How to Present Results of Regression Models to Clinicians

Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University School of Medicine
f.harrell@vanderbilt.edu
biostat.mc.vanderbilt.edu/FHHandouts

Outline

1. Regression models: Uses and problems
2. Example of interpreting coefficients
3. Goals of software to assist analyst
4. Software
5. Example binary logistic regression fit
6. Typesetting model fits
7. Summarizing effects of predictors
8. Displaying shape of effect of a predictor
9. Displaying how predicted values are formed
10. Tree approximations
11. Interactively obtaining predictions

Uses of Models for Individual Patients

- Probability of having a specific disease
- Probability of getting a disease by t years
- Lifetime risk of disease
- Expected time until disease development
- Life expectency given risk factors, disease damage
- Risk reduction due to a new drug

Problems in Obtaining Predictions

- Many predictor variables
- Predictor variables act nonlinearly
- Predictor variables act non–additively (e.g., multiply)
- Often make a nonlinear transformation of the linear predictor ($X\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$) to obtain predicted response
- Want **patterns** and **predictions** to be understandable to clinicians
- Regression models are not “black boxes”

Example: Interpreting Coefficients

$$\begin{aligned} E(Y|age, sex) &= \beta_0 + \beta_1 age + \beta_2(sex = f) \\ &+ \beta_3 age(sex = f), \end{aligned}$$

where $sex = f$ is a dummy indicator variable for sex=female, i.e., reference cell is sex=male.

Parameter	Meaning
β_0	$E(Y\|age = 0, sex = m)$
β_1	$E(Y\|age = x + 1, sex = m) -$
	$E(Y\|age = x, sex = m)$
β_2	$E(Y\|age = 0, sex = f) -$
	$E(Y\|age = 0, sex = m)$
β_3	$E(Y\|age = x + 1, sex = f) -$
	$E(Y\|age = x, sex = f) -$
	$[E(Y\|age = x + 1, sex = m) -$
	$E(Y\|age = x, sex = m)]$

Goals of Software

1. Document and represent fitted model algebraically
2. Estimating odds ratios for meaningful changes in predictors, in presence of nonlinearity, interaction
3. Displaying multiple–level confidence bars
4. Understanding predictor transformation/strength
5. Easily obtaining predicted values

Software

- R (`www.r-project.org`),
  S-PLUS (Insightful Corp.)
- Object–oriented, user–extendible
- `Design` add–on package (Harrell)
- `Design` freely available (as is R)

Example Logistic Regression Model

- Re–analysis of bacterial vs. viral meningitis

  (Spanos et al. JAMA 1989)

- $n = 422$ (217 bacterial)

```
time.summer ← function(x)
  pmin(abs(x-7.5), abs(x+12-7.5))

f ← lrm(abm ∼  lsp(age, c(1, 2, 22)) +
               time.summer(month) +
               rcs(glratio, 5) +
               tpolys^0.33333)
```

	Coef	S.E.	Wald Z	P
Intercept	6.4887	7.73326	0.84	0.4014
age	2.1031	0.96654	2.18	0.0296
age’	-4.7292	1.58966	-2.97	0.0029
age’’	2.5289	0.76804	3.29	0.0010
age’’’	0.1754	0.06120	2.87	0.0042
month	0.4928	0.11878	4.15	0.0000
glratio	-26.5169	33.19172	-0.80	0.4243
glratio’	37.0256	114.38492	0.32	0.7462
glratio’’	-44.7037	275.91374	-0.16	0.8713
glratio’’’	-2.2335	264.53421	-0.01	0.9933
tpolys	0.2543	0.05976	4.26	0.0000

```
anova(f)
```

	χ^2	$d.f.$	P
age	33.1	4	< 0.0001
Nonlinear	32.9	3	< 0.0001
month	17.2	1	< 0.0001
glratio	22.8	4	< 0.0001
Nonlinear	22.4	3	0.0001
tpolys	18.1	1	< 0.0001
TOTAL NONLINEAR	49.3	6	< 0.0001
TOTAL	81.9	10	< 0.0001

```
latex(f)
```

$$\text{Prob}\{\text{abm} = 1\} = \frac{1}{1 + \exp(-X\beta)}, \text{ where}$$

$$\begin{aligned} X\hat{\beta} = & \\ & 6.49 \\ & +2.103\ \text{age} - 4.729(\text{age} - 1)_+ \\ & +2.529(\text{age} - 2)_+ + 0.175(\text{age} - 22)_+ \\ & +0.493\ \text{time.summer}(\text{month}) \end{aligned}$$

$$
\begin{aligned}
&-26.52\ \text{glratio} \\
&+42.61(\text{glratio} - 0.0395)_+^3 \\
&-51.45(\text{glratio} - 0.2712)_+^3 \\
&-2.57(\text{glratio} - 0.4954)_+^3 \\
&-7.24(\text{glratio} - 0.6317)_+^3 \\
&+18.65(\text{glratio} - 0.9716)_+^3 \\
&+0.254\ \text{tpolys}^{0.33333}
\end{aligned}
$$

and $(x)_+ = x$ if $x > 0,\ 0$ otherwise.

Summarizing Effects of Predictors

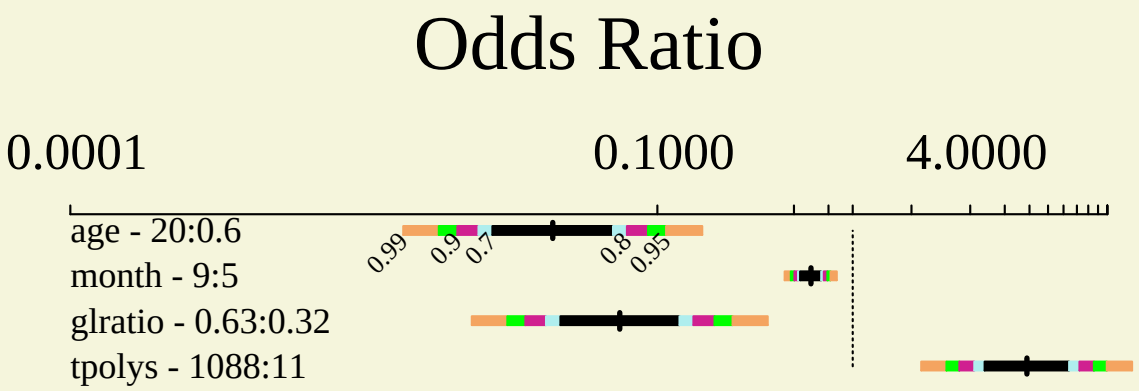
- Even if model is simple, must choose increment carefully
- Odds ratio for 1–cell increase in leukocyte count?
- If predictor is nonlinear and is represented as multiple model terms, can't look at individual coefficients
- Odds ratio for a 1–unit increase in age^2 holding age constant?
- Default: inter–quartile–range ORs
- Get predicted log odds at lower and upper quartiles, take difference, anti–log
- Set levels of interacting factors

```
s ← summary(f, month=c(1.5, 7.5))
print(s)
```

```
    Effects                  Response : abm

      Factor    Low      High Effect  S.E.
 age           0.600   20.000  -3.53  0.68
  Odds Ratio   0.600   20.000   0.03
 month         1.500    7.500  -2.96  0.71
  Odds Ratio   1.500    7.500   0.05
 glratio       0.321    0.629  -2.74  0.68
  Odds Ratio   0.321    0.629   0.06
 tpolys       10.950 1087.500   2.05  0.48
  Odds Ratio 10.950 1087.500   7.77
```

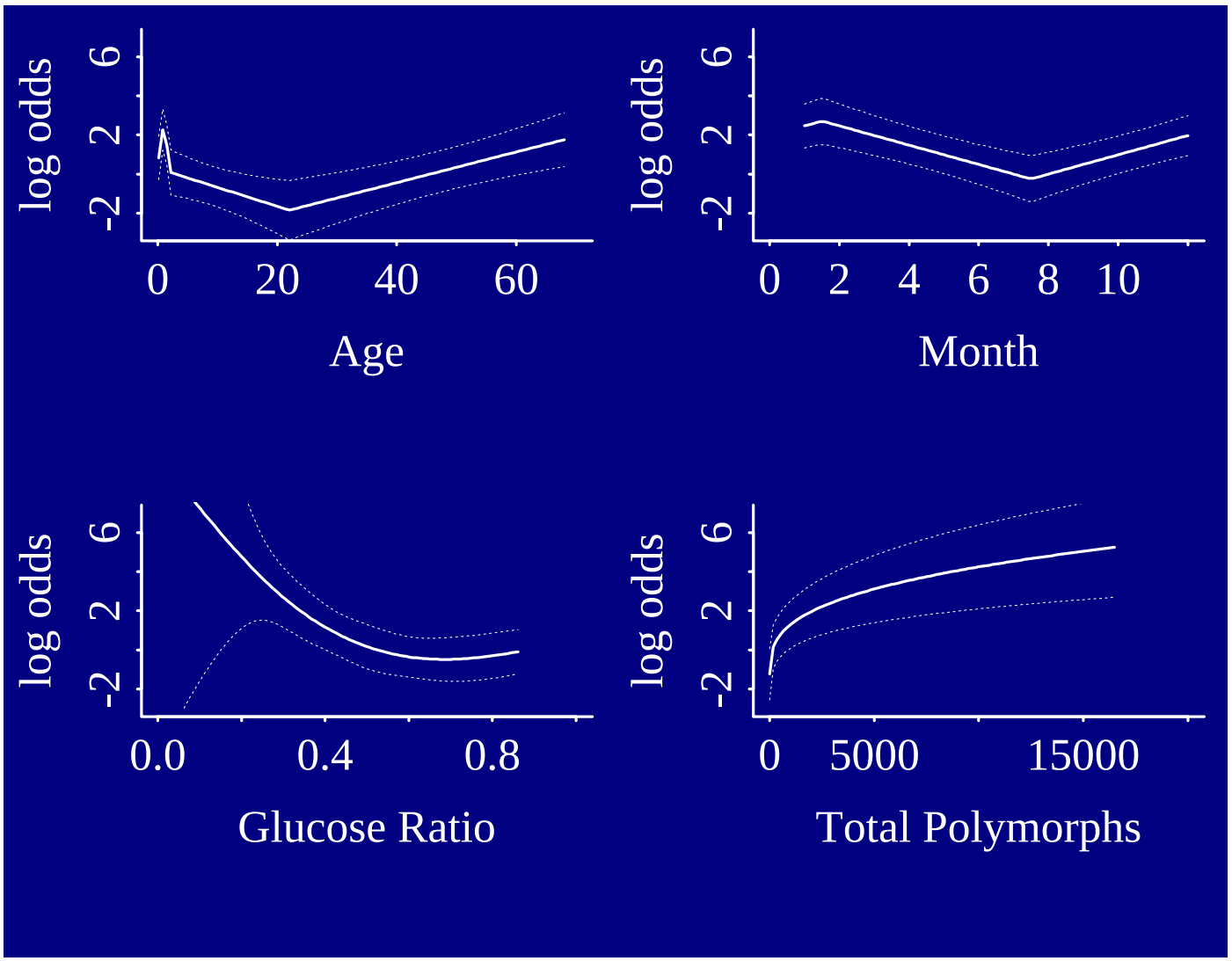
plot(s)


Odds Ratio
0.0001
0.1000
4.0000
age - 20:0.6
month - 9:5
glratio - 0.63:0.32
tpolys - 1088:11
0.99
0.9
0.7
0.8
0.95

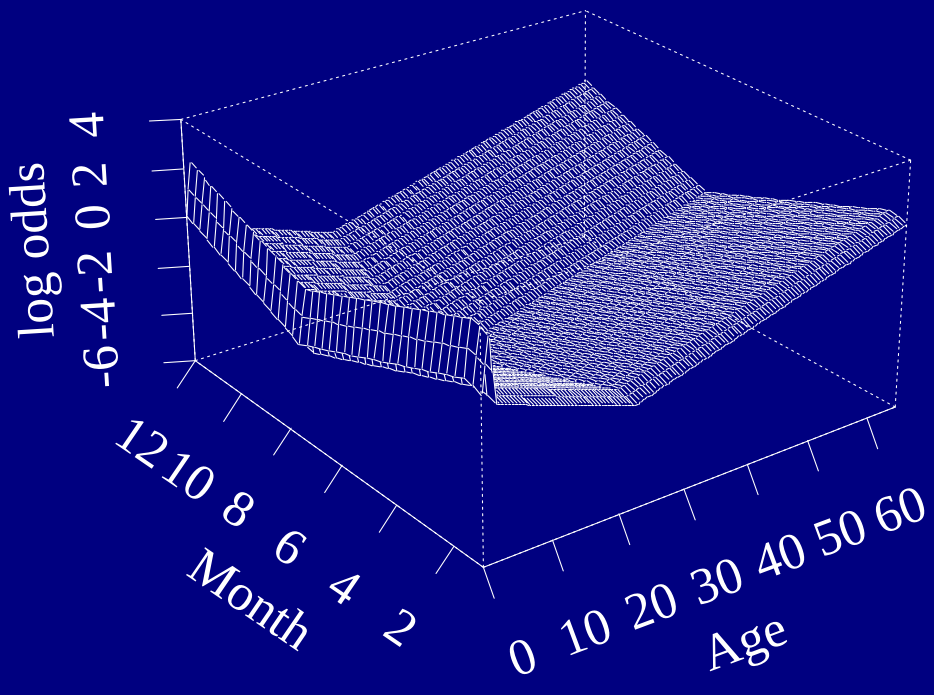
Displaying Shape of Effect

- Fix levels of interacting factors
- Vary X, show how X is related to logit or prob.
- Just a problem in getting predictions, remembering all variable transformations

```
plot(f)
```


log odds
-2
2
6
0
20
40
60
Age
log odds
-2
2
6
0
2
4
6
8
10
Month
log odds
-2
2
6
0.0
0.4
0.8
Glucose Ratio
log odds
-2
2
6
0
5000
15000
Total Polymorphs

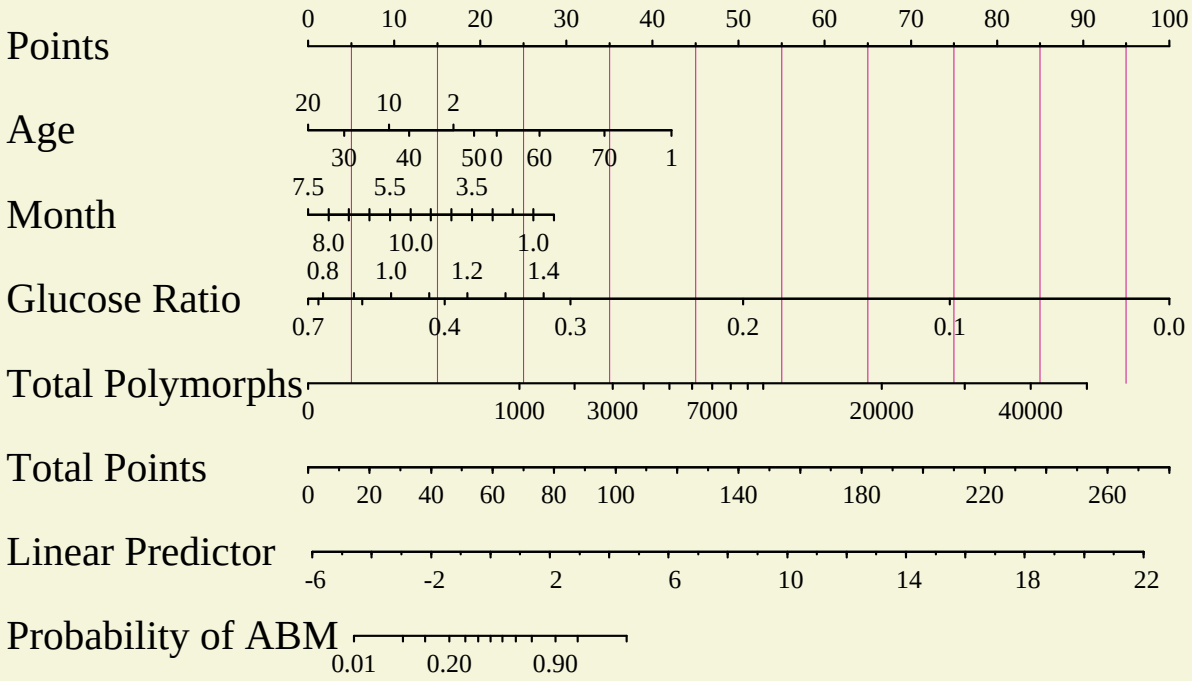
plot(f, age=NA, month=NA)



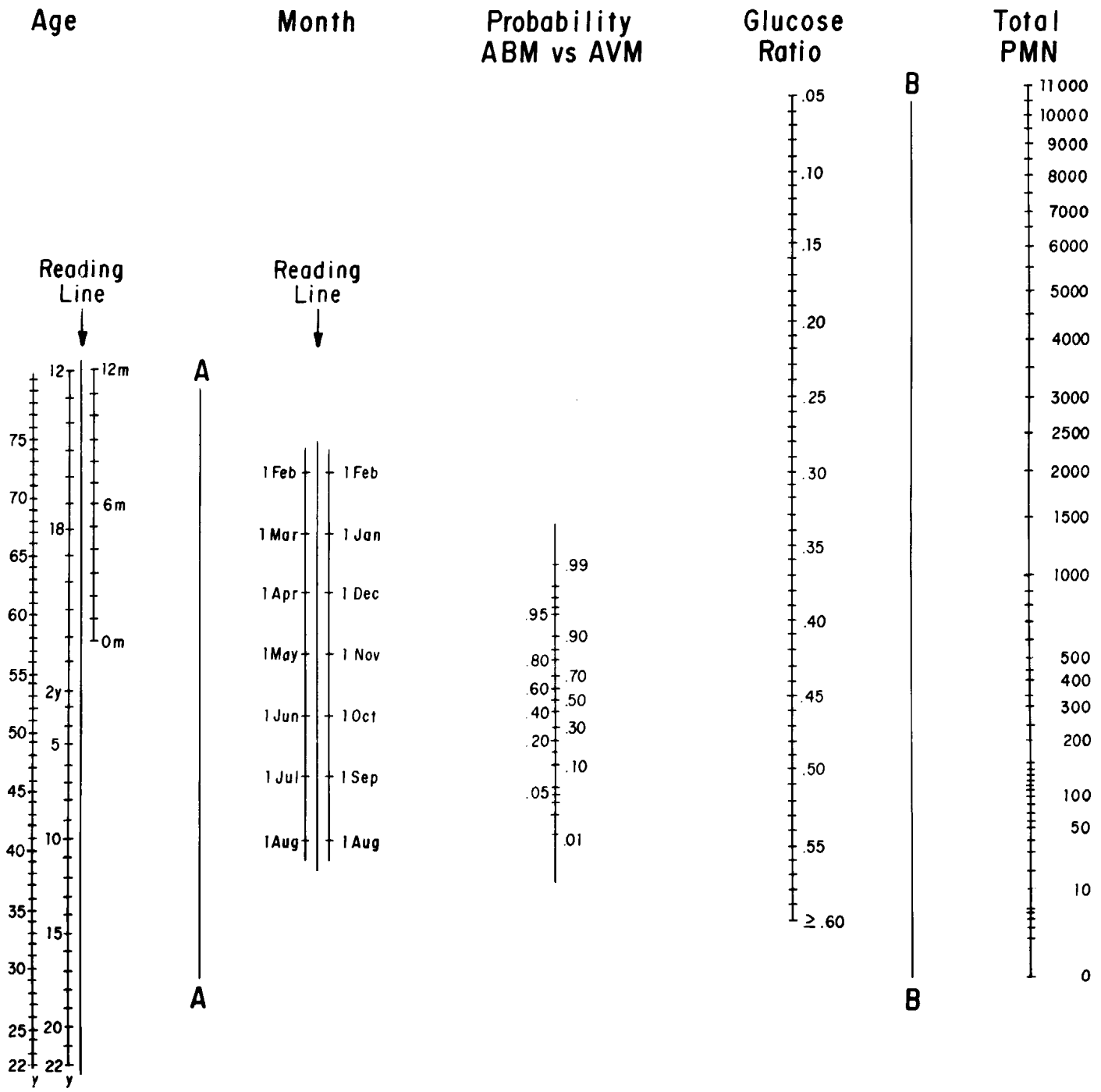
Displaying How Predictions are Formed

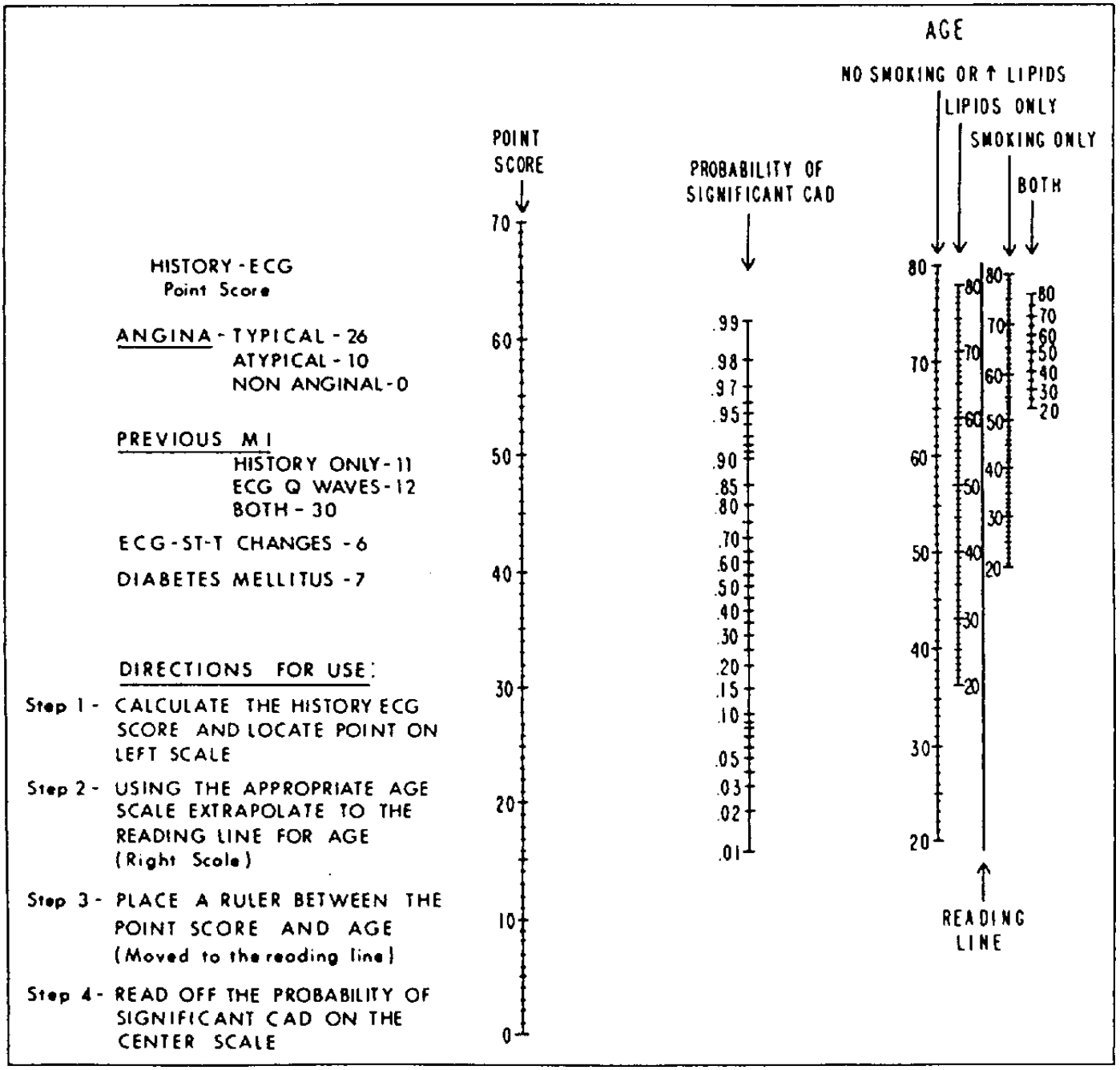
- Nomogram
- One axis per predictor
- And per level of interacting factors
- Use “point scores”
- Show $X\hat{\beta}$ and $\widehat{\mathrm{Prob}}$

```
nomogram(f,
         fun=function(x) 1/(1+exp(-x)),
         funlabel='Probability of ABM')
```


Points
0 10 20 30 40 50 60 70 80 90 100
Age
20 10 2
30 40 50 0 60 70 1
Month
7.5 5.5 3.5
8.0 10.0 1.0
Glucose Ratio
0.8 1.0 1.2 1.4
0.7 0.4 0.3 0.2 0.1 0.0
Total Polymorphs
0 1000 3000 7000 20000 40000
Total Points
0 20 40 60 80 100 140 180 220 260
Linear Predictor
-6 -2 2 6 10 14 18 22
Probability of ABM
0.01 0.20 0.90

Examples of Manually Drawn Nomograms

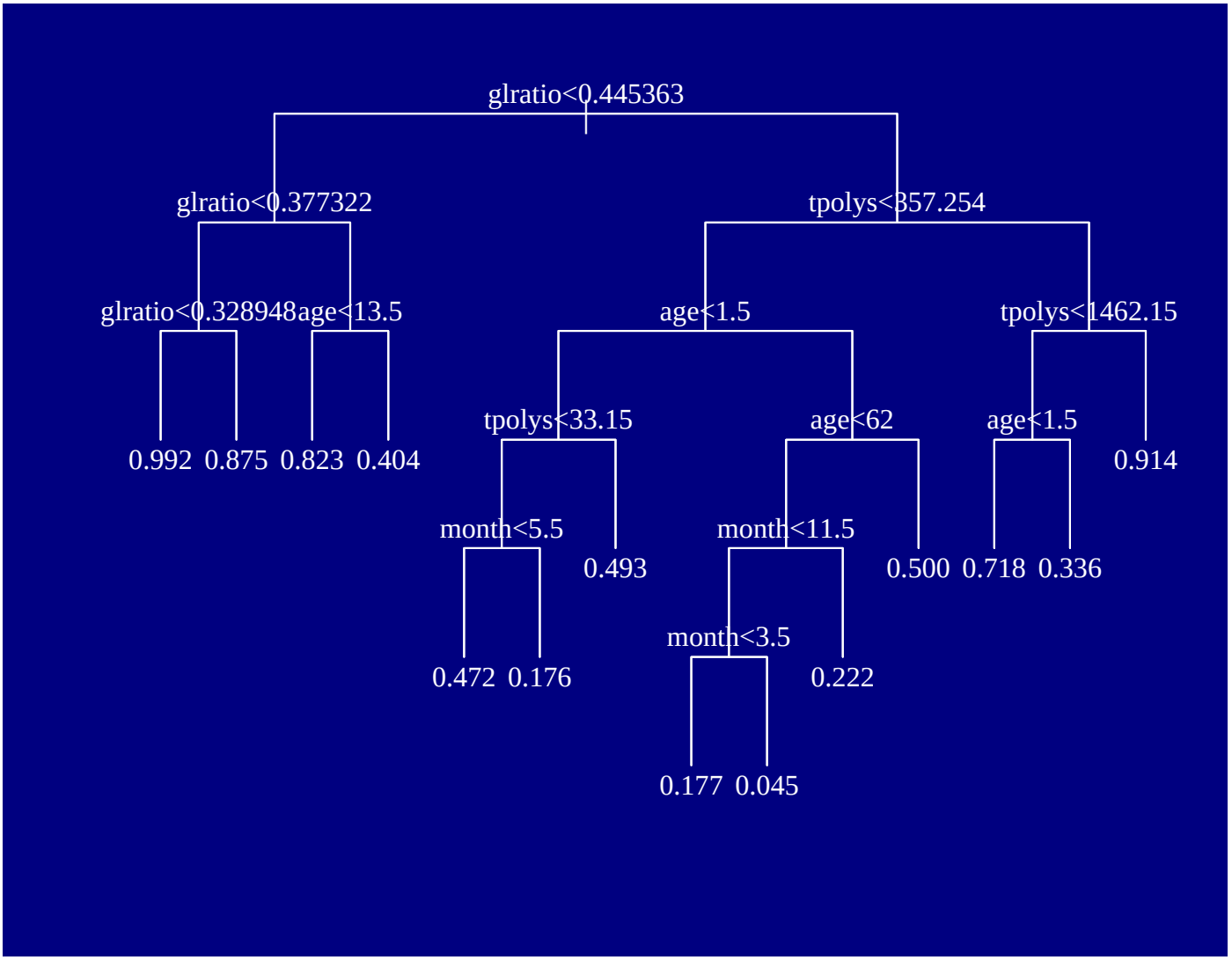



HISTORY - ECG
Point Score
ANGINA - TYPICAL - 26
ATYPICAL - 10
NON ANGINAL - 0
PREVIOUS MI
HISTORY ONLY - 11
ECG Q WAVES - 12
BOTH - 30
ECG - ST-T CHANGES - 6
DIABETES MELLITUS - 7
DIRECTIONS FOR USE:
Step 1 - CALCULATE THE HISTORY ECG SCORE AND LOCATE POINT ON LEFT SCALE
Step 2 - USING THE APPROPRIATE AGE SCALE EXTRAPOLATE TO THE READING LINE FOR AGE (Right Scale)
Step 3 - PLACE A RULER BETWEEN THE POINT SCORE AND AGE (Moved to the reading line)
Step 4 - READ OFF THE PROBABILITY OF SIGNIFICANT CAD ON THE CENTER SCALE
POINT SCORE
70
60
50
40
30
20
10
0
PROBABILITY OF SIGNIFICANT CAD
.99
.98
.97
.95
.90
.85
.80
.70
.60
.50
.40
.30
.20
.15
.10
.05
.03
.02
.01
AGE
NO SMOKING OR ↑ LIPIDS
LIPIDS ONLY
SMOKING ONLY
BOTH
80
70
60
50
40
30
20
READING LINE

Tree Approximations

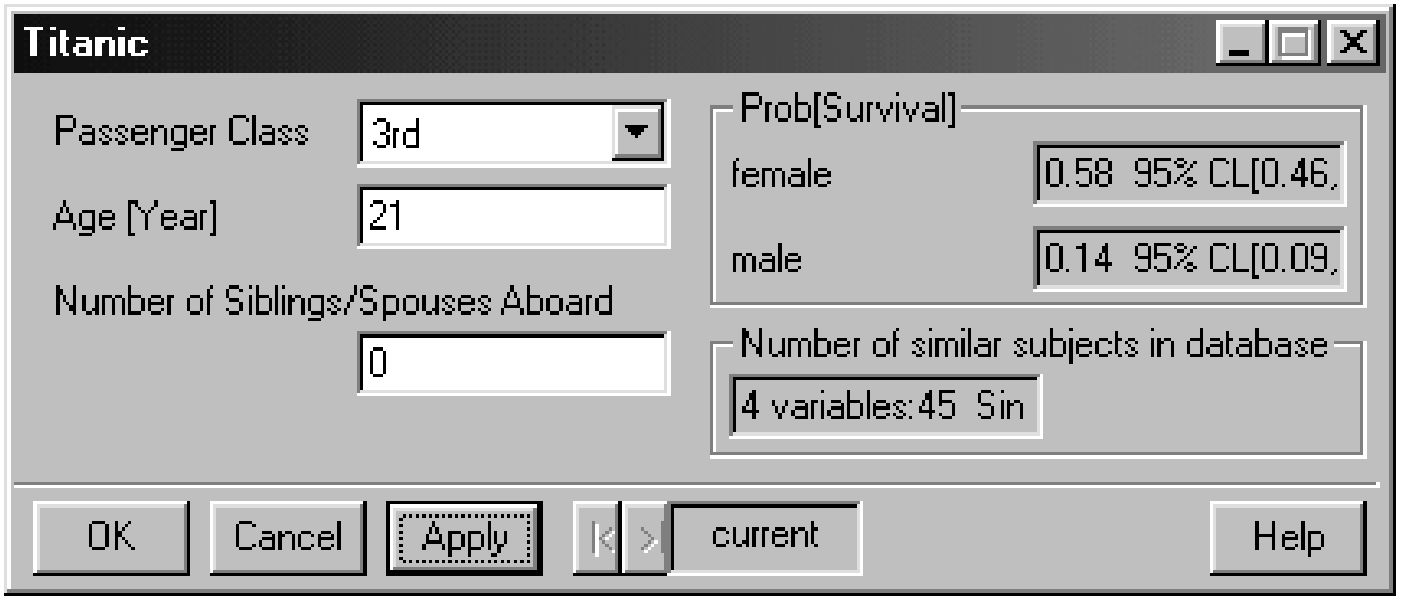
- Use a Regression Tree to approximate $\widehat{\mathrm{Prob}}$ to a desired degree of accuracy
- Can compute estimation error as a function of # nodes

```
pred ← predict(f, type='fitted')
f.approx ← tree(pred ~  age + month +
                        glratio + tpolys)
```


glratio<0.445363
glratio<0.377322
tpolys<357.254
glratio<0.328948
age<13.5
age<1.5
tpolys<1462.15
tpolys<33.15
age<62
age<1.5
0.914
0.992
0.875
0.823
0.404
month<5.5
0.493
month<11.5
0.500
0.718
0.336
month<3.5
0.472
0.176
0.222
0.177
0.045

Interactively Obtaining Predictions

```
drep ← dataRep( ~ roundN(age,10) + sex +
                  pclass +
                  roundN(sibsp, clip=0:1))
Dialog(fitPar('f.mi', lp=F,
       fun=list('Prob[Survival]'=plogis)),
       limits='data', basename='Titanic',
       vary=list(sex=c('female','male')),
       datarep=drep)
runmenu.Titanic()
```



Summary

- Automatic typesetting useful for simplifying nonlinear terms and interactions
- Estimates of effects should take into account scale, nonlinearities, interactions
- Variety of graphical methods useful for displaying strength, shape, predicted values
- Special interactive programs allow one to obtain predictions and confidence limits

Abstract

Regression models can flexibly summarize relationships of multiple variables to clinical outcomes. With increased flexibility, such as incorportation of nonlinear effects or interactions, comes greater difficulty of interpreting model coefficients. In this talk I will discuss graphical and semi-graphical methods for describing a fitted regression model, using the binary logistic model as an example. The methods demonstrated include typesetting the model equation, summarizing the effects of predictors with odds ratio charts, displaying the shape of the effect of a predictor, drawing nomograms, and using trees to approximate the model. These methods will be demonstrated using a re–analysis of a model to predict the probability that a patient will have bacterial meningitis (Spanos et al., JAMA 1989).