Examining validity and precision of prognostic models.

Dan McGee

Department of Statistics

Florida State University

dan@stat.fsu.edu

Acknowledgements

- The National Heart, Lung, and Blood Institute. Funding: HL67640

- The Diverse Populations Collaboration

- Validity
- Classification Efficacy
- Predictive Accuracy

DPC Collaborating Centres

USA
15 cohorts
>230,000 participants

ICELAND
1 cohort
>18,000 participants

SCOTLAND
2 cohorts
>22,000 participants

NORWAY
1 cohort
>48,000 participants

CHINA
1 cohort
>7,000 participants

Hawaii
1 cohort
>8,000 participants

PUERTO RICO
1 cohort
>9,000 participants

DENMARK
1 cohort
>10,000 participants

YUGOSLAVIA
1 cohort
>6,000 participants

ISRAEL
4 cohorts
>35,000 participants

- 21 Studies
- 49 strata (gender, race, etc.)
- 50+ CVD deaths (within 10 years)in each strata

- 219,973 Observations
  - 78,980 Female
  - 9,938 CVD deaths (within 10 years)

Some Published Framingham Risk Models.

Reference	Sample	Cases/Total	Model
1971 (Section 27)	2 year risk, people free of CHD, pool of exams 1-8.	Men: 370/31,704 Women: 206/41,834	Logistic
1973 (Section 28)	8 year risk, people free of CVD, pool of exams 2 and 6	Men 350/3813 Women 212/4960	Logistic
1987 (Section 37)	8 year risk, people free of CVD, pool of exams 2, 6, and 10	Men 523/4970 Women 359/6570	Logistic
1991 AHA (Circulation)	Pool of Exam 11 of cohort and Exam 1 of offspring free of CHD (12 year follow-up)	Men 385/2590 Women 241/2983	Accelerated Failure Time
1998 (Circulation)	Pool of Exam 11 of cohort and Exam 1 of offspring free of CHD.	Men 383/2489 Women 227/2856	Proportional Hazards, categorical data.

The Logistic Model

$$\Pr(Y=1|\mathbf{x}_i)=\pi_i=\frac{1}{1+\exp\left(-\sum_{j=0}^{p}x_{ij}\beta_j\right)}$$

$$\operatorname{logit}(\pi_i)=\log\left(\frac{\pi_i}{1-\pi_i}\right)=\sum_{j=0}^{p}x_{ij}\beta_j$$

$\mathbf{x}_i=(x_{0i,}x_{1i},\ldots,x_{pi})'$, a vector of characteristics, with $x_{i0}=1$

Age, age^{2,} Log(age), Log(age/74)
Cholesterol, Log(chol/hdl)
SBP, hypotensives, Diabetes, Smoker
Hypot.*SBP, Chol*age,
LVH-ECG, Atrial Fibrillation

Predict CVD death (10 years) based on:

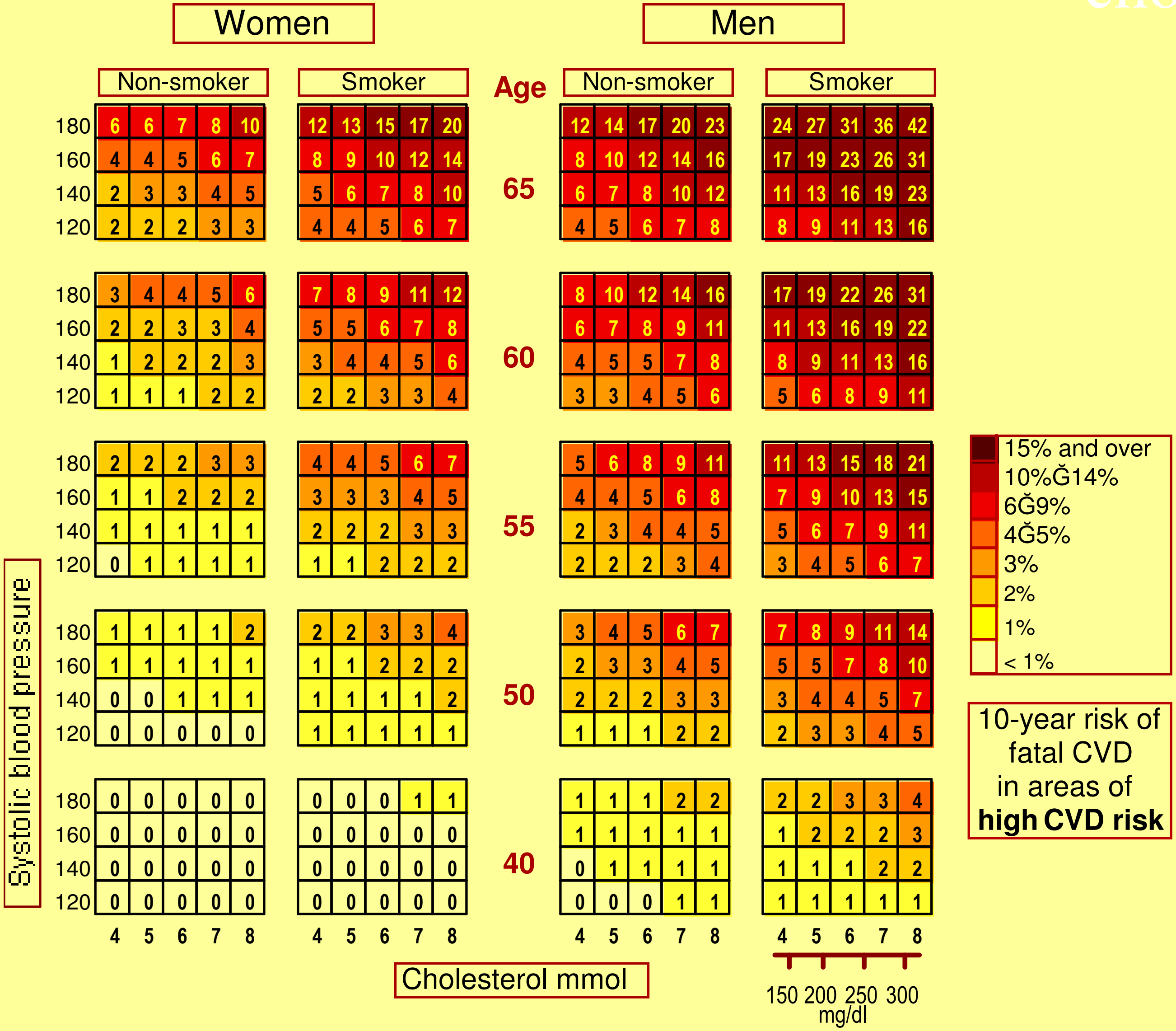
Age
Systolic blood pressure
Serum cholesterol
Diabetic status
Smoking status (yes/no)

Altman D and Royston P: What do we mean by validating a prognostic model? *Statist Med* 2000; **19**:453-473.

- Inform patients and their families.
- Create clinical risk groups for stratification.
- Inform treatment or other decisions for individual patients.

- Usefulness is determined by how well a model works in practice.

High CVD risk regions, risk based on total cholesterol



Age	Systolic blood pressure	Women										Men									
		Non-smoker					Smoker					Non-smoker					Smoker				
		4	5	6	7	8	4	5	6	7	8	4	5	6	7	8	4	5	6	7	8
65	180	6	6	7	8	10	12	13	15	17	20	12	14	17	20	23	24	27	31	36	42
	160	4	4	5	6	7	8	9	10	12	14	8	10	12	14	16	17	19	23	26	31
	140	2	3	3	4	5	5	6	7	8	10	6	7	8	10	12	11	13	16	19	23
	120	2	2	2	3	3	4	4	5	6	7	4	5	6	7	8	8	9	11	13	16
60	180	3	4	4	5	6	7	8	9	11	12	8	10	12	14	16	17	19	22	26	31
	160	2	2	3	3	4	5	5	6	7	8	6	7	8	9	11	11	13	16	19	22
	140	1	2	2	2	3	3	4	4	5	6	4	5	5	7	8	8	9	11	13	16
	120	1	1	1	2	2	2	2	3	3	4	3	3	4	5	6	5	6	8	9	11
55	180	2	2	2	3	3	4	4	5	6	7	5	6	8	9	11	11	13	15	18	21
	160	1	1	2	2	2	3	3	3	4	5	4	4	5	6	8	7	9	10	13	15
	140	1	1	1	1	1	2	2	2	3	3	2	3	4	4	5	5	6	7	9	11
	120	0	1	1	1	1	1	1	2	2	2	2	2	2	3	4	3	4	5	6	7
50	180	1	1	1	1	2	2	2	3	3	4	3	4	5	6	7	7	8	9	11	14
	160	1	1	1	1	1	1	1	2	2	2	2	3	3	4	5	5	5	7	8	10
	140	0	0	1	1	1	1	1	1	1	2	2	2	2	3	3	3	4	4	5	7
	120	0	0	0	0	0	1	1	1	1	1	1	1	1	2	2	2	3	3	4	5
40	180	0	0	0	0	0	0	0	0	1	1	1	1	1	2	2	2	2	3	3	4
	160	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	2	2	2	3
	140	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2
	120	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1

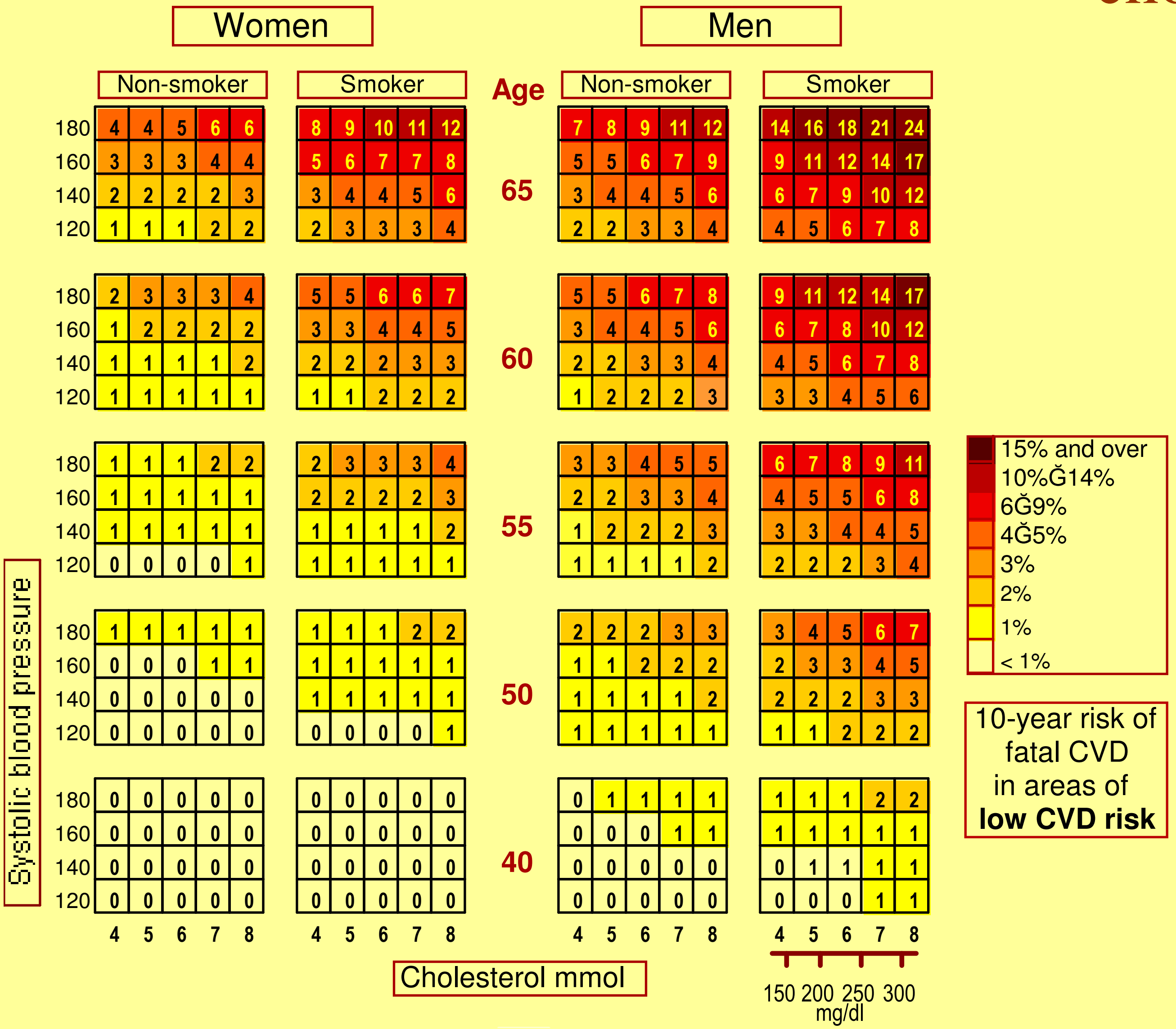
Cholesterol mmol

150 200 250 300
mg/dl

- ■ 15% and over
- ■ 10%Ğ14%
- ■ 6Ğ9%
- ■ 4Ğ5%
- ■ 3%
- ■ 2%
- ■ 1%
- □ < 1%

10-year risk of
fatal CVD
in areas of
high CVD risk

Low CVD risk regions, risk based on total cholesterol



		Women										Men									
		Non-smoker					Smoker					Non-smoker					Smoker				
Age	Systolic blood pressure	4	5	6	7	8	4	5	6	7	8	4	5	6	7	8	4	5	6	7	8
65	180	4	4	5	6	6	8	9	10	11	12	7	8	9	11	12	14	16	18	21	24
	160	3	3	3	4	4	5	6	7	7	8	5	5	6	7	9	9	11	12	14	17
	140	2	2	2	2	3	3	4	4	5	6	3	4	4	5	6	6	7	9	10	12
	120	1	1	1	2	2	2	3	3	3	4	2	2	3	3	4	4	5	6	7	8
60	180	2	3	3	3	4	5	5	6	6	7	5	5	6	7	8	9	11	12	14	17
	160	1	2	2	2	2	3	3	4	4	5	3	4	4	5	6	6	7	8	10	12
	140	1	1	1	1	2	2	2	2	3	3	2	2	3	3	4	4	5	6	7	8
	120	1	1	1	1	1	1	1	2	2	2	1	2	2	2	3	3	3	4	5	6
55	180	1	1	1	2	2	2	3	3	3	4	3	3	4	5	5	6	7	8	9	11
	160	1	1	1	1	1	2	2	2	2	3	2	2	3	3	4	4	5	5	6	8
	140	1	1	1	1	1	1	1	1	1	2	1	2	2	2	3	3	3	4	4	5
	120	0	0	0	0	1	1	1	1	1	1	1	1	1	1	2	2	2	2	3	4
50	180	1	1	1	1	1	1	1	1	2	2	2	2	2	3	3	3	4	5	6	7
	160	0	0	0	1	1	1	1	1	1	1	1	1	2	2	2	2	3	3	4	5
	140	0	0	0	0	0	1	1	1	1	1	1	1	1	1	2	2	2	2	3	3
	120	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	2	2	2
40	180	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2
	160	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
	140	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
	120	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

- 15% and over
- 10%Ğ14%
- 6Ğ9%
- 4Ğ5%
- 3%
- 2%
- 1%
- < 1%

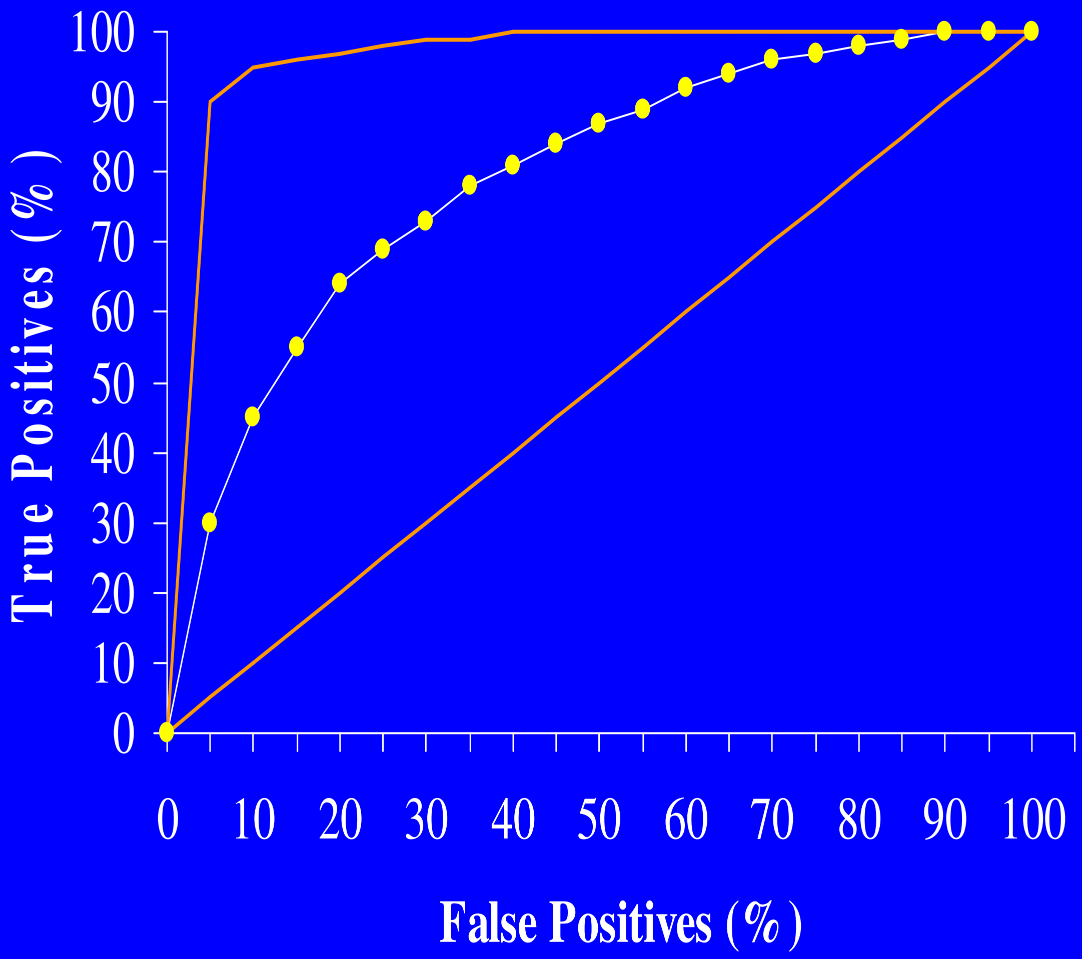
10-year risk of fatal CVD in areas of **low CVD risk**

Reliable **classification** of patients into different groups with different prognosis.

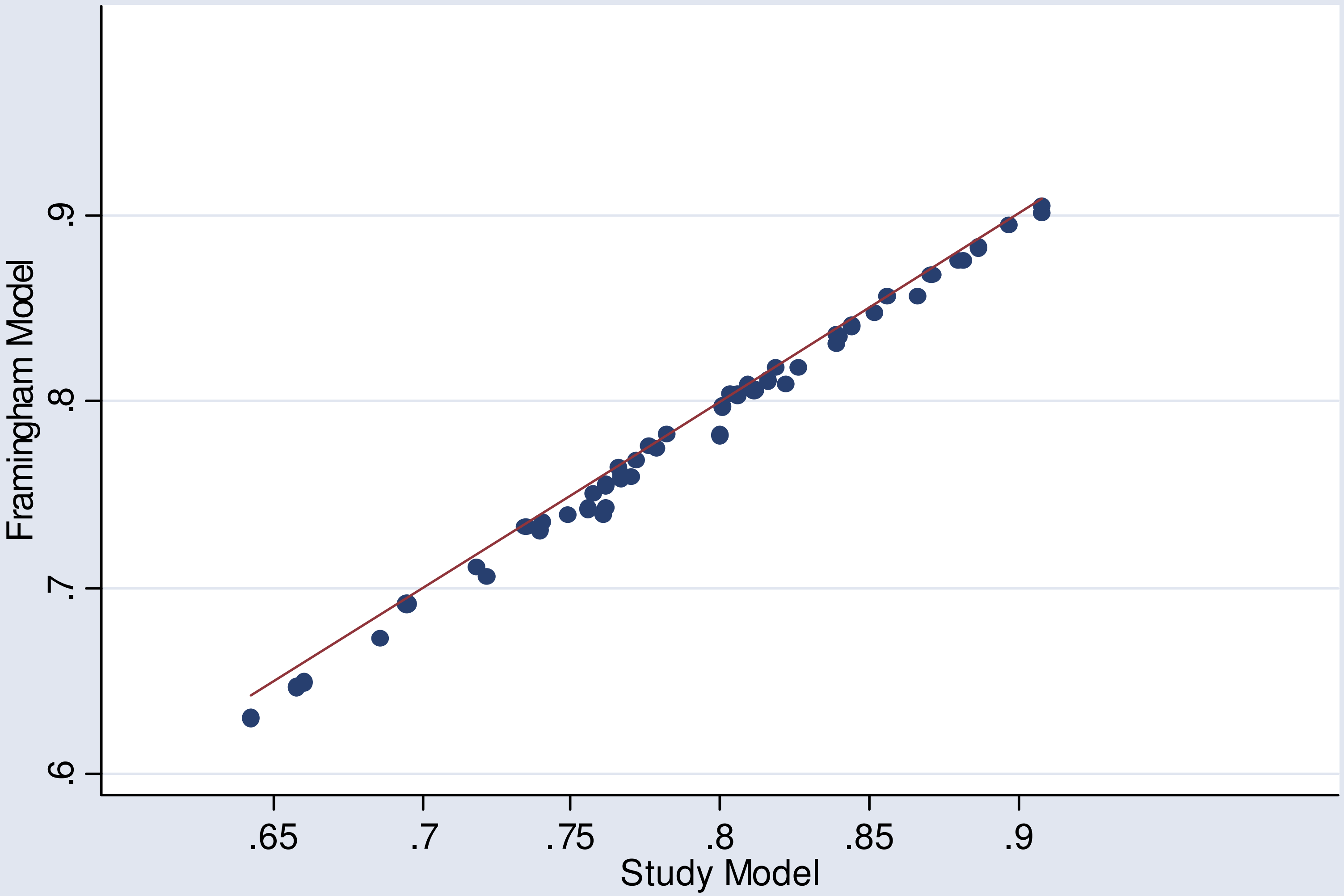
Area under the Receiver Operator Characteristic Curve

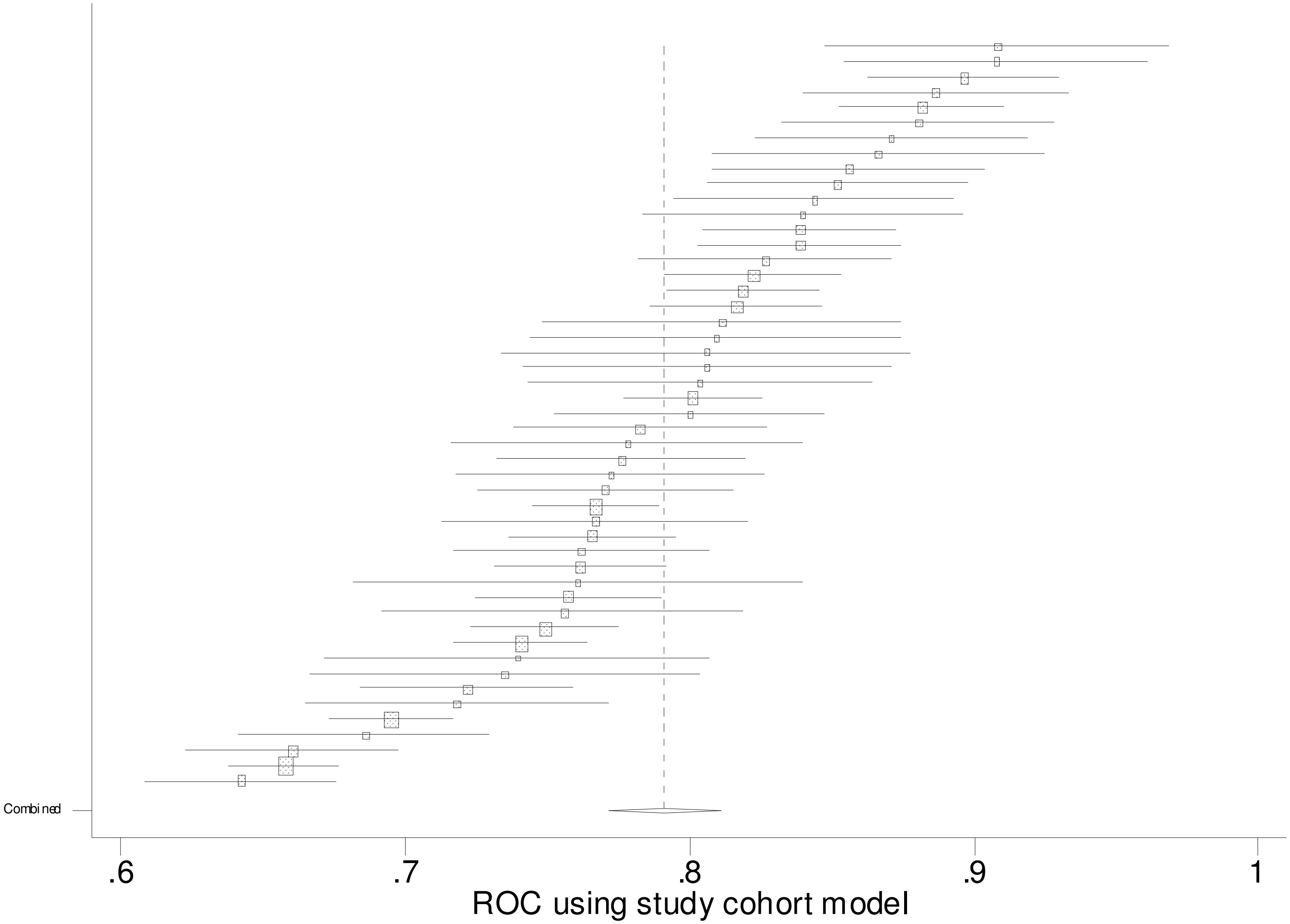
c-statistic, statistic of concordance.

Receiver Operating Characteristic (ROC) analysis



Area Under the ROC Curve





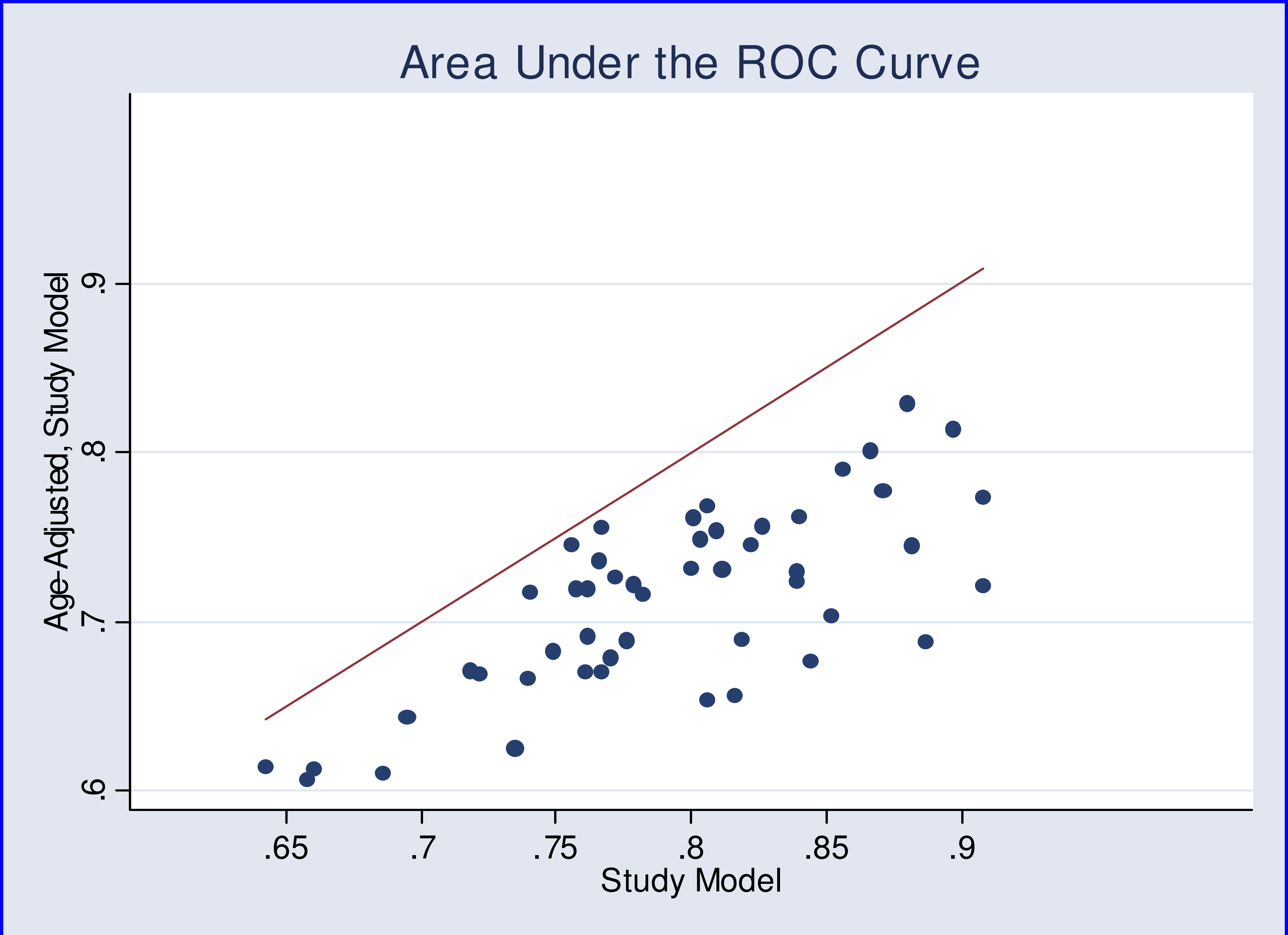
Random effects summary: .79 (.77,.81)

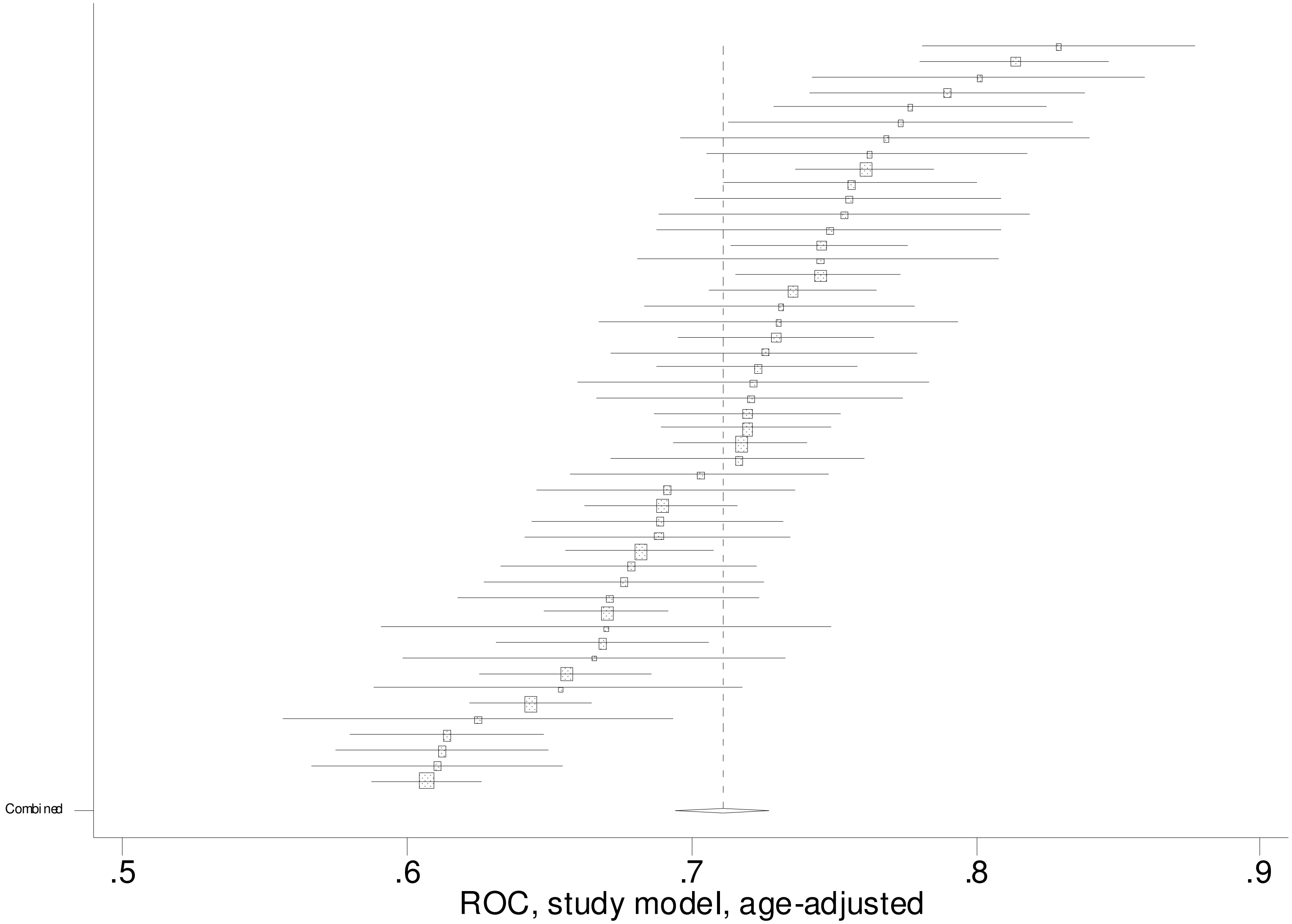
Ordering:

$$\hat{\beta}_0 + age * \hat{\beta}_1 + sbp_i * \hat{\beta}_2 + chol_i * \hat{\beta}_3 + smoking_i * \hat{\beta}_4 + diabetes_i * \hat{\beta}_5$$

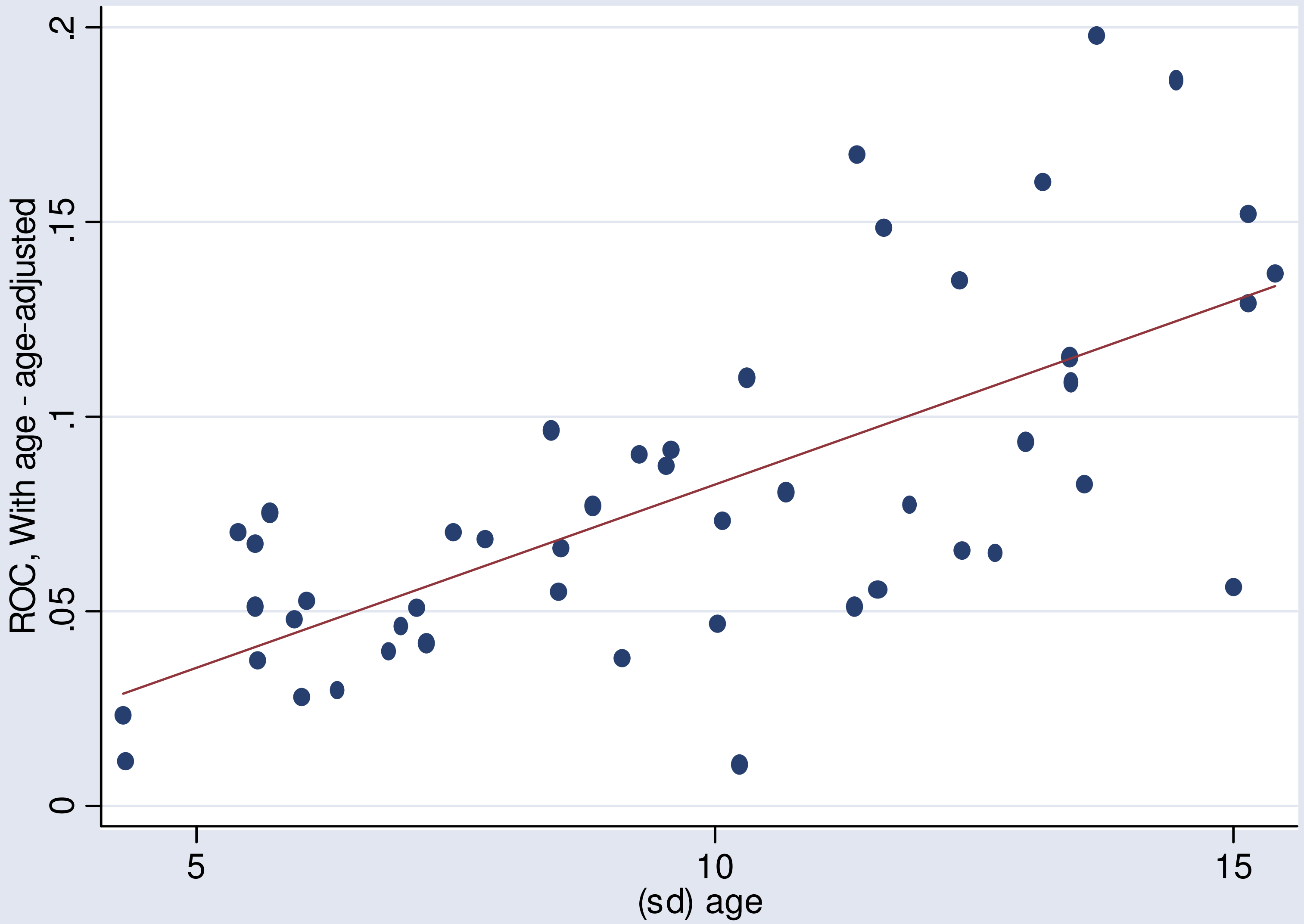
If everyone were the same age, the ordering would be determined by:

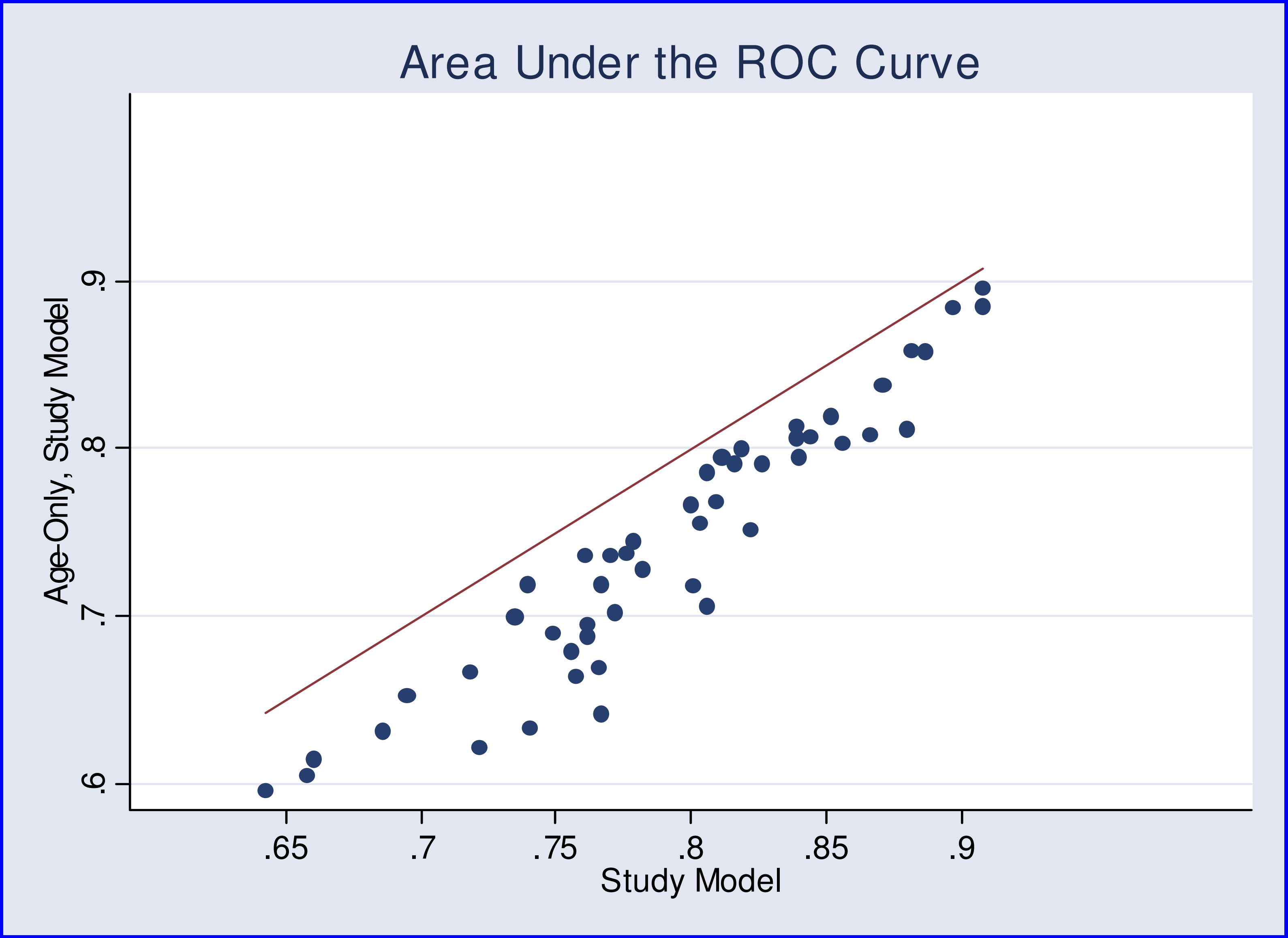
$$sbp_i * \hat{\beta}_2 + chol_i * \hat{\beta}_3 + smoking_i * \hat{\beta}_4 + diabetes_i * \hat{\beta}_5$$


Area Under the ROC Curve



Random effects summary: .71 (.70, .73)


ROC, With age - age-adjusted
0
.05
.1
.15
.2
5
10
15
(sd) age

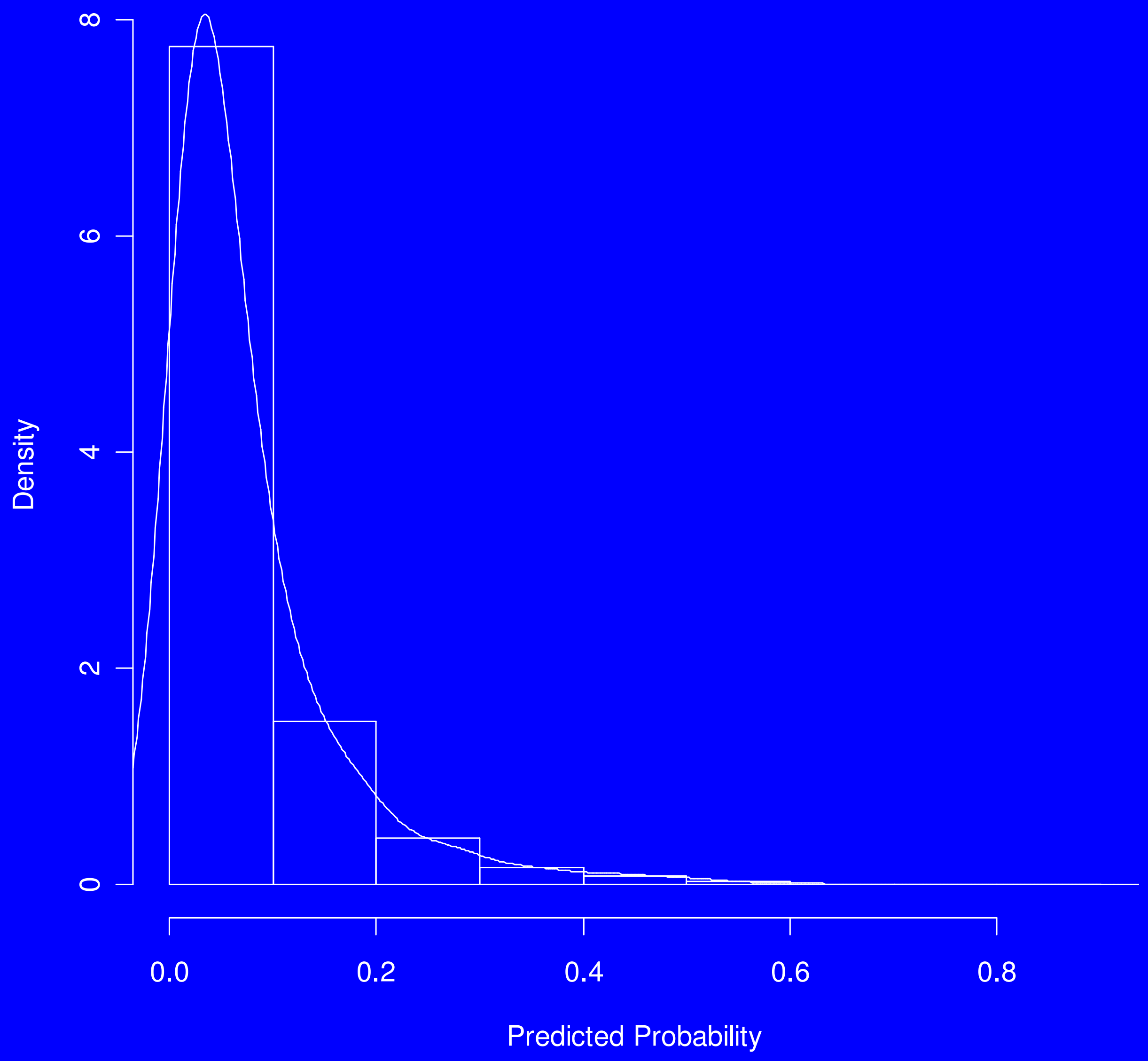


Area Under the ROC Curve

Classification Model (Gordon 1979)

Each person belongs to either one group or another.

Estimated probabilities tend to be a unimodal right-skewed distribution.



Framingham Males

How close are the estimated probabilities to the observed values.

Predictive Accuracy
Goodness of Fit
Explained Variation
Strength of association

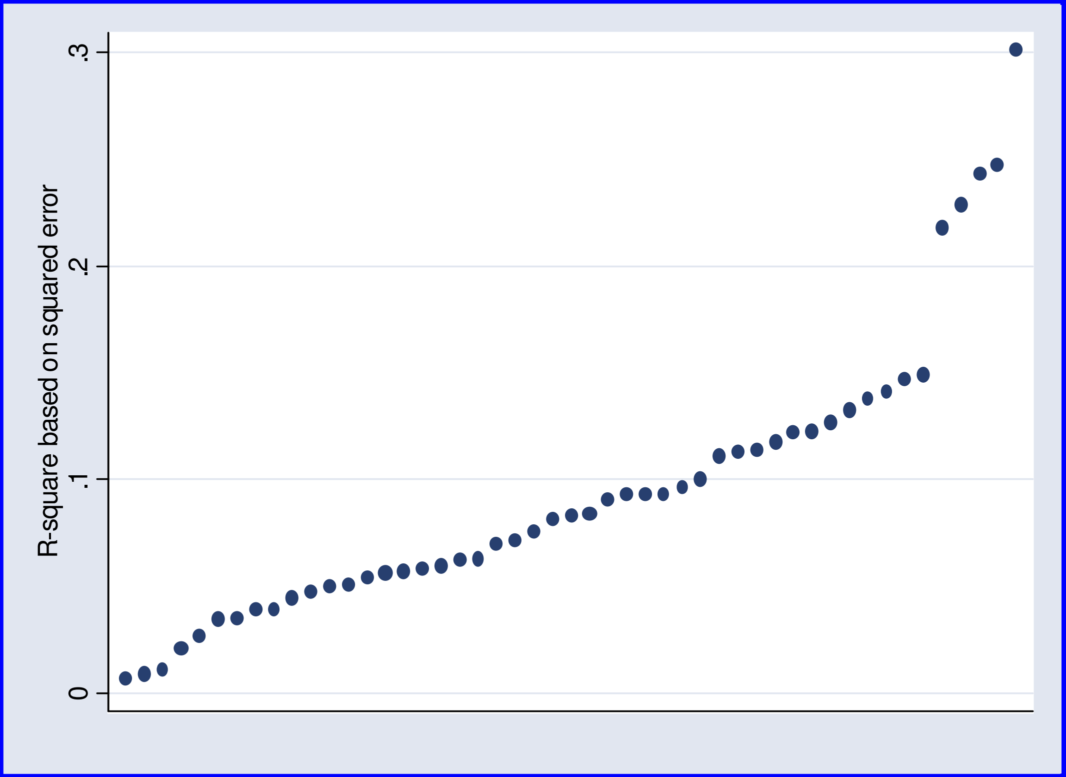
R^2

Ordinary Least Squares (OLS)

R^2

Coefficient of determination
Explained variance
Squared correlation, observed, predicted

$$R_0^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - p_i)^2}{\sum_{i=1}^{n} (y_i - \bar{p}_i)^2}$$



Average: .095

Gordon (1979)

p_i from a Beta Distribution with:

$$\alpha, \beta > 1$$

$$\bar{p} \leq \frac{1}{2}$$

$$R_{\mathrm{O}}^2 \leq \frac{2/3 - \bar{p}}{1 - \bar{p}}$$

$$\mathrm{R}_{\mathrm{O}}^{2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

minimizing $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ is not the criteria for developing estimates

$\mathrm{R}_{\mathrm{O}}^{2}$ can decrease with additional information (or even be negative)

The error sum of squares is the only reasonable criteria for judging residual variation in OLS. (Efron 1978)

Several exist for dichotomous dependent variables.

(Menard 2000)

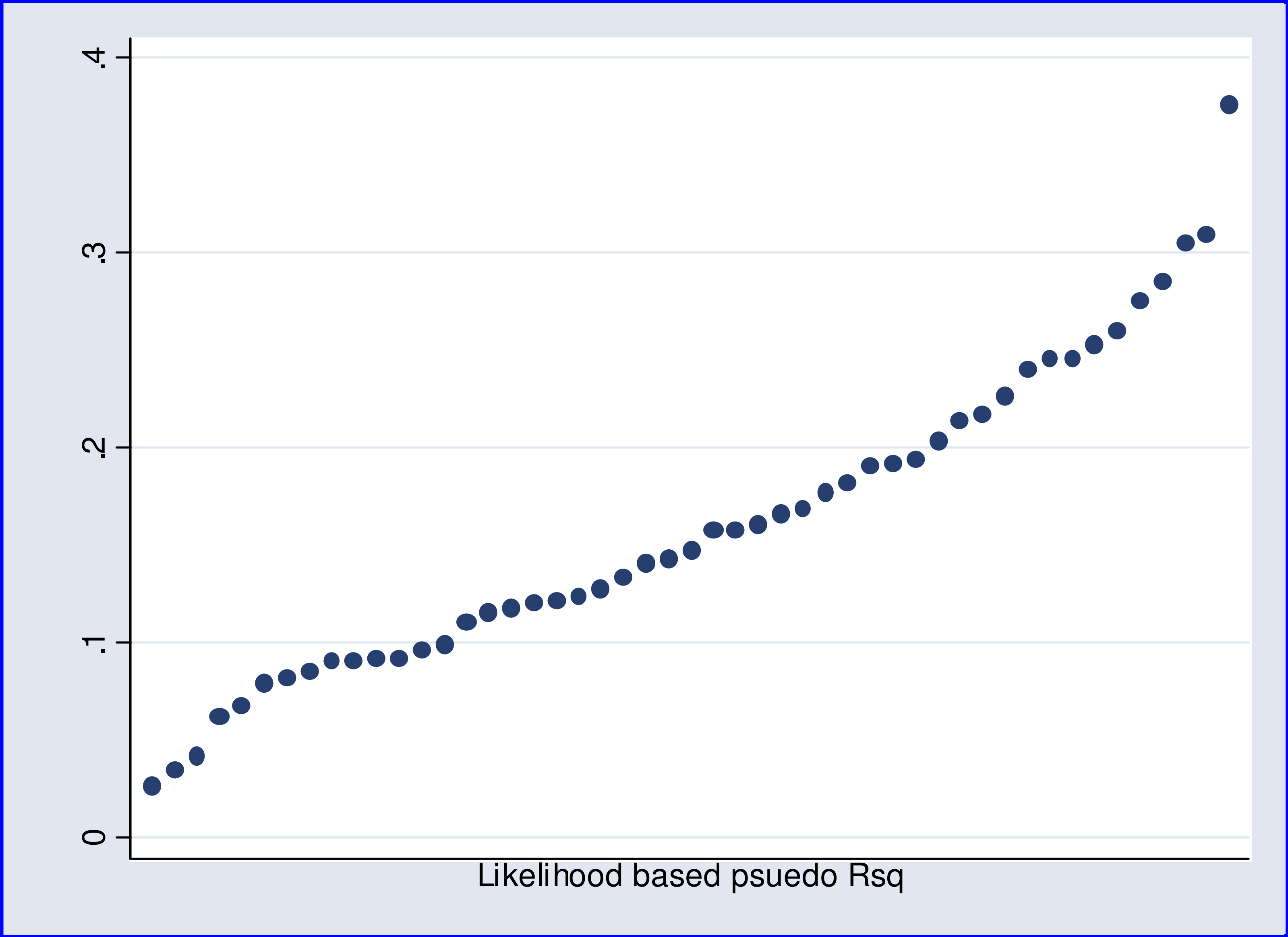
$$\mathrm{l}_{\mathrm{p}} = -\sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

(Negative log likelihood of p variable model)

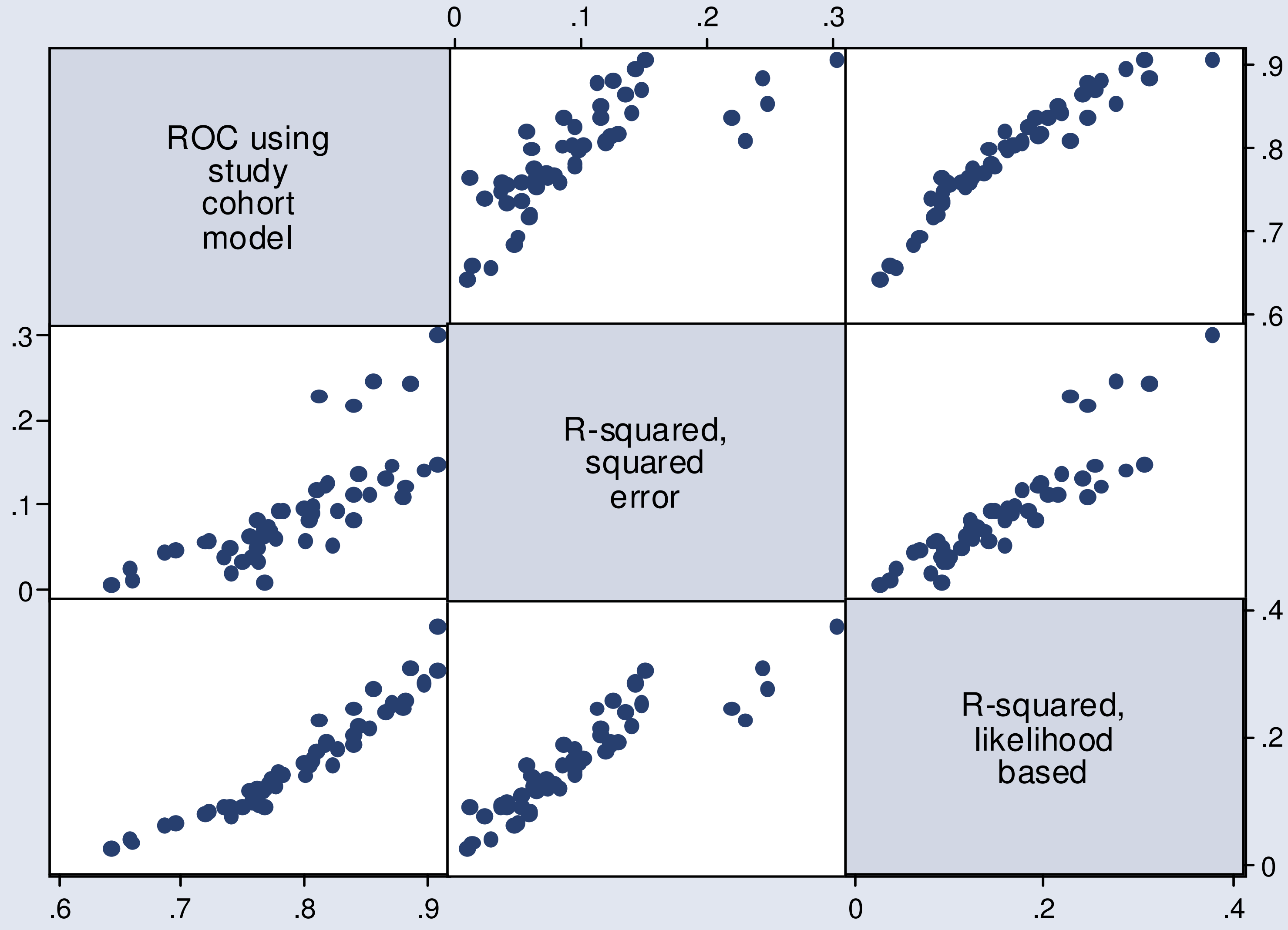
$$\mathrm{l}_{0} = -\sum_{i=1}^{n} y_i \log(\bar{y}) + (1 - y_i)\log(1 - \bar{y})$$

(Negative log likelihood of intercept only model)

$$\mathrm{R}_{\mathrm{L}}^{2} = 1 - \frac{\mathrm{l}_{\mathrm{p}}}{\mathrm{l}_{0}}$$



Average: .16


ROC using study cohort model
R-squared, squared error
R-squared, likelihood based
0
.1
.2
.3
.9
.8
.7
.6
.3
.2
.1
0
.4
.2
0
.6
.7
.8
.9
0
.2
.4