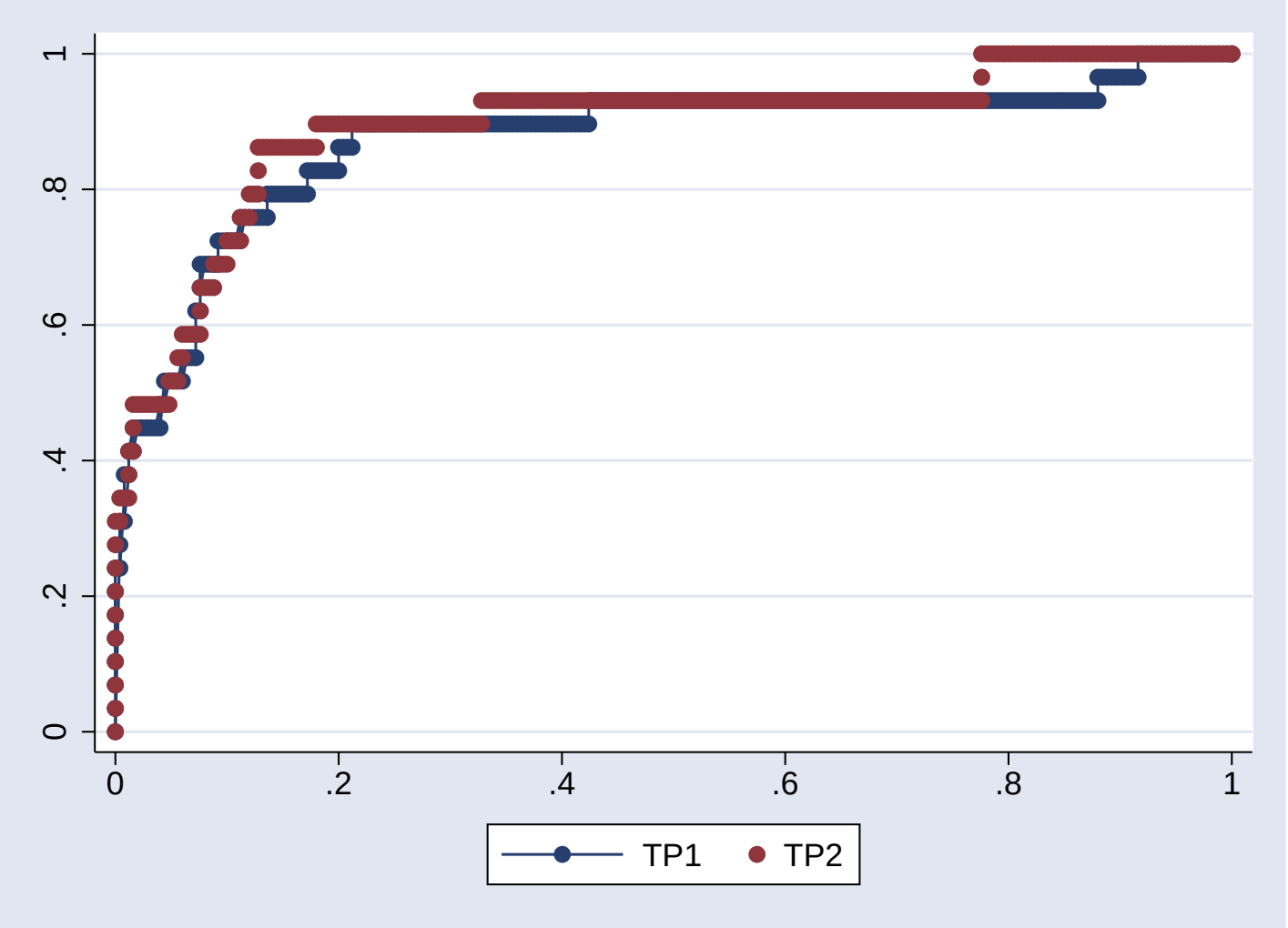




Example: 2-year mortality and Mayo PBC



Example: 2-year mortality and Mayo PBC

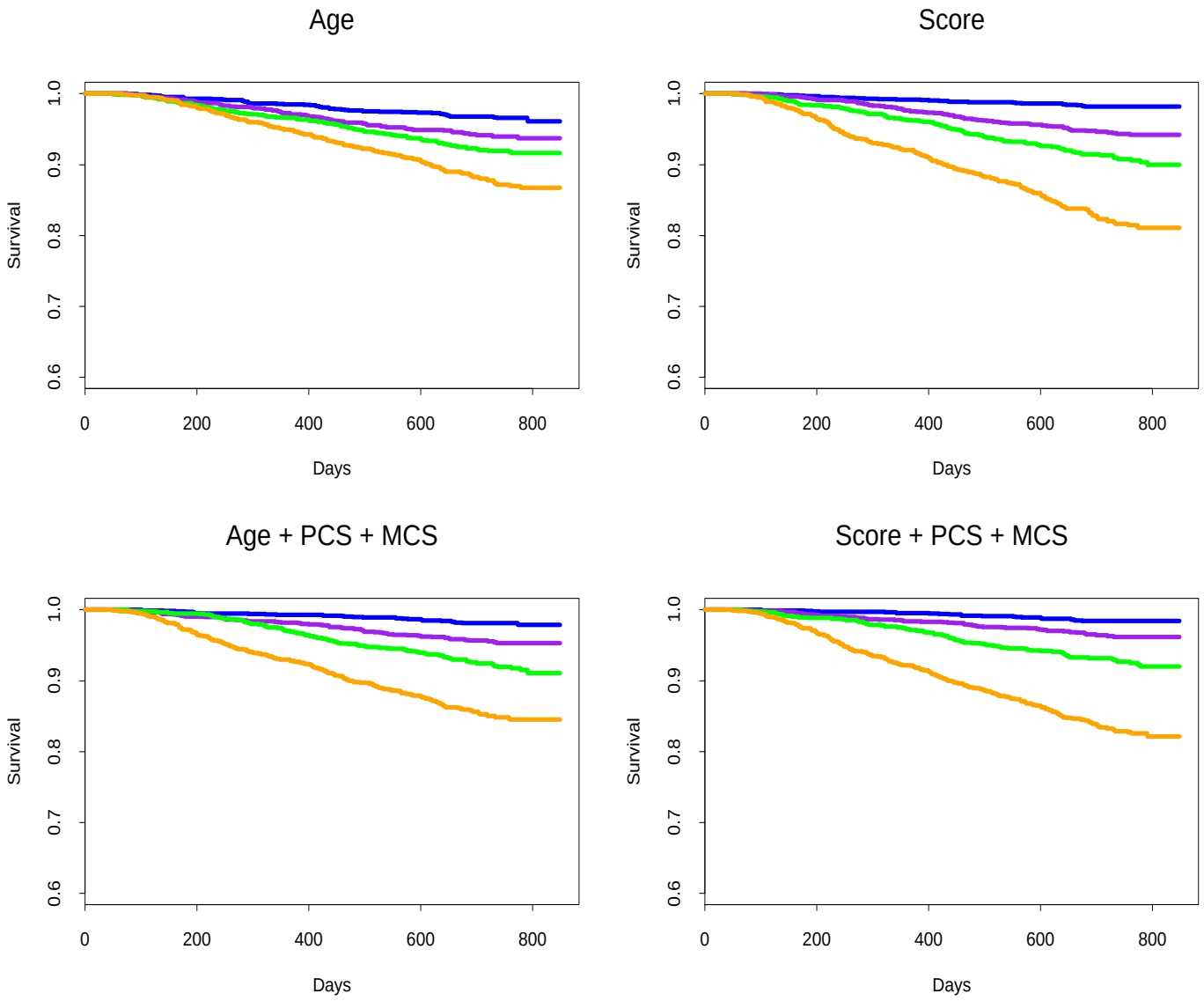


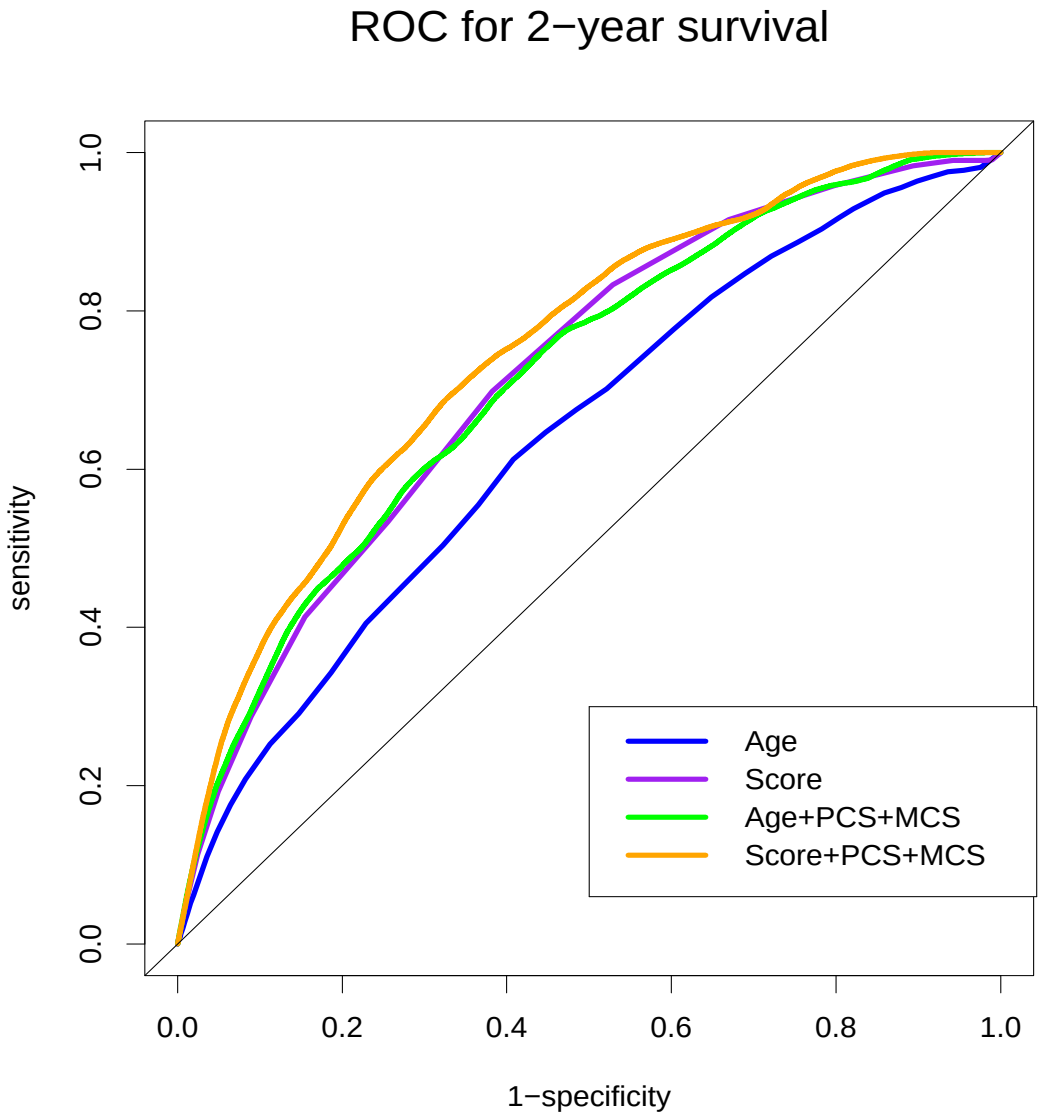
Example: VA ACQUIP Data

- Cox model with `age` only.
$$M_1 = \widehat{\beta}_1 X_1 \text{ for validation data}$$
- Cox model with `age` and `score`.
$$M_2 = \widehat{\beta}_2 X_2 \text{ for validation data}$$
- Cox model with `age` and `(PCS, MCS)`.
$$M_3 = \widehat{\beta}_3 X_3 \text{ for validation data}$$
- Cox model with `age`, `(PCS,MCS)` and `score`.
$$M_4 = \widehat{\beta}_4 X_4 \text{ for validation data}$$

Survival Model: VA ACQUIP Data

- 5,469 subjects used for model development of comorbidity index [SIC = Seattle Index of Comorbidity].
- 5,478 subjects used for model validation.
- AUC for model 3 = 0.71
- AUC for model 4 = 0.74, $p < 0.005$ for difference.
- See Fan et al. (2002) *J. Clin Epi.*


Age
Score
Age + PCS + MCS
Score + PCS + MCS
Survival
Days
0.6
0.7
0.8
0.9
1.0
0
200
400
600
800


ROC for 2−year survival
sensitivity
1−specificity
0.0
0.2
0.4
0.6
0.8
1.0
Age
Score
Age+PCS+MCS
Score+PCS+MCS

Accuracy: Some other proposals

R^2 Generalizations

- Korn and Simon (1990)
- Schemper and Henderson (2000)
- O'Quigley and Xu (2001)

TP, FP, ROC Generalizations

- Etzioni et al. (1999); Slate and Turnbull (1999)
- Heagerty, Lumley, and Pepe (2000)

Concordance (c-index)

- Harrell et al. (1996)
- Heagerty & Zheng (2005)

R^2: Schemper and Henderson (2000)

Idea: $D(t) = 1(T \leq t)$ with $E[D(t)] = 1 - S(t)$

	Without Covariates	With Covariates
variance	$S(t)[1 - S(t)]$	$S(t \mid X)[1 - S(t \mid X)]$
average (X)		$E_X \{S(t \mid X)[1 - S(t \mid X)]\}$
average (T)	$\int_t S(t)[1 - S(t)]f(t)dt$	$\int_t E_X \{S(t \mid X)[1 - S(t \mid X)]\} f(t)dt$
	$\downarrow$	$\downarrow$
	V_0	V_X

Proposal: $R^2 = (V_0 - V_X)/V_0$

Some Comments

- Schemper and Henderson (2000), p. 249:

  "Consequently, there have been a number of attempts to develop measures akin to R^2 for Cox proportional hazards models [[*references*]], though as yet, none have been generally accepted."

- Their R^2 is not about variance in T.
- Natural to think of survival through counting process $N(t)$.
- Uncommon to use R^2 for logistic regression / binary classification.

Summary: Predictive Models

- Clear goal of **prediction** (how will this be used?)
- Need to define an **"error"** measurement scale
  - ▷ Distance for a continuous measurement
  - ▷ Classification errors for discrete
- Need to obtain **honest** estimates of error rates
  - ▷ Test data (external)
  - ▷ Cross-validation
- Computational problem of searching candidate models.